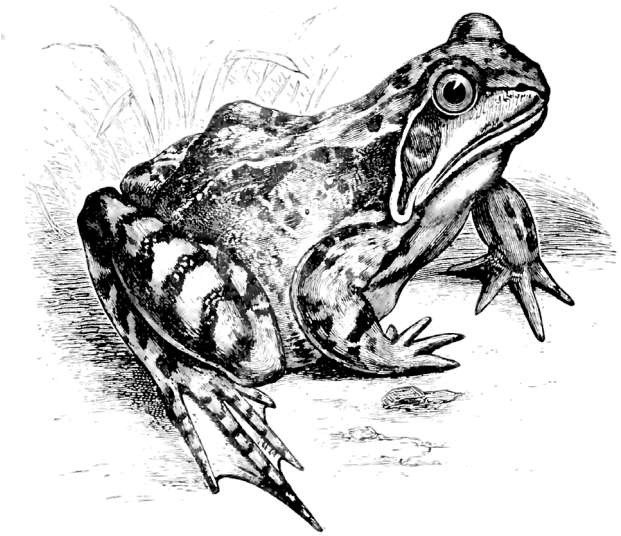
DOCTORAL THESIS
IN COMPUTATIONAL LINGUISTICS

WHY THE POND IS NOT *OUTSIDE* THE FROG?

Grounding in contextual representations by neural language models

Mehdi Ghanimifard



DEPARTMENT OF PHILOSOPHY, LINGUISTICS AND THEORY OF SCIENCE

UNIVERSITY OF GOTHENBURG

When we describe the location of objects in an image, we relate them by their physical location and by the nature of their interaction. This thesis examines how artificial neural networks learn what information is relevant to spatial descriptions. Favouring "the frog is outside the pond" rather than "the pond is outside the frog" is possible by considering the knowledge about the world and human interactions in language models. The findings of this thesis benefit the design of systems that automatically generate image descriptions and search engines and lead to a more natural human-robot interaction.

ISBN 978-91-7833-916-7 (PRINT)
ISBN 978-91-7833-917-4 (PDF)

Why the pond is not *outside* the frog? Grounding in contextual representations by neural language models

Doctoral Thesis in Computational Linguistics

Why the pond is not *outside* the frog? Grounding in contextual representations by neural language models

Mehdi Ghanimifard

Department of Philosophy, Linguistics and Theory of Science
The Centre for Linguistic Theory and Studies in Probability (CLASP)

UNIVERSITY OF GOTHENBURG
May 2020

Mehdi Ghanimifard
Why the pond is not outside the frog?
Grounding in contextual representations by neural language models
Doctoral Thesis in Computational Linguistics, May 2020
Main supervisor: Simon Dobnik

University of Gothenburg
Department of Philosophy, Linguistics and Theory of Science
Box 100, Gothenburg, Sweden

The cover drawing from 'The Common Frog' by Mivart (1881). Not in copyright, scanned at Harvard University, Museum of Comparative Zoology, Ernst Mayr Library. The cover designed by Boshra Khoshnevis

The research reported in this thesis was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

ISBN: 978-91-7833-916-7 (PRINT)
ISBN: 978-91-7833-917-4 (PDF)

Part I of the publication is also available in full text at:
http://hdl.handle.net/2077/64095

Abstract

In this thesis, to build a multi-modal system for language generation and understanding, we study grounded neural language models. Literature in psychology informs us that spatial cognition involves different aspects of knowledge that include visual perception and human interaction with the world. This makes spatial descriptions a compelling case for the study of how spatial language is grounded in different kinds of knowledge. In seven studies, we investigate *what* and *how* neural language models (NLM) encode spatial knowledge.

In the first study, we explore the traces of functional-geometric distinction of spatial relations in uni-modal NLM. This distinction is essential since the knowledge about object-specific relations is not grounded in the visible situation. Following that, in the second study, we inspect representations of spatial relations in a uni-modal NLM to understand how they capture the concept of space from the corpus. The predictability of grounding spatial relations from contextual embeddings is vital for the evaluation of grounding in multi-modal language models. On the argument for the geometric meaning, in the third study, we inspect the spectrum of bounding box annotations on image descriptions. We show that less geometrically biased spatial relations are more likely to deviate from the norm of their bounding box features. In the fourth study, we try to evaluate the degree of grounding in language and vision with adaptive attention. In the fifth study, we use adaptive attention to understand if and how additional bounding box geometric information could improve the generation of relational image descriptions. In the sixth study, we ask if the language model has an ability of systematic generalisation to learn the grounding on the unseen composition of representations. Then in the seventh study, we show the potentials in using uni-modal knowledge for detecting metaphors in adjective-nouns compositions.

The primary argument of the thesis is built on the fact that spatial expressions in natural language are not always grounded in direct interpretations of the locations. We argue that distributional knowledge from corpora of language use and their association with visual features constitute grounding with neural language models. Therefore, in a joint model of vision and language, the neural language model provides spatial knowledge that is contextualising the visual representations about locations.

Sammanfattning (Abstract)

I denna avhandling, för att bygga ett multimodalt system för språkgenerering och förståelse, studerar vi förankrade neurala språkmodeller. Litteratur i psykologi informerar oss om att rumslig kognition involverar olika aspekter av kunskap som inkluderar visuell uppfattning och mänsklig interaktion med världen. Detta gör att rumsliga beskrivningar är ett bra fall för att studera hur rumsligt språk är förankrat i olika typer av kunskap. I sju studier undersöker vi *hur* neurala språkmodeller (NLM) kodar rumslig kunskap, och *vad* de kodar.

I den första studien undersöker vi spåren av den funktionella-geometriska distinktion av rumsliga relationer i unimodala NLM. Denna distinktion är väsentlig eftersom kunskapen om objektspecifika relationer inte är baserad i den synliga situationen. Därefter, i den andra studien, inspekterar vi representationer av rumsliga relationer i unimodala NLM för att förstå hur de representerar begreppet rymd från en korpus. Förutsägbarheten av grundläggande rumsliga relationer från kontextuella representationer är avgörande för utvärderingen av förankring i multimodala språkmodeller. I den tredje studien undersöker vi argument för den geometriska betydelsen genom att inspektera spektrumet av avgränsningsruteannoteringar för bildbeskrivningar. Vi visar att geometriska relationer med en mindre grad av rumslighet är mer benägna att avvika från normen av avgränsningsfunktionens särdrag. I den fjärde studien försöker vi utvärdera graden av förankring i språk och syn med adaptiv uppmärksamhet. I den femte studien använder vi adaptiv uppmärksamhet för att förstå om och hur ytterligare geometrisk information om avgränsningsrutorna kan förbättra generationen av relationella bildbeskrivningar. I den sjätte studien frågar vi om språkmodeller har en systematisk generaliseringsförmåga att lära sig förankring av osedda sammansättningen av representationer. Sedan i den sjunde studien visar vi att unimodal kunskap har potential för att upptäcka metaforer i adjektiv-substantivkompositioner.

Avhandlingens huvudargument bygger på det faktum att rumsliga uttryck i naturligt språk inte alltid är baserade på direkta tolkningar av platser. Vi hävdar att distributionell kunskap från korpusar om språkbruk och deras associering med visuella funktioner utgör förankring för neurala språkmodeller. Därmed, en modell som använder både visuell information och språk,

tillhandahåller neurala språkmodeller rumslig kunskap som kontextualiserar visuella representationer av platser.

List of Publications

Study 1 Simon Dobnik, Mehdi Ghanimifard and John Kelleher. Exploring the Functional and Geometric Bias of Spatial Relations Using Neural Language Models. *In Proceedings of the First International Workshop on Spatial Language Understanding, pp. 1-11. 2018.*

Study 2 Mehdi Ghanimifard and Simon Dobnik. *What* a neural language model tells us about spatial relations. *In Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP), pp. 71-81. 2019.*

Study 3 Simon Dobnik and Mehdi Ghanimifard. Spatial descriptions on a functional-geometric spectrum: the location of objects. *Preprint - under review, 2020.*

Study 4 Mehdi Ghanimifard and Simon Dobnik. Knowing When to Look for What and Where: Evaluating Generation of Spatial Descriptions with Adaptive Attention. *The European Conference on Computer Vision (ECCV) Workshops, pp. 153-161. Springer, Cham, 2018.*

Study 5 Mehdi Ghanimifard and Simon Dobnik. What Goes Into A Word: Generating Image Descriptions With Top-Down Spatial Knowledge *In INLG 2019-12th International Conference on Natural Language Generation-Long papers. 2019.*

Study 6 Mehdi Ghanimifard and Simon Dobnik. Learning to Compose Spatial Relations with Grounded Neural Language Models. *In IWCS 2017-12th International Conference on Computational Semantics-Long papers. 2017.*

Study 7 Yuri Bizzoni, Stergios Chatzikyriakidis and Mehdi Ghanimifard. "Deep" Learning: Detecting Metaphoricity in Adjective-Noun Pairs. *In Proceedings of the Workshop on Stylistic Variation, pp. 43-52. 2017.*

Acknowledgement

This collection of articles was a result of a long journey that started in 2015 when I joined the PhD program at CLASP. Producing such a piece is not a one-person job. I was fortunate to have the support of my supervisors, colleagues, friends and family along the way.

First of all, I would like to recognize the invaluable assistance of my supervisors. I want to express my most profound appreciation to Simon Dobnik, my primary advisor, for his steady support of my trembling steps on this path. A large body of this work became possible with his direct guidance. I wish to acknowledge the efforts of Staffan Larsson, my second advisor, whose feedback helped me at various times during this project. I am incredibly grateful to John Kelleher, my third advisor, our long passionate conversations shaped many ideas presented here. In addition to the help I received from supervisors, I was lucky for being able to seek advice from other fantastic people. I am indebted to Richard Johansson for his counselling which led me to apply for the PhD program. I had the privilege to receive his support since I was a master student. I am also grateful to Stergios Chatzikyriakidis for his encouraging feedback. I have always been able to rely on his friendly advice. Also, I want to thank Lena Eriksson for her counselling. Her thoughtful advice proved to have a significant influence on the success of my studies. I wish to thank, Desmond Elliott, my thesis assessor, for his constructive criticism and helpful feedback on an earlier version of the thesis.

I want to extend my sincere thanks to friends and colleagues for building such an excellent social and research environment. Shalom Lappin, for his unparalleled support and care; Robin Cooper, for sharing extensive knowledge in our bi-weekly meetings at the language and perception reading

group; Jean-Philippe Bernardy, for his constructive advice from time to time; Yuri Bizzoni, for being a colleague and a friend with a significant positive atmosphere around him. Also, thanks to Johan Gross and Julia Forsberg, with Yuri, we had a relaxed PhD life at the office! Same regards to Adam Ek and Vidya Somashekarappa my officemates for almost a year. Although we only shared office for a short period, we made unforgettable memories. I literally received a collection of life lessons from Vidya and Adam after their thoughtful prank! I also wish to thank all the incredible people who made Dicksonsgatan 4 like a family house for me. Haris Themistocleous, Chris Howes, Ellen Breitholtz, Sharid Loáiciga, Asad Sayeed, Sandro Maskharashvili, Rasmus Blanck, Wafia Adouane, Vlad Maraev, Bill Noble, Sylvie Saget, Kathrein Abu Kwaik, Susanna Myyry, Iines Turunen, Aram Karimi. We had a lot of meaningful spontaneous conversations in the kitchen, corridors and the meeting room. I learned something memorable from each of you. I won't miss a chance if I get to work with you again.

The extent of the fantastic colleagues that I wish to thank them for their company goes beyond the circle of CLASP, at GU and Chalmers. Conversation with these friendly faces kept me on the track! I want to thank Mio Cibulka, Leila El-Alti, Richard Endörfer, David Alfter, Luis Nieto Piña, Ildikó Pilán, Prasanth Kolachina, Herbert Lange.

I cannot begin to express my thanks to my family and friends whose support cannot be overestimated. These extraordinary people kept me sane during my PhD life. In Sweden, I had incredible times with Alireza Pazirandeh, Ala Pazirandeh, Niklas Arvidsson and My Högblom. I wish to thank my sisters, Marzieh and Mona, and my mother and father, Shokoufeh and Siamack, for profound belief in my work and my abilities.

Lastly, I would like to thank my wife, Boshra Khoshnevis (without forgetting the help of Biloo and Ghoughies). She is the reason that kept me going on. The completion of my dissertation without her support would not be possible.

Contents

Part I

Thesis Frame

1 Introduction

The success of a class of machine learning algorithms in a wide variety of computer vision tasks led to the emergence of the deep learning paradigm over the past decade. This transformation in computer vision has affected the design and application of perceptual systems in other domains, such as human-robot interactions and computational linguistics. Designing a system based on a data-driven module, such as a neural network classifier, would be challenging without a proper explanation of how the system would reason in situated environments. The potential applications of deep neural networks and questions about different models and how to apply them initiated and motivated this research. In this chapter, we describe the aims of the thesis, the research questions and the objectives. Then, we briefly address the contributions and findings of the studies within this thesis. In the next chapters, we provide more detailed background information.

1.1 Aims

The main objective of this thesis is to pave the way for the development of language generation and language understanding systems with deep neural networks that are grounded in both linguistic and visual inputs. Several use cases of such systems involve spatial language, including automatically describing the location of objects or locating objects based on a description of their whereabouts. These use cases have guided this research to focus on spatial language as an original puzzle in the domain of computational linguistics.

Our goal of building such systems yielded several challenges and fundamental questions about learning representations using deep neural networks. Therefore, the primary aims of the thesis are to assess and investigate the application of neural networks in automatically learning multi-modal representations and the grounding of linguistic units in such representations.

1.2 Research questions

Q1: What spatial knowledge is learned in neural language models? A fundamental question about the representation of modalities is whether these modalities are sufficient for solving the relevant tasks. Language modelling yields several other questions about the encoding of knowledge. Is the knowledge encoded in these modalities extractable using specific learning algorithms and models? In the context of spatial language, is there appropriate geometric knowledge in representations that can be learned with neural models (such as convolutional neural networks)? Is the geometric knowledge in relevant modalities enough to understand or generate a spatial description of a situation?

Q2 How are spatial descriptions associated with spatial knowledge? The measure of success in a data-driven inference cannot be just the accuracy of correct predictions. To have a valid data-driven procedure, there must be an explanation of which evidence is used to conclude the decision. Therefore, we ask the following questions. How do different modalities contribute to a grounded neural language model? How much do visual inputs affect the results in comparison to language inputs? What is the balance between the contributions of different modalities? Why do different modalities contribute differently to the performance of the model?

Q3 Is there a systematic generalisation? The assessment of learning with neural networks cannot be limited to the ability of the system to perform specific tasks in specific examples. This question is often asked in the form of learned compositionality. We investigate the generalisability of learning outcomes by looking at learned representations. To what degree are the expected structures, compositions and relations in linguistic and visual-cognitive representations learned with neural networks? If unexpected relations within neural representations emerge from learning distant tasks, such as language modelling, we need to show if the learned representations imposed from the data are transferable knowledge or if the relations are intrinsic to the neural structures and can be generalised. If there are intrinsic neural structures that capture the compositional relations in neural representations, we ask if these structures are aligned with expected structures. For example, if a two-arity relation is not symmetric, is the order

of arguments preserved in learned representations? Is the order in neural structures aligned with the expected order?

1.3 Contributions and findings

Our findings are distributed over seven studies. In the first three studies (Dobnik et al., 2018; Ghanimifard and Dobnik, 2019a; Dobnik and Ghanimifard, 2020), we inspect unimodal neural language models and the geometric features of bounding boxes to see what latent knowledge on spatial relations is encoded in the models and data **(Q1)**.

Then, in studies 4 and 5 (Ghanimifard and Dobnik, 2018, 2019b), we examine the contributions of visual features and contextual embeddings from a language model in generating image descriptions. We show that the contributions of visual features in producing spatial relations are lower than nominal parts of speech. We also show that top-down localisation has the highest contribution to performance compared to any other top-down feature representation **(Q2)**. In study 5, the question about top-down representations in neural networks resolves whether the intrinsic structures in neural networks affect learning **(Q3)**.

The last two studies Ghanimifard and Dobnik (2017); Bizzoni et al. (2017), focus on the capability of neural language models to learn compositional knowledge. We show that, when using neural networks as the composition function, the models can generalise from limited samples of language use to new word compositions **(Q3)**. In study 6, we found that composing unseen word sequences and decomposing unseen single word representations are possible if the neural language model is trained with enough coverage of vocabulary. However, this is dependent on combinatorial properties and the complexity of compositional meaning **(Q2)**. In study 7, we found that distributional knowledge from neural language models, abstracted from visual/sensory grounding, can recognise metaphors. This finding is consistent with our findings in study 2 —that unimodal language models can capture spatial knowledge in distributional patterns **(Q1)**.

1.4 The thesis frame

This dissertation consists of two parts, the thesis frame and the studies. The first part comprises a synthesis of ideas, background case studies, concepts, a summary of studies and the conclusions of the thesis (chapters 1 to 5). The second part comprises the seven articles of the research project (chapters 6 to 12).

In chapter 2, we briefly explore the background of spatial language in image descriptions. In chapter 3, we extensively discuss the concept of modelling meaning representations with neural networks. chapter 4 summarises the seven studies in connection with the research questions and their findings. Finally, chapter 5 provides the final summary and discussion of the thesis findings and its connection to the published studies.

2 Spatial Expressions In Image Descriptions

The importance of spatial language is embedded in fundamental domain questions about parsing visual sensory inputs into meaningful representations. In its purest form, the individuation of objects would be a spatial cognition task; to recognise where objects begin and where they end in space, how their parts are spatially composed and how the spatial properties (including shapes and locations) interact with each other. In chapter 1, we stated that the use cases of grounded language generation and understanding brought us to the study of spatial language. In a large body of work on referring expression generation (Dale and Reiter, 1995; Krahmer and van Deemter, 2011) with applications that describe scenes (Viethen and Dale, 2008) and images (Mitchell et al., 2012; Elliott and Keller, 2013), describe visible objects (Mitchell et al., 2013) and engage in human-robot dialogue (Kelleher and Kruijff, 2006; Dobnik, 2009), spatial grounding language is inseparable from the tasks. In this chapter, we discuss the relevant topics and concepts regarding how human language can describe the space, the location of objects and the relation between them in an image.

2.1 Spatial expressions

There are many ways to describe a situation and convey information about the location of important matters. The most direct form of denoting the location of objects in natural language is to use *locative expressions*. The simplest type of locative expression is composed of three constituents —a locative preposition and two noun phrases. One of the noun phrases is the subject of the preposition and the other is the object. For example, in 'the frog next to the pond' the preposition 'next to' describes the location of the subject 'the frog' with respect to the object 'the pond'. The subject/object are also known by different names, such as *referent/relatum* (Miller and Johnson-Laird, 1976), *figure/ground* (Talmy, 1983) or *located object/reference*

The *frog next to* the *pond*.

The *frog* is *next to* the *pond*.

There is a *frog next to* the *pond*.

The *frog next to* the *pond* is watching us.

[a]The drawing from *Grimm's Fairy Tales* by Grimm and Andre (1899)

Figure. 2.1: Describing the location of target and landmark —⟨ *frog*,*next to*,*pond*⟩.

object (Herskovits, 1986; Gapp, 1994a; Dobnik, 2009). In this work, we refer to them as TARGET/LANDMARK. When referring to a situation with the structure ⟨TARGET, RELATION, LANDMARK⟩, the expression may be combined with a copulative verb, an existential quantifier or other additional information (Figure 2.1).

In English, there are a small class of words with meanings that denote spatial relations between targets and landmarks. This includes simple words (*on*, *in*, *over*, *under*) and compound phrases (*on top of*, *to the left of*, *to the right of*, *in front of*, and etc.) Some of these relations are compositional, which means they can be combined to produce new relations (*above* and *far from*). Based on the list of prepositions in (Landau and Jackendoff, 1993) and alternative compositional and compound relations discussed in (Herskovits, 1986, p. 156), we created a dictionary of 75 spatial relations. Considering their alternative forms, with a minor difference in their spatial sense, they constitute 1,194 entries[1] (Table 2.1). For example, '*to the left of*' could be one form of several possible alternative multi-words with close spatial meaning:

- {*at*/*on*/*in*/*to*/*by*} *the left* {{*hand*} *side*} *of* → *to the left of*.

[1]The source code to generate the collection of multi-words expressions is available in the online repository of published studies including `https://github.com/GU-CLASP/functional-geometric-lm`.

about	above	across	after	afterward
against	ago	along	alongside	amid
among	apart	around	at	away
back	backward	before	behind	below
beneath	beside	between	bottom	by
down	downstairs	downward	during	east
from	front	here	in	inside
into	inward	left	near	nearby
next	north	off	on	onto
out	outside	outward	over	parallel
perpendicular	right	side	sideways	since
south	there	through	throughout	to
together	top	toward	under	underneath
until	up	upon	upstairs	upward
via	west	with	within	without

Table. 2.1: The vocabulary of 75 spatial relations from Landau and Jackendoff (1993) and Herskovits (1986).

2.2 Functional/geometric meaning

Expressing location and describing space is not limited to the prepositional relations in spatial expressions. Other verbs in referring expressions can indicate the relative location between subject and object. With different degrees, these relations might have a strong or weak association with the location of the subject and object. For example, '*ride*' entails a specific spatial configuration between the subject and the object, depending on their shape. Other relations, such as '*touch*', '*sit on*', and '*jump over*' indicate a specific spatial configuration of subject and object. Nevertheless, their direct meaning is not just the location of objects; it indicates other kinds of relations, which consequently entail specific spatial arrangements.

In the same way, the meaning of spatial prepositions is not purely geometrical; it entails other relations and associations between subject and object that are functional. The relation '*over*' does not just describe a geometric location; it also indicates a function —the subject provides protection or shelter for the object/s. The functional sense of relations includes specific interactive relations between entities that are not dependent on the location and spatial configurations.

A simple representation of the geometric sense of relations is based on the acceptability ratings of individual locations with respect to the landmark. Logan and Sadler (1996) suggest that the mental representation of a geometric meaning could be a template projection of locations on a map, where the landmark is in the middle and each location has its degree of acceptability for the target object. A study on location acceptabilities for different prepositions shows that each spatial preposition has a different degree of dependency on object-specific relations (Coventry et al., 2001). The meaning of each relation is an interplay between the functional and geometric relations of two objects. For instance, *'above'* has both geometric locational meaning and functional sense. When it is used in different context it can have different degree of functional and geometric acceptability. Spatial relations in natural language have a spectrum of geometricity, with different degree of favouring geometric bias or functional bias. Another way to study the object-specific sense of relations is to consider the distributional dependency between relations and objects in image descriptions (Dobnik and Kelleher, 2013, 2014). In our studies, we consider these aspects of by examining language models.

2.3 Image descriptions

In a simple *show-and-tell* task, when provided an image, the agent must generate a description of the image. Since the early works on human-robot interactions, this task became the centre of interests for natural language grounding (Roy, 2002). In recent years, several large datasets have been developed, in which crowd-sourced human annotators describe images from freely available datasets of photographed scenes over the Internet.

Datasets Common datasets of image caption tasks, such as MSCOCO (Lin et al., 2014) with more than 300,000 images and Flickr30k (Young et al., 2014) with 30,000 images, each provide five alternative descriptions per image. However, the variation and number of geometric spatial relations in the dataset is limited; '*to the left of*' and '*to the right of*' are rarely used in the dataset. On the other hand, the Visual Genome (Krishna et al., 2017) provides 50 region descriptions and triplet annotations per image, for a total of over 108,000 images. The annotation schema, in this case, was slightly different from captioning, as it asks annotators to describe specific parts of

the images or the relation between two object areas in the images. In this dataset, description and relation annotations are associated with relevant bounding boxes in the image.

Grounding spatial descriptions Both generating and understanding a spatial description with three components —TARGET, RELATION, LANDMARK —requires several types of knowledge: (1) object identification, (2) comprehension of geometric configuration, (3) capturing object-specific relations between objects and (4) a frame of reference for projective relations '*to the right of*' and '*below*'. When people describe image contents, they commonly use spatial expressions. A scene can be described correctly using any spatial relation fitting the same objects depending on the intent of the speaker. However, the image description task may use the knowledge about the scene in a specific way.Precisely, object identification and the capture of object-specific relations in the picture might be enough to describe an image with spatial expressions.

2.4 Summary

In this chapter, we described the concept of locative expressions and its connection with image descriptions. Spatial relations denote locations in scenes. However, their meaning, to some extent, is also dependent on object-specific relations. In our studies on grounding spatial descriptions, we will use datasets of images with descriptions, including MSCOCO(Lin et al., 2014), Flickr30k(Young et al., 2014) and Visual Genome (Krishna et al., 2017).

3 Modelling And Meaning Representation

In this chapter, we address the concepts and the theoretical framework for modelling meaning with uncertainty and grounding natural language in multi-modal representations.

3.1 Terminology of modelling

Whenever we need to make a systematic prediction based on observations and evidence, we use a set of assumptions. A *model* is the encoding of these assumptions in the form of a function. This function takes given evidence as input and produces the prediction as an output. Formally, modelling y based on x with the function f would be as follows:

$$y = f_\Theta(x) \tag{3.1}$$

where f_Θ is called the *model function*, which is parametrised with Θ.

The parameters are part of the assumption about the model function. For example, by assuming the rules of physics, the position of a falling object at a specific time in the future can be modelled given the current evident position of the object. However, the formula of the location still requires the important parameters of velocity and acceleration of the object in the model. More often, determined prediction of an outcome is not enough. We need to associate each prediction with an uncertainty measure. Such a model is a *probabilistic model*. Instead of modelling the predictable outcome, a measure of uncertainty for any possible outcome is modelled; the density of possible outcomes is conditioned with the observable evidence.

$$\Pr(Y = y|X = x) = f_\Theta(y, x) \tag{3.2}$$

where f_Θ is the model function, which assigns a degree of uncertainty for predicting y grounded on an observable x. To simplify the probability annotations, we do not write the complete propositions $(Y = y)$; instead, we use shorthand —$\Pr(y|x)$. Commonly, the probability distributions used in this work are categorical, in which Y is a bounded discrete vector of the items. Therefore, the common implementation of function f_Θ is conducted using a module with a vectorised output the same size as Y:

$$f_\Theta(y, x) = \texttt{modules}(x)[y] \tag{3.3}$$

where y is a category in distribution, the output of $\texttt{modules}(.)$ is a vector with the size of all possible categories and the square bracket annotation $[y]$ indicates a lookup operation to select the value for y-key.

Representing assumptions about the future in a function requires a *framework of modelling* to acquire the model function. A constructive *proof*, a search algorithm over a class of functions or its parameters, is the path to building the model function from these assumptions. When a set of data points drives the search algorithm, the process of fitting a function according to these data points is called *learning* or *training* the model.

3.2 Modelling in deep neural networks

In this work, we study the framework of artificial neural networks to encode and build the model function. *Deep learning* (DL) and *artificial neural networks* (ANN) refer to a modelling framework in which a composition of differentiable functions form the model function. The learning occurs through parameters of the function with the *backpropagation algorithm*. The backpropagation algorithm is a data-driven optimisation algorithm that uses a measure of error loss over the training data to gradually update the model parameters toward lower error. The differentiability of the loss enables this method to apply the chain rule of derivatives to aim the parameter updates toward reducing errors for the training data. In a nutshell, the critical assumptions needed to build a deep learning model are the assumed model function (the composition of modules or the neural network architecture), the assumption that the training dataset has relevant knowledge for the task and the assumption about the error function (the loss function).

There are several learning paradigms, such as supervised and unsupervised learning. These distinctions mainly concern the difference in annotation on the training data, the error function and how they are related to the predictable variable of the model. The most common paradigm is supervised learning, in which the training dataset is a set of annotated inputs and the predicted output of the model $D = \{(y_i, x_i)\}$. Unsupervised learning, on the other hand, uses an unlabelled training dataset $D = \{x_i\}$. The error function in these cases provides additional assumptions about how data points are internally connected. Any internal data structures that indirectly relate to the predictable outcome of the model could be the basis for an error function in unsupervised learning, such as unsupervised clustering of data points for a classification task without supervised data.

The most common loss function for deep learning models is the surprisal of the training data. The surprisal of a random variable $(X = x)$ is defined as: $I_X(x) = -\log(\Pr(X = x))$. With a given dataset, such as $D = \{(y_i, x_i)\}$ and a model function f, an ideal search algorithm for finds the best fitting parameters that minimise the loss:

$$J_\Theta(f, D) = \sum_{(y_i, x_i) \in D} -\log(f_\Theta(y_i, x_i)) \tag{3.4}$$

$$\Theta = \underset{\Theta}{\operatorname{argmin}} J_\Theta(f, D) \tag{3.5}$$

while the search algorithm looks for the best fitting parameters Θ, there is usually more than one answer or there is no converging path to an acceptable error level with backpropagation. In the most straightforward form —backpropagation in a *gradient descent algorithm* —the gradient of the error function with a pre-defined learning rate updates all parameters iteratively until the error converges to an acceptable threshold. To overcome technical difficulties in processing large datasets and parameter space, other variations of this algorithm may process data in batches, using *stochastic gradient descent* and the momentum of past updates. For simplicity, each step of mini-batch training can be formulated as an updating operation for Θ parameters as follows:

$$\Theta = \Theta - \eta \cdot \nabla_\Theta J_\Theta(f; D_{batch})$$

where η is a hyper-parameter for the learning rate and ∇_Θ, is a notation for a stochastic deferential operation over every parameter in Θ.

The concept of indirect learning from a function different from the goal prediction of the model is also related to the concept of *multi-task learning* and *transfer learning* in neural networks (Goodfellow et al., 2016, Chapter 15). In summary, the parameters of a model function learned from a different dataset or a different goal or a different task encode relevant assumptions needed to make our intended prediction model. Therefore, the pre-trained modules can be reused or composed into the model function in the neural network. The final training steps with much less training data is then known as fine-tuning or in some context referred to as domain adaptation phase.

3.3 Grounding in representations

So far, we have identified that modelling is a way to represent assumptions about the world in the form of the parameters of a predicting function and that DL models encode assumptions inferred from data as a representation space. The goal of understanding and comprehension is to connect two types of representations —sensory representations and abstract concepts. A model that provides the link between two representations is a model of grounded meaning. A probabilistic model of representations can potentially formulate such links. However, the question remains about the generalisability of the learning (see section 3.4).

By definition, the model represents the uncertainty of connecting observables to their representation; therefore, it can be used as a model of *grounded representations*. Within the paradigms of DL, instead of having a given strict symbolic representation for concepts and their internal associations, these representations must be learned. The architectural design of the neural networks and the training datasets impose restrictions on how these representations are interconnected. In the machine learning community, this has become a field of study called *representation learning* (Bengio et al., 2013).

When modelling conversational agents, linguistic expressions are grounded in internal representations of the agent. There are at least two proba-

bilistic models of meaning for (1) generating and (2) understanding utterances:

$$\text{Speaker model} : \Pr(u|r) = f_{\Theta_s}(u, r) \tag{3.6}$$

$$\text{Listener model} : \Pr(r|u) = f_{\Theta_l}(u, r) \tag{3.7}$$

$\Pr(u|r)$ is the measure of uncertainty in choosing the utterance u, referring to the given representation r. $\Pr(r|u)$ is the measure of uncertainty in interpreting the given utterance u as if it meant representation r for the listener. In other words, the grounded meaning of each natural language utterance is what it denotes in the representation space according to the model. Without any probabilistic sampling in composition of modules in neural networks, there is unambiguous mapping of sensory representations onto grounded representation space. However, the link between grounded representations and natural language utterances is uncertain, with a linking degree of uncertainty on all possible outcomes. The learning process establishes the degree of certainty of the link between utterances and representation space and builds the fitting map between the agent's primitive sensory and motor representations and grounded representation space.

Later, in section 3.5, we provide additional discussion about the link between meaning and representations in the speaker model. In a speaker model, to express what is in an image, the sensory data for the situation is first mapped onto the representation space ($\Rightarrow v$). Then, the speaker model assigns a measure of goodness to the utterance predictions:

$$\Pr(\text{“there is a frog”}|v) = f_{\Theta_s}([\text{“there”}, \text{“is”}, \text{“a”}, \text{“frog”}], v)$$

In other words, the model of grounded meaning is also a model for connecting utterances to sensory evidence. The internal representations are not directly connected to external references. They are interpretations of the sensory readings internal to the agent. The mapping function between visual sensory inputs and internal agent representations could be modelled with a pre-trained convolutional network enriched with other contextual information about the situation. The model establishes uncertain links between primitive sensory readings and utterances of natural language, for that reason it functions as a model of grounding. In section 3.6, we describe some properties of this representation space and how sensory features would be mapped onto this representation space.

3.4 Modelling compositionality

One of the challenges of a model of grounding is formulated as Harnad's symbol grounding problem (Harnad, 1990). The argument is that the capacity of learning from limited data is problematic when it is expected to impose new links to potentially unlimited compositions in a symbolic representation system. The challenge is to infer grounded meanings for new representations (e.g. 'ZEBRA') from known bottom-up representations learned from images (e.g. 'HORSE' and 'STRIPES') when a symbolic link between them ('ZEBRA' and a composition of two others) is established in natural language ('ZEBRA' = 'HORSE' + 'STRIPES'). In other words, compositionality as an ability to construct new representations linked to both sensory and abstract linguistic representations creates a generalisation problem for bottom-up learning. The underlying premise of Harnad's formulation of the problem is that human language is symbolic; therefore, when human behaviour shows the capability of learning from language input, the establishment of such links came from new symbolic rules imposed by new statements of the natural language.

Without this explicit premise about the nature of language, the recent literature of language modelling stretches the notion of compositionality. In one account, any function in a semantic vector space is a model of compositionality (Mitchell and Lapata, 2010). In another direction, the semantic parse trees of linguistic expressions are used for composing neural network modules (Andreas et al., 2016).

The notion of compositionality in natural language that we use in this thesis is simply the extent of generalisation in bottom-up training. The grounding of known representations is the learned link between natural language utterances and their internal neural representations. The representation space imposes the compositional and structural links ('ZEBRA' = 'HORSE' + 'STRIPES'), which are either learnable without intrinsic structures in space from data or learnable with extended structural or top-down control or design of the intrinsic properties' representation units.

3.5 Generative language model

When the target of predictions in Equation 3.2 is a linguistic unit, the model is what we call a *probabilistic language model* —predicting the next word given previous sequences of words.

$$\Pr(w_{t+1}|w_{1:t}) = f_\Theta(w_{t+1}, w_{1:t}) \tag{3.8}$$

where $w_{1:t}$ represents the given sequence of words at time step t and the random variable is the target token at time $t+1$, here represented with a shorthand annotation for the probability of w_{t+1}. The sampling process from this model can potentially be part of a model designed to generate sequences of any length:

$$\Pr(w_{1:T}) = \prod_{t=1}^{T-1} \Pr(w_{t+1}|w_{1:t}) \tag{3.9}$$

when coupled with a search algorithm, such as beam search, can be used for language generation. When the model function is based on recurrent neural networks, we call it a *recurrent generative language model*, shortened to a *recurrent language model*:

$$\Pr(w_{t+1}|w_{1:t}) = f_\Theta(w_{t+1}, h_t) \tag{3.10}$$

$$h_t = \texttt{rnn}_{\theta_1}(h_{t-1}, w_t) \tag{3.11}$$

$$f_\Theta(w_{t+1}, h_t) = \texttt{softmax}(g_{\theta_2}(h_t))[w_{t+1}] \tag{3.12}$$

where h_t represents the recurrent state at time t. This could also be interpreted as an agent representation in Equation 3.6 for generating each token. Two important modules of the language model are $\texttt{rnn}_{\theta_1}$, the recurrent module, and g, the top module, often a multi-layer perceptron, the output of which is a vector with the size of the vocabulary. In the end, `softmax` is the final activation function over all possible tokens in the vocabulary:

$$\texttt{softmax}(V) = [\frac{e^x}{\sum_{x' \in V} e^{x'}}]_{x \in V} \tag{3.13}$$

where V represents a vector of vocabulary size. After activation with `softmax`, the output resembles a categorical probability distribution of the vocabulary. The notation $\texttt{softmax}(\cdot)[w]$ represents the predicted probability

for w at the output vector. The unfolded representation of the model function would be as follows:

$$\Pr(w_{t+1}|w_{1:t}) = f_\Theta(w_{t+1}, \texttt{rnn}_{\theta_1}(\texttt{rnn}_{\theta_1}(...\texttt{rnn}_{\theta_1}(h_0, w_1)..., w_{t-1}), w_t))$$

The parameter set of the model, Θ, is comprised of two partitions, θ_1, θ_2, from the two main modules of the model. The most common recurrent neural network we will use in this thesis is long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997). We often add a trainable embedding layer in addition to the recurrent neural network to learn token representations. When a generated utterance is supposed to describe the content of an image or other situation, the generative language model in Equation 3.8 and 3.9 could be written as a conditional probability similar to the speaker model in Equation 3.6:

$$\Pr(w_{t+1}|w_{1:t}, c) = f_\Theta(w_{t+1}, h_t, c) \tag{3.14}$$

$$\Pr(w_{1:T}|c) = \prod_{t=1}^{T-1} \Pr(w_{t+1}|w_{1:t}, c) \tag{3.15}$$

where c represents the encoding of the situation —visual features and the fusion of two representations, h_t. c is the agent's grounded representation in the speaker model in Equation 3.6.

In these models, the uncertainty measures of the language model could also be interpreted as the degree of acceptability of linguistic units. With the speaker rationality assumption, the distribution of utterances in training data should not be very different from the acceptability judgment rankings. The generative language model learned from this data is also a model of acceptability judgments. If the language model can accurately predict acceptability judgments, it can be considered a model of syntax. It has been argued that such a language model is also an implementation of syntax without underlying categorical syntactic rules (Lau et al., 2017, Section 3).

On the other hand, predicting the categorical distribution of tokens in their positions is, in fact, a model for the substitutability of words (i.e. Equation 3.14). Such a model loosely simulates substitutability of tokens. Therefore, the vector representations learned for tokens and words in these models loosely posses the attributes for lexical-semantic representations. This notion

of meaning representation is consistent with the grounded representation discussed in section 3.3. Based on these two arguments, a neural language model must be able to encode knowledge about syntax and semantics in the form of the neural network modules' structure and parameterised representations. To predict the model outputs, parameters such as embeddings and intermediate representations of modules (contextualised embeddings) encode relevant knowledge learned from the training data.

3.6 Modelling convolutional neural networks

In section 3.3, we mentioned the possibility of using a function such as convolutional networks to map visual inputs onto a representation space for language grounding. Here we discuss two aspects of using convolutional neural networks as a feature extraction function:

(1) How do convolutional neural networks process images?
(2) What types of knowledge are encoded in convolutional representations?

The role of convolutional networks (ConvNets) as a mapping function is to take basic two-dimensional pixel representation of images from a colour feature space, then project it onto another feature space that can discriminate images based on their content. The most basic form of visual understanding is to recognise objects and entities in images. A set of features that can distinguish visual differences between objects would be enough for most tasks. For this reason, the most common way to use ConvNets in a variety of visual processing tasks is to train it as a module in an object recognition model, then use it as feature extraction module in other tasks and models. The success of ConvNets in an object recognition task (Krizhevsky et al., 2012) with a large ImageNet dataset (Deng et al., 2009) was the landmark deep learning success in computer vision.

Conceptually, an object recognition model has the following modular design:

$$\Pr(category|I) = f_\Theta(category, I) \tag{3.16}$$

$$\mathbb{V}_I = \texttt{ConvNet}_{\theta_1}(I) \tag{3.17}$$

$$f_\Theta(category, I) = \texttt{softmax}(\texttt{mlp}_{\theta_2}(\texttt{flatten}(\mathbb{V}_I)))[category] \tag{3.18}$$

where f is the uncertainty model for recognising a category of the object given the image I. The model has two modules —$\texttt{ConvNet}_{\theta_1}$, the convolutional network for feature extraction and $\texttt{mlp}_{\theta_2}$, the multi-layer perceptron for classification based on convoluted features. The most important property of ConvNets is that it can transfer object recognition knowledge to find patterns of local structures (LeCun et al., 2010).

For recognising objects, we need a feature representation that would be, to some extent, invariant to spatial transformations. Geometric transformations such as shifting, rotating and rescaling have limited effects on recognising an object. Notable image representations such as the scale-invariant feature transform (SIFT) (Lowe, 2004) and the histogram of oriented gradients (HOG) (Dalal and Triggs, 2005) were motivated by this requirement. However, object recognition is not strictly invariant to geometric transformations. Spatial compositions at a global level can change the interpretation of smaller patterns. For these reasons, ConvNets was designed with the prior knowledge that identifying local patterns in different scales is essential. Then, for all possible regions of the image, as pre-defined granularities, their receptive fields would be mapped onto a new feature space. The feature mapper (or the kernel function) with input as small as 3x3 pixels interprets local regions into a new representation space. The term *convolutional network* refers to multiple stages of feature mapping followed by spatial sub-sampling, which finally produces a representation with a coarser space but richer representations. After stacking the modules of feature mapping with sub-sampling, each broader region is mapped to a vector representation.

Although prior knowledge about the task led to its design and the popularity of ConvNets in several tasks, there is limited theoretical understanding about how and what geometric features are encoded in convoluted representations. Based on one account, these features are useful for localisation and object detection tasks, without an algorithmic search (Lenc and Vedaldi, 2015a; Ren et al., 2015). In another account, the recognition tasks are sensitive to local geometric transformations[1] and ConvNets relax the geometric knowledge; therefore, geometric relationships between local parts in lower layers decay when reaching the higher layers (Hinton et al., 2011; Lenc and Vedaldi, 2015b; Kelleher and Dobnik, 2017).

[1] The recognition of a face depends on the spatial relationships between eyes and nose (Hinton et al., 2011)

Based on these two accounts:

(1) The geometric features in convolutional representations are a continuum of relational features among smaller regions and larger superpixels;
(2) These are locally relaxed at the final layers to the extent that convolutional representation may have lost its geometric knowledge. This is an important consideration when we want to ground spatial relations in natural language on these visual representations.

3.7 Conclusion

We explained that any modelling requires a set of assumptions; modelling with neural networks encodes these assumptions into the model architecture, training datasets and objectives. We addressed the neural network modelling for grounded representations, compositionality and language generation. All these models correspond to challenges in meaning representations. In the next chapter, we summarise our studies on spatial knowledge encoded in the language models.

Summary Of Studies

4

In this chapter, we summarise the questions, methods and findings of six studies and discuss their relevance to the main aims of the thesis.

4.1 Study 1: Functional/geometric bias in neural language models

Simon Dobnik, Mehdi Ghanimifard, and John Kelleher. Exploring the Functional and Geometric Bias of Spatial Relations Using Neural Language Models. *In Proceedings of the First International Workshop on Spatial Language Understanding, pp. 1-11. 2018.*

Understanding spatial language is fundamental to human-robot interactions. The meaning of spatial relations in scene descriptions is grounded in the geometry of the scene and the functional relationship between objects. We used a neural language model on a large corpus of image descriptions to investigate the earlier observations about functional bias in spatial relationships. We contributed to understanding that are encoded in unimodal neural language models.

4.1.1 Questions

Does the performance of trained neural language models on relational descriptions in the Visual Genome dataset (Krishna et al., 2017) account for the expectation that more functional spatial relations are more predictable based on the objects they describe?

We propose two hypotheses:

- Descriptions with functional relations (in contrast with geometric relationships) have lower perplexity in their language model over the held-out test suite (more likely gold-standard descriptions) because they describe a functionally common situation. The target/landmark

object pairs in the dataset are more specific to functional relations compared to geometric relations.

- When modifying spatial relations with any alternatives, the phrases with the initial choice of functional relations gain increased perplexity because they are more contextually dependent on targets and landmarks compared to geometric relations.

4.1.2 Method

We trained the neural language model on image descriptions in Visual Genome. Then, we measured the perplexity of the model on held-out descriptions based on their spatial relations. In our experiments, we examined the hypothesis on both natural occurring descriptions in the dataset and the down-sampled balanced dataset.

4.1.3 Findings and conclusions

We observed from the perplexity of the language model that functionally-biased spatial relationships are more predictable when the model was trained on the dataset with a naturally occurring frequency of descriptions. However, training the model on a down-sampled dataset did not result in the expected outcome of perplexities for each test group.We reported a more detailed examination of sensitivity of the language model. Our observation showed that the degree of sensitivity for target and landmark is not the same in the two groups of spatial relations. A possible explanation for different sensitivity for targets and landmarks is the misalignment between word order, semantic structure of relations and the cognitive process of choosing related objects as landmark and target. Misalignment of word order and the underlying semantic structure of spatial expressions explains why the forward and backward direction language models have different levels of perplexity.

The second category of the hypothesis was only partially confirmed. Only a few spatial relations confirmed the hypothesis. While some geometric relations, such as *'above'* tend to see a high degree of change in perplexity when replaced with other spatial relations, the dependency of geometric relations on the textual context leaves interesting open questions about the world knowledge and spatial knowledge in neural language models.

Author contributions Mehdi Ghanimifard had the main responsibility for implementing the model, conducting the experiments and reporting it. Simon Dobnik and John Kelleher had shared responsibility for the remaining aspects of this research. All authors read and approved the final manuscript.

4.2 Study 2: Representation of spatial relations in neural language models

Mehdi Ghanimifard and Simon Dobnik. *What* a neural language model tells us about spatial relations. *In Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP), pp. 71-81. 2019.*

We followed up the question about possible encoded knowledge in section 4.1 about spatial relations in unimodal neural language models. In this work, we extend the method to inspect the knowledge of spatial relations in generative language models. One of the methods for measuring lexical knowledge in distributional semantics is analogical reasoning tasks. The knowledge of spatial relations in the image description task may be different from the visual-cognitive knowledge required for human judgment on spatial relations, so we also examined if learned representations are transferable to other tasks.

4.2.1 Questions

- What should we expect from a contextualised model of spatial relations based on textual features in terms of their functional and geometric bias?
- How can we inspect these in a generative language model?
- How do the learned representations from the generative model compare with representations from human judgments?

4.2.2 Method

We trained a generative LSTM language model on region descriptions. Then, we inspected how the language model encodes descriptions of spatial rela-

tions in swapped contexts of target and landmark objects in terms of perplexity. We proposed a method in which word-context vectors are produced based on augmented datasets, swapping spatial relations in the context of other spatial relations. Perplexities of these generated word-context examples built perplexity-based vector space for spatial relations.

We ran analogical tests on these vectors and other textual embeddings to inspect how these representations differed from each other. Finally, we compared them with vector representations of human acceptability judgments and relatedness judgments.

4.2.3 Findings and Conclusions

In the absence of the image, we expected contextual representations to learn object-specific knowledge (functional knowledge). However, the learned representations showed high performance in solving analogical tasks that required also some sense of geometry. Our analysis is that functional knowledge must be complementary to geometric knowledge, which is why language models can partly solve these puzzles. These finding were also confirmed with qualitative inspections, for example the representations of *'left'* and *'right'* were similar to each other and different from *'above'* and *'over'*.

The task of judging acceptabilities, the task of generating descriptions and annotations, and the task of finding related words are three different tasks, perhaps using different kinds of spatial knowledge. The last experiment comparing the similarities of representations in the image description task, acceptability judgments and relatedness experiments hinted that spatial relations in the image description task might use different types of knowledge about space and spatial relations.

The findings of this study raise questions that links this study to study four and five. Knowing that textual context provides discriminative features for identifying spatial relations, we argue that, in language generation tasks, one can ground the word choices in textual evidence. Generative neural language models can encode task-specific knowledge of space, including functional and geometric bias, when describing the relation between two objects. With this insight, the consequence of memorised knowledge is an open

question for multimodal language models —how does a multimodal language model balance the attention to knowledge of the scene and linguistic representations in the task?

Author contributions Mehdi Ghanimifard had the main responsibility for implementing the model, conducting the experiments and reporting it. Mehdi Ghanimifard and Simon Dobnik had shared responsibility for the remaining aspects of this research. Both authors read and approved the final manuscript.

4.3 Study 3: Functional/geometric spectrum in bounding boxes

Simon Dobnik and Mehdi Ghanimifard. Spatial descriptions on a functional-geometric spectrum: the location of objects. *Preprint - 2020.*

In the previous two studies, we investigated the distributional properties of spatial relations in language. We argued that functional knowledge and geometric knowledge are encoded in distributional representation, which can be captured to some extent with neural language models. To complement this study on the grounding of spatial relations for image descriptions, we extended the investigation on the distributional properties of bounding boxes for the spatial relations that describe them. More specifically, in this study, we used bounding boxes to extract the basic geometric features of the relations between two objects. Then, we inspected the geometric feature distribution of each spatial relation.

4.3.1 Questions

- Are the bounding box features extracted from annotated images reliable descriptors for spatial relations; do the extracted features correspond to other geometric representations, such as spatial templates?
- Are geometrically biased spatial relationsobtained in constrained experimental settings reflected in more predictable locations of objects? Can they be mapped into fewer variations of their related object locations?

4.3.2 Methods

For each relation, we collected the pairs of bounding boxes from the relationship dataset in the Visual Genome (Krishna et al., 2017). After standardising the bounding boxes, each pair of objects produced several feature vectors $[x, y, d]$. Inspired by the Attentional Vector Sum (AVS) model Regier and Carlson (2001), the bounded boxes were converted to feature vectors that expressed the geometric relations between individual locations of objects. The expected feature vectors for each spatial relation are comparable with spatial templates. Then, we inspected the variations and skewness of the feature vector distributions from their centroid to determine if thisaccounts for geometric bias. The lower the variation, the more geometrically biased is the relation.

4.3.3 Findings and conclusion

We found that the bounding box features represented as the weighted sum vectors from acceptability scores in spatial templates for projective relations. We found that the feature vectors for geometrically biased relations diverge less from the average vectors compared to their more functionally biased equivalent relations. The distribution of feature vector divergence from the average vector is more skewed toward zero when they describe a geometrical relation. We also inspected the properties of some verbal relations with spatial content. These spatial features indicate spatial regularities in the image description dataset. Practically, the findings of this study would be helpful when designing models for image captioning, as it demonstrates the representations that are relevant for different types of descriptions.

Author contributions Mehdi Ghanimifard had the main responsibility for implementing the model, conducting the experiments and reporting it. Simon Dobnik and Mehdi Ghanimifard had shared responsibility for the remaining aspects of this research. Both authors read and approved the final manuscript.

4.4 Study 4: Evaluating generation of spatial descriptions with adaptive attention

Mehdi Ghanimifard and Simon Dobnik. Knowing When to Look for What and Where: Evaluating Generation of Spatial Descriptions with Adaptive Attention. *The European Conference on Computer Vision (ECCV) Workshops, pp. 153-161. Springer, Cham, 2018.*

The neural network model in (Lu et al., 2017) provides an attention mechanism that expands the domain of attention from spatial attention on visual features to hidden states in the language model. In sections 4.1 and 4.2 we explored the possibility of memorising specific spatial knowledge in a unimodal language model, including functional and task-specific spatial relations between objects. In this study, we wanted to determine how a multimodal language model uses it in a language generation task. The attention on linguistic features when knowledge from different sources generating different parts of speech and, more specifically, on spatial relations can explain the grounding of the generation model in multimodal information, including the contextual representations in the language model memory.

4.4.1 Questions

- How does the attention on visual features and linguistic features change for different parts of speech?
- Is there any difference in the magnitude of attention on visual features between targets and in landmarks?
- Are spatial relations grounded in visual features?

4.4.2 Method

The adaptive attention between visual features and linguistic features compete with each other. The magnitude of attention on linguistic features is a sign of dependence on language and context instead of a grounding in visual features. In this study, we inspected adaptive attention as a source of explanation for grounding. The average of attention on linguistic features

for each part of speech and, for each semantic role in spatial descriptions (target, relation, landmark) is interpreted as an indication for a lack of visual grounding. We compared the ranking of attentions with the rankings of accuracy rates of a uni-modal language model predicting a mismatch, whether a part of speech has been replaced in the FOIL captions (Shekhar et al., 2017). We also qualitatively examine the average spatial attentions of descriptions containing each spatial relation; the spatial attention on target, landmark and spatial relation.

4.4.3 Findings and conclusion

The degree of attention on linguistic features varies depending on the part of speech. In particular, we found that the attention on visual features drops when predicting spatial relations compared to the average attention on noun phrases. The average visual attention on parts of speech partially reflected the results from the FOIL task. For example, nouns are highly visually attended but difficult to predict by the language model, the adpositions (prepositions and postpositions) were ranked among the least visually attended parts of speeach while there were moderately predicatable in the FOIL task.

There are three possible explanations for these results:

(1) Spatial relations are more functional and object dependent in these tasks. Therefore, object-specific spatial features (spatial affordances of objects) encoded in language models are more likely to be predictive of spatial relations than visual clues in the image.
(2) CNNs do not have represent geometric locational information required for grounding spatial relations. As they are trained for object identification, there is some degree of spatial invariance in these features.
(3) Using softmax for modelling attention is a disadvantage in cases where spatial attention is distributed over several objects and their relation. Spatial relations depend on target, landmark and locational features; therefore, the softmax model of attention is noisier when it attends to multiple locations for predicting spatial relations.

Author contributions Mehdi Ghanimifard had the main responsibility for implementing the model, conducting the experiments and reporting it. Mehdi Ghanimifard and Simon Dobnik had shared responsibility for the

remaining aspects of this research. Both authors read and approved the final manuscript.

4.5 Study 5: Generating descriptions with top-down spatial knowledge

Mehdi Ghanimifard, and Simon Dobnik. What Goes Into A Word: Generating Image Descriptions With Top-Down Spatial Knowledge. *In Proceedings of the 12th International Conference on Natural Language Generation. 2019.*

As we continued to question how the neural language model learns spatial knowledge, we investigated the effects of top-down knowledge on space in generating relational image descriptions. As seen in section 4.4, the attention mechanism for the generative recurrent language model can control and explain how different modalities contribute to generation tasks. In this study, we integrated specific geometric and non-geometric features that are considered relevant in top-down computational models of spatial descriptions into the design of the attention model. We compared the effects of three types of top-down spatial knowledge:

(1) Where objects are obtained with a separate localisation procedure;
(2) Which object is the target, and which is the landmark, with prior role assignment;
(3) How they are geometrically related in images by representing their spatial configuration.

4.5.1 Questions

(1) Which types of top-down spatial knowledge improve language generation?
(2) How does each category of features contribute to generating image descriptions?

4.5.2 Method

We experimented on a relationships dataset from the Visual Genome Krishna et al. (2017), training several comparable neural network designs with

different spatial modules and different types of top-down knowledge about spatial relations. We changed the attention module in these models to be able to attend over language model features, visual features and geometric features and enriched the input representations with the additional geometric features representing the spatial configurations of objects. In addition to comparing the performance loss on unseen examples, we inspected the attention module to determine what features had a dominant effect on generating descriptions.

4.5.3 Findings and conclusions

We observed that the overall performance improved with the additional top-down knowledge of space. However, the results showed a substantial contribution of the language model representations in generating descriptions. Among added spatial knowledge localisation had the strongest effect, while the effects of role assignment and geometric spatial features were mixed. The reasons behind this outcome may be the bias in two kinds of regularities in data —the spatial composition of objects in photos in this dataset (location of objects are meaningful from the perspective of the photographer) and the task of describing object relations may have neglected the application of certain geometric relations (*'to the left of'*, *'to the right of'*) but preferring general, less specific spatial relations such as *'close'* and *'with'*.

Author contributions Mehdi Ghanimifard had the main responsibility for implementing the model, conducting the experiments and reporting it. Mehdi Ghanimifard and Simon Dobnik had shared responsibility for the remaining aspects of this research. Both authors read and approved the final manuscript.

4.6 Study 6: Learning to compose grounded spatial relations

Mehdi Ghanimifard and Simon Dobnik. Learning to Compose Spatial Relations with Grounded Neural Language Models. *In IWCS 2017-12th International Conference on Computational Semantics-Long papers. 2017.*

A basic definition of grounding linguistic units in visual perception is to associate words and phrases with visual features. Learning these associations must generalise from limited examples to novel unseen compositions. Compositionality in language imposes a systematic generalisation to the grounding of words and phrases. Due to the broad application of recurrent neural language models in vision and language tasks, this study investigated the capability of a recurrent language model in learning these compositional generalisations in the grounded language.

4.6.1 Questions

- To what extent is the language model trained on single examples can retrieve acceptability representations about the scene?
- Is the recurrent language model capable of generalising from word compositions to phrase compositions and how does it perform over previously unseen word compositions?

4.6.2 Method

Simple spatial relations are composable and can be used to construct new relations, such as *'above and to the right of'*, which denotes a relation constructed from two simple relations *'above'* and *'to the right of'*. We constructed artificially composed spatial templates based on their acceptability scores of the individual the spatial templates (Logan and Sadler, 1996) using known compositional operations. Then, from these templates we generated synthetic examples of individual situations and descriptions based on the aggregated acceptability scores. We tested different learning scenarios by controlling for unseen descriptions. In each experiment, we reconstructed the spatial templates of unseen descriptions based on the model scores over the unfolded predictions of words. Qualitatively, we measured the mismatches between the reconstructed templates and templates that were used to generate the artificial training data.

4.6.3 Findings and conclusions

We found that a grounded neural recurrent language model is capable of generalising when composing and decomposing word sequences both across the language and the perceptual domain.. We investigated the effects of three

factors on the success of the task and found that the degree to which the training data was impoverished had the most substantial effect, the type of composition is an essential factor in learning compositionality, and the presence of ungrounded distractor words had a small effect.

These observations leave an open question —is the performance for certain compositions reliant on intrinsic structures of recurrent neural networks for learning specific functions or the frequency and variation of data due to the semantic and combinatorial properties of compositions? Negation has fewer possible variations compared to 'OR' phrases and 'AND' phrases; therefore, it produces fewer instances and fewer variations in the training data for the negation marker. Distribution of the training data affects the final learned embeddings for the tokens of 'AND', 'OR' and 'NOT'. Learning to encode negation as a function in the recurrent unit might be harder than encoding other functions.

In summary, the combinatorial properties and semantics of different compositions affect the frequency and distribution of all tokens in language. While the distributional effects have potentially challenging consequences for the uniform learning of compositions, it can signal the difference between them. This is why, despite the imbalance in the number of compositions, the model could learn not ground distractor tokens.

Author contributions Mehdi Ghanimifard had the main responsibility for implementing the model, conducting the experiments and reporting it. Mehdi Ghanimifard and Simon Dobnik had shared responsibility for the remaining aspects of this research. Both authors read and approved the final manuscript.

4.7 Study 7: Metaphoricity of compositions with distributional representations

Yuri Bizzoni, Stergios Chatzikyriakidis and Mehdi Ghanimifard. "Deep" Learning: Detecting Metaphoricity in Adjective-Noun Pairs. *In Proceedings of the Workshop on Stylistic Variation, pp. 43-52. 2017.*

Recognising metaphoric use of language requires an understanding of the situation, context and how expressions refer to extra-linguistic knowledge

about the world. On the other hand, distributional knowledge in unimodal language models encodes word-context associations. Even without the presence of extra-linguistic knowledge of situations, distributional knowledge might be able to determine metaphorical adjective-nouns. In this study, we proposed that knowledge of the compositionality of adjective-nouns is encoded in the pre-trained word embeddings of textual corpora and a simple neural network can transfer this knowledge to metaphor recognition tasks. We used methods of vector composition in a neural network design to predict the metaphoricity of adjective-noun compositions.

4.7.1 Questions

- Is it possible to detect metaphoric adjective-noun compositions using pre-trained word embeddings in a shallow neural network?
- Are there any differences in performance between design choices and language model types, including word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and dependency-based embeddings (Levy and Goldberg, 2014)?

4.7.2 Method

We compared different methods of vector composition in a neural network design, similar to Mitchell and Lapata (2010), and used different pre-trained word embeddings. We examined the performance of these models with cross-validation on unseen adjectives and unseen adjective-noun pairs.

4.7.3 Findings and Conclusions

We found that pre-trained word embeddings with simple neural network designs performed better than previous approaches without using word embeddings. This study raises a question if similar designs could expand metaphoricity judgments to other part-of-speech compositions. The high performance of the textual word embeddings up to 93% accuracy confirms that unimodal language models can encode some knowledge of the referential meaning to real situations. However, questions concerning the type of knowledge and it are left for subsequent studies.

Author contributions Mehdi Ghanimifard had the main responsibility for writing the model section of the manuscript. Mehdi Ghanimifard and Yuri Bizzoni had shared responsibility on running the experiments and reporting it. Yuri Bizzoni and Stergios Chatzikyriakidis had shared responsibility for the remaining aspects of this research. All authors read and approved the final manuscript.

4.8 Summary

In the first three studies, we focused on latent extra-linguistic knowledge of spatial relations in unimodal neural recurrent language models and on geometric features as represented by bounding boxes.

In studies 4 and 5, we examined the contribution of visual features, geometric features, and the contextual embeddings of a neural language model when generating image descriptions. We showed that, in training generative neural language models, the spatial knowledge used in the task is also learned latently in language models.

In the last two studies, the focus of the research was on the capability of neural language models to learn compositional knowledge and generalise from limited samples to new word compositions.

Final Discussions

5

5.1 From aims to findings

This thesis aimed to build and examine systems capable of generating and understanding situated language. Using deep neural networks, we may be able to build language models to imitate natural language. However, explanations are required regarding what knowledge is encoded in the models, how the models encode relevant knowledge and if such data-driven methods satisfy the systematic generalisations required for going beyond limited data sets. The recent success of deep learning methods in vision and language tasks are promising and challenge theoretical discussions about language grounding and explainability.

A study on spatial expressions in image descriptions provides challenges and broad applications of a vision and language model for situated language processing. The challenge is to understand how a model should and would ground language in spatial knowledge. Spatial knowledge could include the geometry of a scene and the location of objects; alternatively, it could include causality in physics and the functional affordance of objects in relation to each other. The grounding of linguistic categories in these two types of knowledge presents a challenge for disentangling the representation of two types of knowledge. In the context of deep learning methods, we asked three research questions:

(Q1) What type of spatial knowledge is encoded in language models?
(Q2) How does the model encode semantic knowledge?
(Q3) Is there systematic generalisation of the knowledge?

In seven studies, we contributed to the discussion on grounding and answered the questions regarding the use of neural language models. The first (Dobnik et al., 2018) and the second (Ghanimifard and Dobnik, 2019a) studies focused on unimodal language models for spatial descriptions. The corpus data suggests a statistical dependency between semantic components of a spatial description ⟨TARGET, RELATION, LANDMARK⟩. This explained

with the functional meaning of spatial relations. The fact that spatial descriptions of object pairs are predictable is mostly because of their functional relationship. The overlap between the functional and geometric sense in linguistic categories of spatial relations contributes to the encoding of knowledge about geometry in word distributions as well. On the other hand, both studies suggest the possibility that, in an image description task, spatial expressions tend to explain *what* the objects are in the picture instead of *where* they are. Therefore, the non-geometric sense of spatial descriptions has a strong effect in these corpora.

With a focus on spatial grounding in the spectrum of functional/geometric sense of relations, in our third study, we looked at the geometric properties of bounding box annotations in images and their distributions for different spatial expressions. We found that the variation in the relative location of objects in geometrically biased expressions is lower than in the functionally biased relations. This finding is consistent with the predictability of functional relations from linguistic evidence rather than geometric features. This conclusion has implications for the evaluation of multi-modal language models, which brought us to the fourth and the fifth studies.

The fourth study (Ghanimifard and Dobnik, 2018) examined the possibility of evaluating grounding based on adaptive attention. We found that pretrained convolutional visual features contributed more to the generation of nouns compared to other parts of speech. Some spatial relations are more dependent on contextual language embeddings. This is consistent with our view that spatial relations in image descriptions are less dependent on the location of objects. Due to the opaque representation of space in convolutional features, further studies are required how these contribute to spatial expressions and whether such representations can be improved with feature engineering.

In the fifth study (Ghanimifard and Dobnik, 2019b), we extended the adaptive attention to enrich the visual features with locational information. We found that top-down algorithmic localisation has the most positive effect on language generation among the different methods for enriching visual features. The effect of both a top-down semantic role assignment and geometric feature vectors is positive, but much less than expected. This observation is consistent with our findings in the third and fourth studies on unimodal language models, which indicated that reliable predictability of object relations

without visual features varies depending on the kind of spatial relations in the absence of adequate geometric descriptions. These observations demand further studies, especially beyond image description tasks, for example in visual question answering.

In the sixth study (Ghanimifard and Dobnik, 2017), weexamined the degree of generalisationa recurrent language model learned compositional descriptions. We found that the generalisation depends on both the combinatorial and semantic properties of the compositions. The combinatorial properties of compositions change the variations and frequencies of possible phrases (unary vs binary compositions). The semantics of the compositions shape the acceptable space. , for example conjunction and disjunction result in different frequencies. Both combinatorial and semantic properties of compositions contribute to token distributions in language.

In the seventh study (Bizzoni et al., 2017), we examined if the knowledge from a unimodal language model could recognise the metaphoricity of adjective-noun compositions. This contributed to an understanding of the type of knowledge that could be encoded in the language model. Distribution of tokens, as seen in the first study, could affect the generalisation in language grounding. According to this study, it contributes to encoding of deeper non-perceptible knowledge, such as metaphors.

5.2 Knowledge and grounding

One of the central claims of this thesis is that some spatial knowledge is encoded in neural language models. Then, despite the fact that representations in language are not linked to primitive sensory representations, we used the term *grounding* for spatial descriptions that are explainable with a language model instead of perceptual inputs. This argument requires a more in-depth discussion about the definition of *spatial knowledge* and *grounding*.

Spatial knowledge In this thesis, the term knowledge was extensively used to describe language grounding in (1) *spatial knowledge*, (2) *geometric knowledge*, (3) *functional knowledge* and (4) *distributional knowledge* or *knowledge in language models*. The main argument of the thesis is that spatial language projects onto the representations in language models. Therefore,

distributional knowledge of spatial relations encodes the spatial knowledge (findings of studies 1, 2, 4 and 5 concerning Q1). Nevertheless, the distinction between functional and geometric knowledge implies that there are two different types of spatial expressions. Geometric knowledge is a literal sense of space and functional knowledge is an abstraction of non-spatial relations between objects. We expect that, by capturing regularity in language use, distributional knowledge captures functional knowledge (study 1 concerning Q1). However, the distributional distinction between functional and geometric use is entangled in datasets. Therefore, distributional knowledge captures regularities that seem more geometric than functional.

This distributional property is an artefact of the entangled concept of space. The skewness of spatial relations in datasets is a result of this entanglement. For example, the reason the functional sense of the relation *'over'* as a sheltering relation is possible is because of its geometric properties and the rules of physics. Similarly, the reason why some objects are functionally related is because of their geometric shapes and their geometric capacity of being in that position. The entangled relation between functional and geometric meaning calls for a better understanding of spatial knowledge. Without spatial reasoning, the functional meaning of the relations is not possible. In this, we argued that language models capture spatial knowledge, but also that there are different types of spatial knowledge and what their implications are for descriptions in different contexts. The evaluation of the relation between the knowledge kinds in different contexts in which spatial descriptions are made should be addressed in future work.

Grounding The conclusion of this thesis with regard to grounding is that any prediction based on the available evidence is a form of grounding. The representation of this evidence varies in models. Predictions of a show-and-tell system are grounded in both situated features and the modelling assumptions, such as the function and design of the model (composition of modules), its training data (sufficient data for model convergence) and its learning goals. Therefore, when the system makes predictions without relying on situated features or with minimal attention to these situated features, there are two explanations of this performance:

(1) Some assumptions used building the model are erroneous, such as the assumption around what training data provides sufficient knowledge for constructing the model.

(2) The situated features do not contain independent encoding of knowledge required for the task. In other words, the task relies on other information, such as world knowledge. This kind of missing information could be included as different representations of the model. This can be done by:
   (i) a module that fuses language and vision to exploit the fusion of the situated features with other representations, or
   (ii) exclusively encoding some knowledge about the task in modules, such as language models, that provide distributional evidence.

In the case of generating spatial descriptions, the predictability of relations from textual evidence or with minimal contribution from visual features has two explanations:

(1) There might be mistakes in the model design, feature representations or the assumption that training data has appropriate information for the task. For example, knowledge about embodied actions and interactions between objects may be missing from show-and-tell datasets.
(2) Some spatial knowledge is encoded in the distributional knowledge of language, in addition to situated visual knowledge, such as functional knowledge and frame of reference. The neural language model encodes this knowledge in its parameters. The composition of the visual module and the language module contextualises the representations based on the training data.

With this conclusion about grounding, we can examine the future improvements of vision and language models.

5.3 Future work

The discussion of grounding and learning representations with deep learning methods opens several directions for future research:

- The current model designs use simple tools of modality fusion, such as embedding representations, attention mechanisms and simple vector manipulations, including concatenation or multiplication. More research on modality fusion is required in future studies.
- Our attempts to understand what is learned in neural language models can be expanded with additional methods of explainability and probing

representations. The question of what representations are learned and what are the effect of parameters is beneficial for improving the algorithms.

- In addition to investigating explainability and developing better modules, more rigorous testing of models is required to measure their success. Such a study would lead to development of better learning goals and loss functions for the model. Instead of language modelling with token level loss, new loss functions related to task problem-solving, such as spatial navigation, may be able to learn different aspects of meaning in language models.
- This requires a better understanding of the data. We found two types of bias in image description datasets:
  (i) The bias in the task constrains the words to specific senses. In the image description task, spatial relations have a strong bias towards relating *what* is in the picture, instead of relating *where* objects are in respect to each other.
  (ii) The bias in the visual composition of images. The images in image captioning datasets are focused on objects in regions of interest. This suggests that other datasets should also be examined, such as those collected from ego centric robotic sensory and imaging data, which lack such a focus of attention on objects as a property of image compositions.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *CVPR*, 3(5):6.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *journal of machine learning research*, 3(Feb):1137–1155.

Yuri Bizzoni, Stergios Chatzikyriakidis, and Mehdi Ghanimifard. 2017. "deep" learning : Detecting metaphoricity in adjective-noun pairs. In *Proceedings of the Workshop on Stylistic Variation*, pages 43–52, Copenhagen, Denmark. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Guillem Collell and Marie-Francine Moens. 2018. Learning representations specialized in spatial knowledge: Leveraging language and vision. *Transactions of the Association of Computational Linguistics*, 6:133–144.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kenny R Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V Richards. 2004. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In *International Conference on Spatial Cognition*, pages 98–110. Springer.

Kenny R Coventry, Merce Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of memory and language*, 44(3):376–398.

Navneet Dalal and Bill Triggs. 2005. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik and Amelie Åstbom. 2017. (Perceptual) grounding as interaction. In *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–26, Saarbrücken, Germany.

Simon Dobnik and Robin Cooper. 2017. Interfacing language, spatial perception and cognition in Type Theory with Records. *Accepted for Journal of Language Modelling*, n(n):1–30.

Simon Dobnik and Mehdi Ghanimifard. 2020. Spatial descriptions on a functional-geometric spectrum: the location of objects. *Preprint - under review*.

Simon Dobnik, Mehdi Ghanimifard, and John Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 1–11, New Orleans. Association for Computational Linguistics.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceed-*

ings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue, pages 24–32, Gothenburg, Sweden.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pages 1–6, Berlin, Germany.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302.

Desmond Elliott and Arjen de Vries. 2015. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 42–52.

Klaus-Peter Gapp. 1994a. Basic meanings of spatial relations: computation and evaluation in 3D space. In *Proceedings of the twelfth national conference on Artificial Intelligence (AAAI'94)*, volume 2, pages 1393–1398, Menlo Park, CA, USA. American Association for Artificial Intelligence, AAAI Press/MIT Press.

Klaus-Peter Gapp. 1994b. A computational model of the basic meanings of graded composite spatial relations in 3D space. In *Advanced geographic data modelling. Spatial data modelling and query languages for 2D and 3D applications (Proceedings of the AGDM'94)*, Publications on Geodesy 40, pages 66–79. Netherlands Geodetic Commission.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Mehdi Ghanimifard and Simon Dobnik. 2018. Knowing when to look for what and where: Evaluating generation of spatial descriptions with adaptive attention. In *Proceedings of the 1st Workshop on Shortcomings in Vision and Language (SiVL'18), ECCV, 2018*.

Mehdi Ghanimifard and Simon Dobnik. 2019a. What a neural language model tells us about spatial relations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, pages 71–81, Minneapolis, Minnesota. Association for Computational Linguistics.

Mehdi Ghanimifard and Simon Dobnik. 2019b. What goes into a word: generating image descriptions with top-down spatial knowledge. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 540–551, Tokyo, Japan. Association for Computational Linguistics.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

Herbert Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 41–58. Academic Press.

Wilhelm Grimm and Richard Andre. 1899. *Grimm's fairy tales : retold in one-syllable words*. McLoughlin Brothers. https://archive.org/details/grimmsfairytalesgrim/page/42/mode/2up.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. 2011. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

Audun Jøsang and David McAnally. 2005. Multiplication and comultiplication of beliefs. *International Journal of Approximate Reasoning*, 38(1):19–51.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *CLASP Papers in Computational Linguistics: Proceedings of the Conference on Logic and Machine Learning in Natural Language*.

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1041–1048, Sydney, Australia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014a. Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014b. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105.

Barbara Landau. 1996. Multiple geometric representations of objects in languages and language learners. *Language and space*, pages 317–363.

Barbara Landau and Ray Jackendoff. 1993. Whence and whither in spatial language and spatial cognition? *Behavioral and brain sciences*, 16(2):255–265.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

Yann LeCun, Koray Kavukcuoglu, Clément Farabet, et al. 2010. Convolutional networks and applications in vision. In *ISCAS*, volume 2010, pages 253–256.

K. Lenc and A. Vedaldi. 2015a. R-cnn minus r. In *BMVC*, pages 5.1–5.12. BMVA Press. Oral Presentation.

Karel Lenc and Andrea Vedaldi. 2015b. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 991–999.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 302–308, Baltimore, Maryland. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423*.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A. Miller and Philip N. Johnson-Laird. 1976. *Language and perception*. Cambridge University Press, Cambridge.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1184, Atlanta, Georgia. Association for Computational Linguistics.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

St. George Jackson Mivart. 1881. *The common frog*. Macmillan and co. https://archive.org/details/commonfrog1881miva/page/n5/mode/2up.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648. ACM.

Will Monroe, Noah D Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. *arXiv preprint arXiv:1606.03821*.

Richard Montague. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Arnau Ramisa, JK Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220. Association for Computational Linguistics.

Frank P Ramsey. 1931. Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198.

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

Deb Roy and Niloy Mukherjee. 2005. Towards situated speech understanding: Visual context priming of language models. *Computer Speech & Language*, 19(2):227–248.

Deb K Roy. 2002. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*, 16(3-4):353–385.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. Vision and language integration: moving beyond objects. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Leonard Talmy. 1983. How language structures space. In Herbert L. Pick Jr. and Linda P. Acredolo, editors, *Spatial orientation: theory, research, and application*, pages 225–282. Plenum Press, New York.

Marc Tanti, Albert Gatt, and Kenneth P Camilleri. 2018a. Quantifying the amount of visual information used by neural caption generators. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.

Marc Tanti, Albert Gatt, and Kenneth P Camilleri. 2018b. Where to put the image in an image caption generator. *Natural Language Engineering*, 24(3):467–489.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Part II

Studies

6 Functional/Geometric Bias In Neural Language Models

> **Simon Dobnik, Mehdi Ghanimifard and John Kelleher.** Exploring the functional and geometric bias of spatial relations using neural language models, *In Proceedings of the First International Workshop on Spatial Language Understanding, pp. 1-11. 2018.*

Abstract The challenge for computational models of spatial descriptions for situated dialogue systems is the integration of information from different modalities. The semantics of spatial descriptions are grounded in at least two sources of information: (i) a geometric representation of space and (ii) the functional interaction of related objects that. We train several neural language models on descriptions of scenes from a dataset of image captions and examine whether the functional or geometric bias of spatial descriptions reported in the literature is reflected in the estimated perplexity of these models. The results of these experiments have implications for the creation of models of spatial lexical semantics for human-robot dialogue systems. Furthermore, they also provide an insight into the kinds of the semantic knowledge captured by neural language models trained on spatial descriptions, which has implications for image captioning systems.

6.1 Introduction

Spatial language understanding is fundamental requirement for human-robot interaction through dialogue. A natural task for a human to request a robot to fulfil is to retrieve or replace an object for them. Consequently, a particularly frequent form of spatial description within human-robot interaction is a

locative expression. A locative expression is a noun phrase that describes the location of one object (the *target object*) relative to another object (the *landmark*). The relative location of the target object is specified through a prepositional phrase:

$$\text{Bring me } \underbrace{\underbrace{\textit{the big red book}}_{\textit{Target}} \; \textit{on} \; \underbrace{\textit{the table}}_{\textit{Landmark}}}_{\substack{\textit{Prepositional} \\ \textit{Phrase}}} \underbrace{\qquad\qquad}_{\textit{Locative Expression}}.$$

In order to understand these forms of spatial descriptions a robot must be equipped with computational models of the spatial semantics of prepositions that enable them to ground the semantics of the locative expression relative to the context of the situated dialogue.

A natural approach to developing these computational models is to define them in terms of scene *geometry*. And, indeed, there is a tradition of research that follows this path, see for example Logan and Sadler (1996); Kelleher and Costello (2005, 2009). However, there is also a body of experimental and computational research that has highlighted that the semantics of spatial descriptions are dependent on several sources of information beyond scene geometry, including *functional semantics* (which encompasses a range of factors such as world knowledge about the typical interactions between objects, and object affordances) Coventry and Garrod (2004). We can illustrate this distinction between geometric and functionally defined semantics using a number of examples. To illustrate a geometric semantics: assuming a spatial meaning, anything can be described as *to left of* anything else so long the spatial configuration of the two objects is geometrically correct. However, as Coventry et al. (2001) has shown the spatial description *the umbrella is over the man* is sensitive to the protective affordances of the umbrella to stop rain, and is appropriate in contexts where, the umbrella is not in a geometrically prototypical position above the man, so long as the umbrella is protecting the man from the rain.

A further complication with regard to modelling the semantics of spatial descriptions is that experimental results indicate that the contribution of geometrical and functional factors is not the same for every spatial relation (Garrod et al., 1999; Coventry et al., 2001). This experimental work shows that there is an interplay between function and geometry in the definition of

spatial semantics and therefore the spatial meaning of given spatial relation is neither fully functional nor fully geometric. Rather, spatial terms can be ordered on a spectrum based on the sensitivity of their semantics to geometric or functional factors.

Given the distinction between geometric and functional factors in shaping spatial semantics, a useful analysis that would inform the design and creation of computational models of spatial semantics is *to identify the particular semantic bias (geometric/functional) that each spatial term evinces*. However, such an analysis is difficult. Native speakers do not have strong intuitions about the bias of prepositions and such bias had to be established experimentally Coventry et al. (2001); Garrod et al. (1999) or through linguistic analysis (Herskovits, 1986, p.55).[1] Reviewing the literature on this experimental and analytic work reveals that prepositions such as *in*, *on*, *at*, *over*, *under* have been identified as being functionally biased, whereas *above*, *below*, *left of* and *right of* are geometrically biased. Other spatial relations may be somewhere in between. In this paper we will use these relations as ground-truth pointers against which our methods will be evaluated. If the method is successful, then we are able to make predictions about those relations that have not been verified for their bias experimentally. Knowing the bias of a spatial relation is useful both theoretically and practically. Theoretically, it informs us about the complexity of grounded semantics of spatial relations. In particular, it engages with the "what" and "where" debate where it has been argued that spatial relations are not only spatial (i.e. geometric) Landau and Jackendoff (1993); Coventry and Garrod (2004); Landau (2016). Practically, the procedure to estimate the bias is useful for natural language generation systems, for example in situated robotic applications that cannot be trained end-to-end. Given that a particular pair of objects can be described geometrically with several spatial relations, the knowledge of functional bias may be used as a filter, prioritising those relations that are more likely for a particular pair of objects, thereby incorporating functional knowledge. This approach to generation of spatial descriptions is therefore similar to the approach that introduces a cognitive load based hierarchy of spatial relations Kelleher and Kruijff (2006) or a classification-based approach that combines geometric (related to the bounding box), textual

[1] The discussion of Herskovits focuses on interaction of objects conceptualised as geometric shapes, for example *on:* contiguity with line or surface. The fact that the interacting objects can be conceptualised as different geometric shapes points and therefore related by a particular prepositions points to their functional nature as discussed here.

(word2vec embeddings) and visual features (final layer of a convolutional network) Ramisa et al. (2015). The functional geometric bias of spatial relations could also be used to inform semantic parsing, for example in prepositional phrase attachment resolution Christie et al. (2016); Delecraz et al. (2017).

Previous work has investigated metrics of the semantic bias of spatial prepositions, see Dobnik and Kelleher (2013, 2014). (Dobnik and Kelleher, 2013) uses (i) normalised entropy of target-landmark pairs to estimate variation of targets and landmarks per relation and (ii) log likelihood ratio to predict the strength of association of target-landmark pairs with a spatial relation and presents ranked lists of relations by the degree of argument variation or strength of the association respectively. The approach hypothesises that functionally biased relations are more selective in the kind of targets and landmarks they co-occur with. The reasoning behind this is that geometrically it is possible to relate a wider range of objects than in the case where additional functional constrains between objects are also applied. Dobnik and Kelleher (2014) generalises over landmarks and targets in WordNet hierarchy and estimates the generality of the types of landmark. Again, the work hypothesises that functional relations are more restricted in their choice of target and landmark objects and therefore are generally more specific in terms of the WordNet hierarchy. Both papers present results compatible with the hypotheses where the functional or geometric nature of prepositions is predicted in line with the experimental studies Garrod et al. (1999); Coventry et al. (2001).

Sensitive to the fact that relations such as *in* and *on* not only have spatial usage but also usages that may be considered metaphoric Steen et al. (2010), both Dobnik and Kelleher (2013) and Dobnik and Kelleher (2014) were based on an analysis of a corpus of image captions. The idea being that descriptions of images are more likely to contain spatial descriptions grounded in the image. For similar reasons, we also employ a corpus of image descriptions (larger than in the previous work).

This paper adopts a similar research hypothesis to Dobnik and Kelleher (2014, 2013), namely that: it is possible to distinguish between functionally biased and geometrically biased spatial relations by examining the diversity of the contexts in which they occur. Defining the concept of context in terms of the *target* and *landmark* object pairs that a relation occurs within, the rationale

of this hypothesis is that: geometrically biased relations are more likely to be observed in a more diverse set of contexts, compared to functionally biased relations, because the use of a geometrically biased relation only presupposes the appropriate geometric configuration whereas the use of a functionally biased relation is also constrained by object affordances or typical interactions.

However, the work presented in this paper provides a more general analytical technique based on a neural language model Bengio et al. (2003); Mikolov et al. (2010) which is applied to a larger dataset of spatial descriptions. We use neural language models as the basic tool for our analysis because they are already commonly used to learn the syntax and semantics of words in an unsupervised way. The contribution of this paper in relation to (i) the previous analyses of geometric and functional aspects of spatial relations is that it examines whether similar predictions can be made using these more general tools of representing meaning of words and phrases; the contribution to (ii) deep learning of language and vision is that it examines to what extent highly specific world-knowledge can be extracted from a neural language model. The paper proceeds as follows: in Section 6.2 we describe the datasets and their processing, in Section 6.3 we describe the basics behind language models and the notion of perplexity, in Section 6.4 and 6.5 we present and discuss our results. We conclude in Section 6.6.

The code that was used to produce the datasets and results discussed in this paper can be found at:
`https://github.com/GU-CLASP/functional-geometric-lm`.

6.2 Datasets

The Amsterdam Metaphor Corpus Steen et al. (2010) which is based on a subsection of a BNC reveals that the spatial sense of prepositions are very rare in genres such as news, fiction and academic texts. For example, *below* only has two instances that are not labelled as a metaphor and more than 60% of fragments with *in*, *on*, and *over* are not used in their spatial sense. For this reason Dobnik and Kelleher (2013) use two image description corpora (IAPR TC-12 Grubinger et al. (2006) and Flickr8k Rashtchian et al. (2010)) where spatial uses of prepositions are common. They apply a dependency parser and a set of post-processing rules to extract spatial relations, target and

landmark object triplets. The size of this extracted dataset is 96,749 instances and is relatively small for training a neural language model. Kordjamshidi et al. (2017) released CLEF 2017 multimodal spatial role labelling dataset (mSpRL) which is a human annotated subset of the IAPR TC-12 Benchmark corpus for spatial relations, targets and landmarks Kordjamshidi et al. (2011) containing 613 text files and 1,213 sentences. While this dataset could not be used to train a language model directly, a spatial role labelling classifier could be trained on it to identify spatial relations and arguments which would then be used to produce a bootstrapped dataset for training a neural language model.

Recently, Visual Genome Krishna et al. (2017) has been released which is a crowd-source annotated corpus of 108K images which also includes annotations of *relationships* between (previously annotated) bounding boxes. Relationships are predicates that relate objects which include spatial relations (2404639, "cup on table"), verbs (2367163, "girl holding on to bear") as well as combinations of verbs and spatial relations (2317920, "woman standing on snow") and others. We use this dataset in the work reported here. Its advantage is that it contains a large number of annotated relationships but the disadvantage is that these are collected in a crowd-sourced setting and are therefore sometimes noisy but we assume these are still of better quality than those from a bootstrapped machine annotated dataset.

To extract spatial relations from the annotated relationships, we created a dictionary of their syntactic forms based on the lists of English spatial relations in Landau (1996) and Herskovits (1986). For the training data we preserve all items annotated as relationships as single tokens ("jumping_over") and we simplify some of the composite spatial relations based on our dictionary, e.g. "left of" and "to the left of" become "left" to increase the frequency of instances. This choice could have affected our results if done without careful consideration. While compound variants of spatial relations have slightly different meanings, we only collapsed those relations for which we assumed this would not affect their geometric or functional bias. Furthermore, Dobnik and Kelleher (2013) show that compound relations cluster with their non-compound variants using normalised entropy of target-landmark pairs as a metric. Finally, some variation was due to the shorthand notation used by the annotators, e.g. "to left of". The reason behind keeping all relation(ships) in the training set is to train the language model on as many targets and landmarks as possible and to learn paradigmatic relations between them.

We normalise all words to lowercase and remove the duplicate descriptions per image (created by different annotators). We also check for and remove instances where a spatial relation is used as an object, e.g. "chair on left". We remove instances where one of the words has fewer than 100 occurrences in the whole dataset which reduces the dataset size by 10%. We add start and end tokens to the triplets ($\langle s \rangle$ $target$ $relation$ $landmark$ $\langle /s \rangle$) as required for training and testing a language model. The dataset is shuffled and split into 10 folds that are later used in cross-validation. In the evaluation, we take 20 samples per spatial relation from the held out data of those relations that are members of the dictionary created previously. This way the average perplexity is always calculated on the same number of samples per each relation.[2]

6.3 Language model and perplexity

6.3.1 Language model

Probabilistic language models capture the sequential properties of language or paradigmatic relations between sequences of words. Using the chain rules of probabilities they estimate the likelihood of a sequence of words:

$$P(w_{1:T}) = \sum_{t=1}^{T} P(w_{t+1}|w_{1:t}) \tag{6.1}$$

Neural language models estimate probabilities by optimising parameters of a function represented in a neural architecture Bengio et al. (2003):

$$\hat{P}(w_{t+1}|w_{1:t} = v_{k_{1:t}}) = f(v_{t-1}; \Theta) = \hat{y}_t \tag{6.2}$$

where Θ represents parameters of the model, f being the composition of functions within the neural network architecture, and $v_{k_{1:t}}$ the words up to time t in the sequence. The output of the function is $\hat{y}_t \in R^n$, a vector of probabilities, with each dimension representing the probability of a word in the vocabulary. The loss of a recurrent language model is the

[2] The reason we use 20 sample is that this is also the size of the 10% test folds in the down-sampled dataset described later. In selecting 20 items for the test-set we also ensure that it contains the vocabulary in the down-sampled training folds.

average surprisal for each batch of data Graves et al. (2013); Mikolov et al. (2010):

$$loss(S) = -\sum_{s \in S}\sum_{t=0}^{|s|} \frac{\log(\hat{y}_t(v_{k_{t+1}}))}{|S| \times |s|} \tag{6.3}$$

Note that our architecture is deliberately simple as we apply it in an experimental setting with constrained descriptions[3]. We use a Keras implementation Chollet et al. (2015), and fit the model parameters with Adam Kingma and Ba (2014) with a batch size of 32 and iterations of 20 epochs. On each iteration the language model is optimised on the loss which is related to perplexity as described in the following section.

6.3.2 Perplexity

Instead of calculating the averages of likelihoods from Equation 6.1, which might get very low on long sequences of text, we use perplexity which is an exponential measure for average negative log likelihoods of the model. This solves the representation problem with floating points and large samples of data.

$$Perplexity(S, P) = 2^{E_S[-log_2(P(w_{1:T}))]} \tag{6.4}$$

where $w_{1:T}$ is an instance in a sample collection S. Perplexity is often used for evaluating language models on test sets. Since language models are optimised for low perplexities[4], the perplexity of a trained model can be used as a measure of fit of the model with the samples.

6.4 Varying targets and landmarks

6.4.1 Hypotheses

As a language model encodes semantic relations between words in a sequence we therefore expect that the distinction between functional and geometric spatial relations will also be captured by it. As functionally biased spatial relations are used in different situational contexts than geometrically biased spatial relations, we expect that a language model will capture this

[3] For more details on the architecture see Section 6.7 in the supplementary material, in particular Figure 6.6 and Equation 6.5.

[4] Equation 6.4 is related to Equation 6.3 as perplexity is 2^{Loss} given a neural model as the likelihood model.

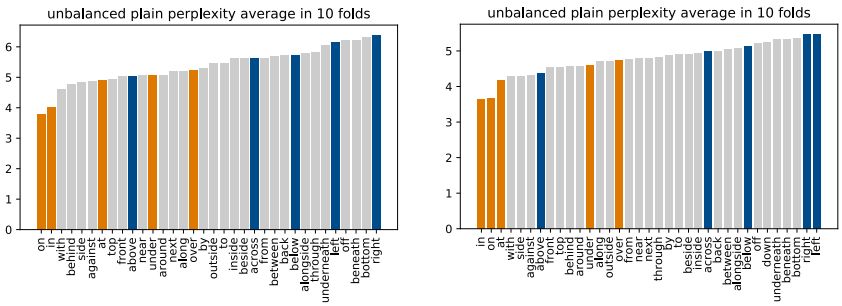
bias in different distributions of target and landmark objects in the forms of the perplexity of phrases. Our weak hypothesis is that the perplexity of phrases on the test set reflects the functional-geometric bias of a spatial relation (Hypothesis 1). We take the assumption that functionally-biased relations are more selective in terms of their target and landmark choice (Section 6.1) and consequently sequences such as `<s> target relation landmark </s>` with functional relations have a higher predictability in the dataset resulting in a lower perplexity in the language model (Hypothesis 2). Related to this hypothesis, there is a stronger hypothesis that target and landmark are predictable with a given functional spatial relation (Hypothesis 3).

6.4.2 Method

We train two language models as described in Section 6.3.1. For training and evaluation 10-fold cross-validation is used and average results are reported. We ensure that the evaluation sets contain no vocabulary not seen during the training. The language model 1 (LM1) is trained on unrestricted frequencies of instances. In training the language model 2 (LM2) we down-sample relations so that they are represented with equal frequencies. The dataset to train LM2 contains 200 instances of each possible relations while the evaluation set contains 20 instances for each spatial relation. Note that using this method some targeted spatial relations might disappear from the evaluation set as their frequency in the held-out data is too low. In addition to the requirement that the evaluation set contains no out-of-vocabulary items, the target and landmarks are included without restriction on their frequency, as they occur with these spatial relations.

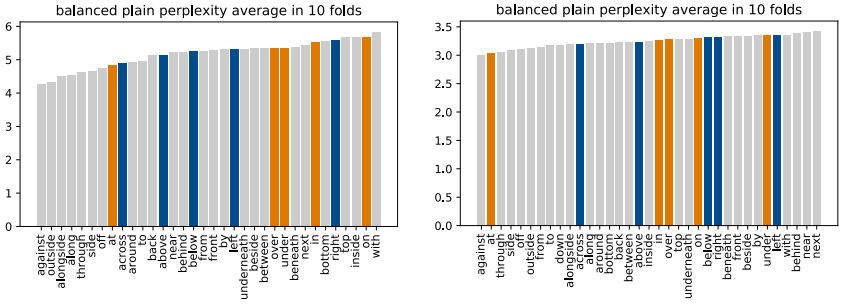
6.4.3 Results

Figure 6.1 shows the estimated average perplexities of a subset of spatial relations, those that satisfy the sampling frequency requirement described in Section 6.4.2. Functionally and geometrically biased spatial relations as identified experimentally in the literature (Section 6.1) are represented with orange and blue bars respectively. There is a tendency that functionally biased relations lead to lower mean perplexity of phrases (Hypothesis 2 is confirmed) and also that there is a tendency that spatial relations of a particular bias cluster together (Hypothesis 1 is also confirmed). We



(a) test-set (b) training set

Figure. 6.1: Mean perplexities of spatial descriptions of LM1 (orange: functionally biased, blue: geometrically biased relations).



(a) test-set (b) training set

Figure. 6.2: Mean perplexities of LM2 by spatial relation (orange: functionally biased, blue: geometrically biased).

report results both on the training set and the test set which show the same tendencies. This means that our model generalises well on the test set and that the latter is representative.

However, in the language model the perplexities are biased by the frequency of individual words: more frequent words are more likely and therefore they are associated with lower LM perplexity. The results show high Spearman's rank correlation coefficient $\rho = 0.90$ between frequencies of spatial relation in the dataset and the perplexity of the model on the test set: on (329,529) $>$ in (108,880) $>$ under (11,631) $>$ above (8,952) $>$ over (5,714) $>$ at (4,890) $>$ below (2,290) $>$ across (1,230) $>$ left (996) $>$ right (891).

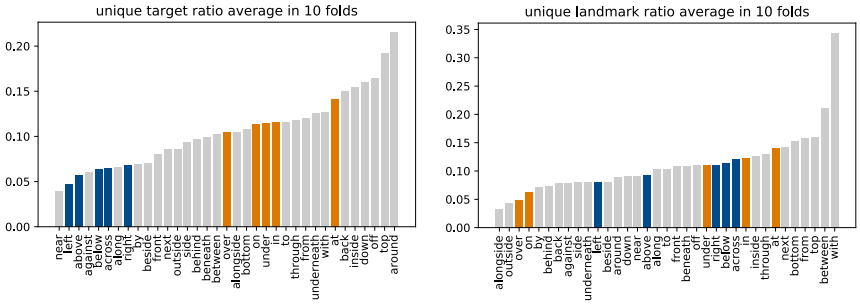
For the purposes of our investigation in predictability of target-landmark pairs (Hypothesis 3) we should avoid the bias in the training set. In order to exclude the bias of frequencies of relations, we evaluate LM2 where spatial relations are presented with equal frequencies in training. Figure 6.2 shows the ranking of spatial relations by the perplexities when the language model was trained with balanced frequencies. The two kinds of spatial relations are less clearly separable as the colours overlap (Hypothesis 3 is not confirmed). In comparison to Figure 6.1 there is an observable trend that all instances lead to lower perplexities in the training set which is the effect of down-sampling on vocabulary size. Figure 6.2 also shows that phrases with geometrically biased spatial relations have a higher change towards lower perplexities.

Noting that the frequency of using functionally-biased spatial relations are higher in English, this bias and our strong hypothesis for predictability of target-landmark pairs can be expressed with simple joint probabilities which we are estimating with the language model:

$$P(\text{target}, \text{relation}, \text{landmark}) = P(\text{relation})P(\text{target}, \text{landmark}|\text{relation})$$

It is possible that targets and landmarks that occur with these relations are very specific to these relations but infrequent with other relations. When we remove their frequency support provided by the frequency of relations these targets and landmarks become infrequent in the dataset and therefore less probable which on overall results in higher perplexities of phrases with functionally-biased relations. Specificity of targets and landmarks can be a source of these results.

To provide (some) evidence for this assumption, Figure 6.3 shows the average ratios of unique types over total types of targets and landmarks in the balanced dataset over 10-folds on which LM2 was trained. There is a very clear division between functionally and geometrically biased spatial relations in terms of the uniqueness of targets, functionally-biased relations are occurring with more unique ones which contributes to higher perplexity of LM2. There is less clear distinction between the two kinds of spatial relations in terms of uniqueness of landmarks. Some functional relations such as *on* occur with fewer unique landmarks than targets (from .11 to .06), some geometric relations such as *right* occur with more unique landmarks than targets (from .07 to .11). The asymmetry between targets and landmarks



(a) targets (b) landmarks

Figure. 6.3: Ratio between unique types and all types per spatial relation in the balanced dataset for LM2.

is expected since the choice of landmarks in the image description task is restricted by the choice of the targets (as well as other contextual factors such as visual salience). They have to be "good landmarks" to relate the targets to. A functional relation-landmark pair is more related to the target through the landmark's affordances whereas a geometric relation-landmark pair is more related to the target through geometry. This might explain for example, why *on* has fewer, but *right* has more unique landmarks than targets. On the other hand there are also relations where the ratio of unique targets and landmarks is very similar, for example *at* (.14 and .14). Overall, it appears that if uniqueness of objects is contributing to the perplexity of the language model of phrases which functionally-biased relations (which in this balanced dataset is the case) then this is more contributed by targets rather than the landmarks.

To further explore the idea of asymmetry between targets and landmarks we re-arranged the targets and landmarks in the descriptions from the balanced dataset that LM2 was trained to `<s> landmark relation target </s>` and trained LM2′. The average perplexities over 10-folds of cross-validation are shown in Figure 6.4. Comparing Figure 6.4 with Figure 6.2 we first observe that the perplexity of LM2′ on the descriptions is overall several magnitudes lower than the perplexity of LM2 (max 0.06, max 140). Secondly, we observe that the perplexities of phrases containing different relations are very similar and that there is no separation of phrases by perplexity depending on the relation bias. The results are in line with our argument

above. Knowing the landmark, it is much easier for the language model to predict the relation (of either kinds) and the target.

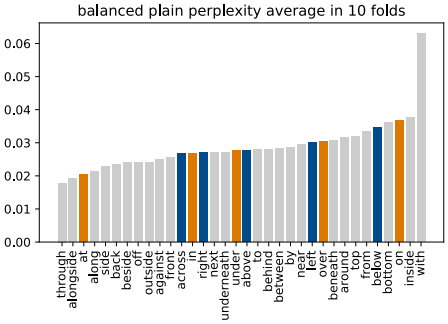


Figure. 6.4: Mean perplexities of LM2′ by spatial relation (orange: functionally biased, blue: geometrically biased)

In conclusion, the explanation why descriptions with functionally-biased relations have a higher perplexity than descriptions with geometrically-biased descriptions appears to be twofold: (i) functionally-biased relations are more selective of their targets as expressed by the uniqueness counts, and (ii) functional relations are also more selective of their landmarks but this fact works against the performance of the language model. As it is trained on the sequence left to right, it has to learn to predict relations only on the basis of targets which in the case of functionally-biased relations are represented by more unique tokens than geometrically-biased relations. The more informative words, the landmarks, that would enable the language model to predict a functional relation, comes last, after the relation has already been seen. The possible reason why geometrically-biased relations lead to lower perplexities of a language model on descriptions is because they have fewer unique targets. Hence, our Hypothesis 1 which linked selectivity of functionally-biased relations to low perplexity of phrases can be refuted. In spatial relations the order of the semantic interpretation of tokens (that we want to capture in these experiments) is different from the linear syntactic order of order which can be captured by the language model. When this order is changed as in LM2′ our predictions come closer to the hypothesis (Figure 6.4).[5]

[5]Modulo that landmarks are, as discussed above, well-predictive of relations of both kinds.

By removing the frequency bias on spatial relations in LM2 we fix the distribution of spatial relations and examine the effect of distribution of targets and landmarks on perplexities of phrases (spatial relation as fixed context). In the following section, we fix the distributions of targets and landmarks of each spatial relation and examine the perplexity of phrases when another spatial relation is projected in this context (targets-landmarks as fixed context).

6.5 Varying spatial relations

6.5.1 Hypotheses

Given a particular spatial relation, the distribution of targets and landmarks that occur with it creates a particular signature of targets and landmarks, the target-landmark context of a spatial relation. In this experiment, we investigate the effect on perplexity of phrases when another spatial relation is projected in such a target-landmark context. Given different selectivity of functionally- and geometrically-biased spatial relations, namely the functionally-based spatial relations are more selective of their targets and landmarks and therefore create more specific contexts, we should observe differences in perplexities of phrases when other spatial relations are projected in these contexts. In particular, we hypothesise that geometrically-biased spatial relations are more easily swappable than functionally-biased spatial relations as measured by the perplexity of a language model trained on the original, non-swapped phrases (Hypothesis 4).

6.5.2 Method

We use LM2 from Section 6.4 (trained on the balanced frequencies of spatial relations) with no additional training from the previous experiment. We group descriptions in the evaluation set by spatial relation. For each phrase containing a particular spatial relation, we replace it with every other spatial relation and estimate the perplexity of the resulting phrase using a language model. Finally, we calculate the mean of perplexities over all phrases. We use 10-fold cross-validation and report the final means across the 10 folds.

6.5.3 Results

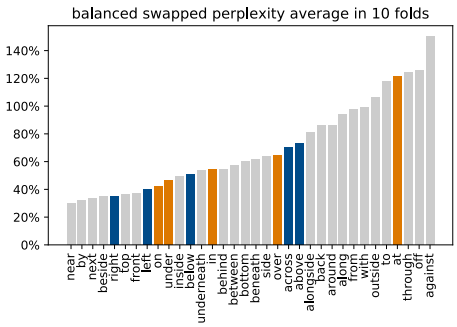


Figure. 6.5: %-increase in perplexities of LM2 shown per context of the original preposition when swapped with another one.

Figure 6.5 shows a %-increase in mean complexities from those in Figure 6.2 when LM2 is applied on phrases with swapped relations in the contexts of the original relations. Hence, the column "at" shows the %-increase in perplexities of phrases that originally contained *at* in the validation dataset but this was replaced by all other spatial relations. Comparing with Figure 6.2 the estimated perplexities are higher across all cases which is predictable. There is a weak tendency that replacing functionally-biased relations with other relations leads to higher perplexities of spatial descriptions than replacing geometrically-biased relations, but the distinction is not clear cut (Hypothesis 4 partially confirmed). The lack of a clear distinction between two classes of descriptions confirms our previous observations about landmarks and targets: the LM has learned particular contexts for both kinds of descriptions.

6.6 Discussion and conclusion

We explored the degree that the functional and geometric character of spatial relations can be identified by a neural language model by focusing on spatial descriptions of controlled length and containing normalised relations. Our first question was about the implications of using a neural language model for this task. The previous research Dobnik and Kelleher (2013) used normalised entropy of target-landmarks per relation and log likelihood

ratio between target-landmarks and relations to test this. These are focused measures that estimate the variation and the strength of association of words in a corpus. On the other hand, a language model provides a more general probabilistic representation of the entire description. As such it captures any kind of associations between words in a sequence. The other important observation is that it captures sequential relations in the direction left to right and as we have seen the sequential nature of the language model does not correspond precisely with the order in which semantic arguments of spatial relations are interpreted. However, nonetheless we can say that language models are able to capture a distinction between functional and geometric spatial relations (plus other semantic distinctions) to a similar degree of success as previously reported measures. Our initial hypothesis about the greater selectivity of spatial relations for its arguments is correct but it is exemplified in a greater perplexity of a language model in the context of balanced spatial relations. We argued that this has to do with the fact that the targets are more unique to these relations (which is consequence of a greater specificity for arguments of functionally biased relations) and is also related to the way a sequential language model works. In the unbalanced dataset, the perplexity of the language model is reversed (it is lower with functionally biased relations) because the specificity of targets to relations is boosted with greater frequency of functionally-biased relations. The fact that functionally-biased relations are more frequent is probably related to the fact that such descriptions are more informative than purely geometric ones if available for a particular pair of objects.

We can only report tendencies based on the perplexities of our language models as our conclusions. This is because the functional-geometric bias is graded, because the predictions are highly dependent on the quality and the size of the dataset, and because other semantic relations might also be expressed by this measure. We chose a large contemporary dataset of image descriptions because we hope that it contains a high proportion of prepositions used as spatial relations. However, there is no guarantee that all prepositions in this dataset are used this way. We observe that there is considerable variation of obtained values across the 10-folds of cross-validation and we report the mean values over all folds. As an illustration, in the supplementary material (Section 6.7) we give an example of graphs from two intermediary folds.

Using a language model in this task we have also learned new insights about the way language models encode spatial relations in image descriptions. It has been pointed out (cf. Kelleher and Dobnik (2017) among others) that convolutional neural networks with an attention model are designed to detect objects whereas spatial relations between objects are likely to be predicted by the language model. In this work we show that language models are not only predicting the relation (which is expected) but are able to distinguish between different classes of relations thus encoding finer semantic distinctions. This tells us that language models are able to encode a surprising amount of information about world knowledge with a usual caveat that it is difficult to separate several strands of this knowledge.

The work can be extended in several ways. One way is to study dataset effects on the predicted results. Datasets with descriptions of robotic actions and instructions may be particularly promising as they focus on spatial uses. Different normalisations of spatial relations have a significant effect on the results. In this work composite spatial relations such *on the left side of* are normalised to simple spatial relations such as *left*. However, these could be treated as separate relations as difference between may exist. A more systematic examination of clusters of spatial relations would eventually tell us what other spatial relations not yet identified as functionally or geometrically biased have similar properties to those that have identified as such experimentally.

Acknowledgements

The research of Dobnik and Ghanimifard was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg.

The research of Kelleher was supported by the ADAPT Research Centre. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Funds.

6.7 Appendix: Supplementary material

Language Model Architecture

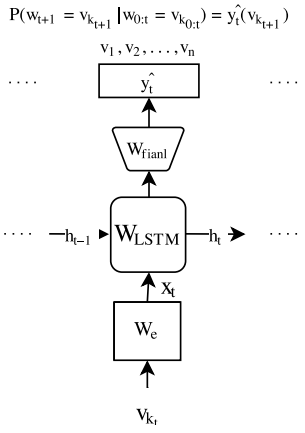


Figure. 6.6: The recurrent language model diagram with LSTM recurrent unit.

The neural language model architecture with the Long-Short Terms Memory (LSTM) function and its parameters, similar to tied weights in Gal and Ghahramani (2016):

- $W_e \in R^{n \times d}$ for word embeddings,
- $W_{LSTM} \in R^{2d \times 4d}$ for parameters of the Long-Short Term Memory function,
- $W_{Final} \in R^{d \times n}$ of the final dense layer with *softmax*.

where n is the vocabulary size for $V = \{v_1, v_2, ..., v_n\}$ and d is both the embeddings size and the memory size in LSTM. For mini-batches from

training data, these parameters are being updated using a stochastic gradient descent to minimise the loss.

$$x_t = \delta_{v_{k_t}} \cdot W_e \tag{6.5}$$

$$\begin{pmatrix} \mathrm{i} \\ \mathrm{f} \\ \mathrm{o} \\ \mathrm{g} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(\begin{pmatrix} \mathrm{x}_t \\ \mathrm{h}_{t-1} \end{pmatrix} \cdot \mathrm{W}_{LSTM} \right) \tag{6.6}$$

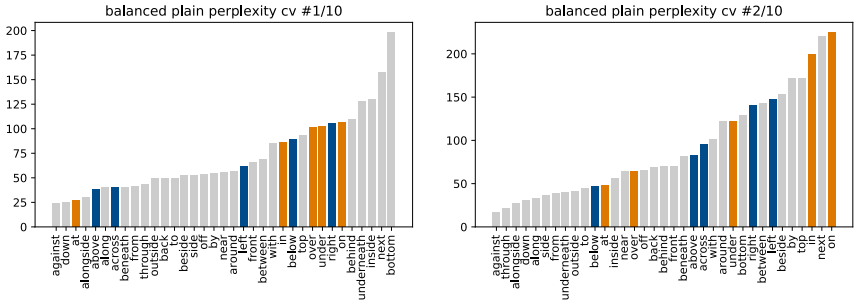
$$c_t = \mathrm{f} \circ c_{t-1} + \mathrm{i} \circ \mathrm{g} \tag{6.7}$$

$$h_t = \mathrm{o} \circ \tanh(c_t) \tag{6.8}$$

$$\hat{y}_t = \mathrm{softmax}(h_t \cdot \mathrm{W}_{final} + b) \tag{6.9}$$

where $\delta_{v_{k_t}}$ represents the one-hot encoding of the t-th word in the sequence. The x_t is the word embedding for this word, and two vectors c_t and h_t represent the states of the recurrent unit. Figure 6.6 illustrates the same equation.

Evaluation



(a) (b)

Figure. 6.7: Mean perplexities of LM2 by spatial relation for (a) folds 1 and (b) 2 (orange: functionally biased, blue: geometrically biased).

Bibliography

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

François Chollet et al. 2015. Keras. https://github.com/keras-team/keras.

Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. 2016. Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. *arXiv*, 1604.02125 [cs.CV].

Kenny R Coventry and Simon C Garrod. 2004. *Saying, seeing, and acting: the psychological semantics of spatial prepositions*. Psychology Press, Hove, East Sussex.

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of Memory and Language*, 44(3):376–398.

Sebastien Delecraz, Alexis Nasr, Frederic Bechet, and Benoit Favre. 2017. Correcting prepositional phrase attachments using multimodal corpora. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 72–77, Pisa, Italy. Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pages 1–6, Berlin, Germany.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

Simon Garrod, Gillian Ferrier, and Siobhan Campbell. 1999. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72(2):167–189.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE.

Michael Grubinger, Paul D. Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR benchmark: A new evaluation resource for visual information systems. In *Proceedings of OntoImage 2006: Workshop on language resources for content-based mage retrieval during LREC 2006*, Genoa, Italy. European Language Resources Association.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

John D. Kelleher and Fintan J. Costello. 2005. Cognitive representations of project prepositions. In *In Proceedings of the Second ACL-Sigsem Workshop on The Linguistic Dimensions of Prepositions and their Used In Computational Linguistic Formalisems and Applications*.

John D. Kelleher and Fintan J. Costello. 2009. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017), Gothenburg, 12 –13 June*, volume 1 of *CLASP Papers in Computational Linguistics*, pages 41–52, Gothenburg, Sweden.

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 1041–1048. Association for Computational Linguistics.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Parisa Kordjamshidi, Taher Rahgooy, Marie-Francine Moens, James Pustejovsky, Umar Manzoor, and Kirk Roberts. 2017. CLEF 2017: Multimodal spatial role labeling (mSpRL) task overview. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 367–376, Cham. Springer International Publishing.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Transactions on Speech and Language Processing*, 8(3):4:1–4:36.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Barbara Landau. 1996. Multiple geometric representations of objects in languages and language learners. *Language and space*, pages 317–363.

Barbara Landau. 2016. Update on "What" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive Science*, 41(S2):321–350.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.

G.D. Logan and D.D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In M. Bloom, P.and Peterson, L. Nadell, and M. Garrett, editors, *Language and Space*, pages 493–529. MIT Press.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220, Lisbon, Portugal. Association for Computational Linguistics.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT 2010 Workshop on creating speech and language data with Amazon's Mechanical Turk*, Los Angeles, CA. North American Chapter of the Association for Computational Linguistics (NAACL).

Gerard J Steen, Aletta G Dorst, J Berenike Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing.

7 Representation Of Spatial Relations In Neural Language Models

> **Mehdi Ghanimifard and Simon Dobnik.** *What* a neural language model tells us about spatial relations, *In Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP), pp. 71-81. 2019.*

Abstract Understanding and generating spatial descriptions requires knowledge about *what* objects are related, their functional interactions, and *where* the objects are geometrically located. Different spatial relations have different functional and geometric bias. The wide usage of neural language models in different areas including generation of image description motivates the study of what kind of knowledge is encoded in neural language models about individual spatial relations. With the premise that the functional bias of relations is expressed in their word distributions, we construct multi-word distributional vector representations and show that these representations perform well on intrinsic semantic reasoning tasks, thus confirming our premise. A comparison of our vector representations to human semantic judgments indicates that different bias (functional or geometric) is captured in different data collection tasks which suggests that the contribution of the two meaning modalities is dynamic, related to the context of the task.

7.1 Introduction

Spatial descriptions such as “the chair is to the left of the table” contain spatial relations “to the left of” the semantic representations of which must

be grounded in visual representations in terms of geometry Harnad (1990). The apprehension of spatial relations in terms of scene geometry has been investigated through acceptability scores of human judges over possible locations of objects Logan and Sadler (1996). In addition, other research has pointed out that there is an interplay between geometry and object-specific function in the apprehension of spatial relations Coventry et al. (2001). Therefore, spatial descriptions must be grounded in two kinds of knowledge Landau and Jackendoff (1993); Coventry et al. (2001); Coventry and Garrod (2004); Landau (2016). One kind of knowledge is referential meaning, expressed in the geometry of scenes (geometric knowledge or *where* objects are) while the other kind of knowledge is higher- level conceptual world knowledge about interactions between objects which is not directly grounded in perceivable situations but is learned through our experience of situations in the world (functional knowledge or *what* objects are related). Furthermore, Coventry et al. (2001) argue that individual relations have a particular geometric and functional bias and "*under*" and "*over*" are more functionally-biased than "*below*" and "*above*". For instance, when describing the relation between a person and an umbrella in a scene with a textual context such as "$an\ umbrella\ \_\_\_\ a\ person$", "*above*" is associated with stricter geometric properties compared to "*over*" which covers a more object-specific extra-geometric sense between the target and the landmark (i.e. *covering* or *protecting* in this case). Of course, there will be several configurations of objects that could be described either with "*over*" or "*above*" which indicates that the choice of a description is determined by the speaker, in particular what aspect of meaning they want to emphasise. Coventry et al. (2001) consider this bias for prepositions that are geometrically similar and therefore the functional knowledge is reflected in different preferences for objects that are related. However, such functional differences also exist between geometrically different relations.

This poses two interesting research questions for computational modelling of spatial language. The first one is how both kinds of knowledge interact with individual spatial relations and how models of spatial language can be constructed and learned within end-to-end deep learning paradigm. Ramisa et al. (2015) compare the performance of classifiers using different multi-modal features (visual, geometric and textual) to predict a spatial preposition. Schwering (2007) applies semantic similarity metrics of spatial relations on geographical data retrieval. Collell et al. (2018) show that word embeddings

can be used as predictive features for common sense knowledge about location of objects in 2D images. The second question is related to the extraction of functional knowledge for applications such as generation of spatial descriptions in a robot scenario. Typically, a robot will not be able to observe all object interactions as in (Coventry et al., 2004) to learn about the interaction of objects and choose the appropriate relation. Following the intuition that the functional bias of spatial relations is reflected in a greater selectivity for their target and landmark objects, Dobnik and Kelleher (2013, 2014) propose that the degree of association between relations and objects in the corpus of image descriptions can be used as filters for selecting the most applicable relation for a pair of objects. They also demonstrate that entropy-based analysis of the targets and landmarks can identify the functional and geometric bias of spatial relations. They use descriptions from a corpus of image descriptions because here the prepositions in spatial relations are used mainly in the spatial sense. The same investigation of textual corpora such as BNC Consortium et al. (2007) does not yield such results as there prepositions are used mainly in their non-spatial sense.[1] Similarly, Dobnik et al. (2018) inspect the perplexity of recurrent language models for different descriptions containing spatial relations in the Visual Genome dataset of image captions Krishna et al. (2017) in order to investigate their bias for objects.

In this paper, we follow this line of work and (i) further investigate what semantics about spatial relations are captured from descriptions of images by generative recurrent neural language models, and (ii) whether such knowledge can be extracted, for example as vector representations, and evaluated in tests. The neural embeddings are opaque to interpretations per se. The benefit of using recurrent language models is that they allow us to (i) deal with spatial relations as multi-word expressions and (ii) they learn their representations within their contexts:

(a) *a cat on a mat*
(b) *a cat on the top of a mat*
(c) *a mat under a cat*

In (a) and (b), the textual contexts are the same "*a cat* ___ *a mat*" but the meaning of the spatial relations, one of which is a multi-word expression,

[1] We may call this metaphoric or highly functional usage which is completely absent of the geometric dimension.

are slightly different. In (c) the context is made different through word order.

The question of what knowledge (functional or geometric) should be represented in the models can be explained in information-theoretic terms. The low surprisal of a textual language model on a new text corpora is an indication that the model has encoded the same information content as the text. In the absence of the geometric knowledge during the training of the model, this means that a language model encodes the relevant functional knowledge. We will show that the degree to which each spatial description containing a spatial relation encodes functional knowledge in different contexts can be used as source for building distributional representations. We evaluate these representations intrinsically in reasoning tests and extrinsically against human performance and human judgment.

The contributions of this paper are:

1. It is an investigation of the semantic knowledge about spatial relations learned from textual features in recurrent language models with intrinsic and extrinsic methods of evaluation on internal representations.
2. It proposes a method of inspecting contextual performance of generative neural language models over a wide categories of contexts.

This paper is organised as follows: in Section 7.2 we describe how we create distributional representations with recurrent neural language models, in Section 7.3 we describe our computational implementations that build these representations, and in Section 7.4 we provide their evaluation. In Section 7.5 we give our final remarks.

7.2 Neural representations of spatial relations

Distributional semantic models produce vector representations which capture latent meanings hidden in association of words in documents Church and Hanks (1990); Turney and Pantel (2010). The neural word embeddings were initially introduced as a component of neural language models Bengio et al. (2003). However, subsequently neural language models such as word2vec Mikolov et al. (2013) and GloVe Pennington et al. (2014) have

become used to specifically learn word embeddings from large corpora. The word embeddings trained by these models capture world-knowledge regularities expressed in language by learning from the distribution of context words which can be used for analogical reasoning[2]. Moreover, sense embeddings Neelakantan et al. (2014) and contextual embeddings Peters et al. (2018) have shown to provide fine-grained representation which can discriminate between different word senses or contexts, for example in substituting synonym words and multi-words in sentences McCarthy and Navigli (2007).

However, meaning is also captured by generative recurrent neural language models used to generate text rather than predict word similarity. The focus of our work is to investigate what semantics about spatial relations is captured by these models. Generative language models use the chain rule of probability for step-by-step prediction of the next word in a sequence. In these models, the probability of a sequence of words (or sometimes characters) is defined as the multiplication of conditional probabilities of each word given the previous context in a sequence:

$$P(w_{1:T}) = \prod_{t=1}^{T-1} P(w_{t+1}|w_{1:t}) \tag{7.1}$$

where T is the length of the word sequence. The language model estimates the probability of a sequence in Equation (7.1) by optimising parameters of a neural network trained over sufficient data. The internal learned parameters includes embeddings for each word token which can be used as word level representations directly.

An alternative way of extracting semantic prediction from a generative neural language model which we are going to explore in this paper is to measure the fidelity of the model's output predictions against a new ground truth sequence of words. This is expressed in the measure of *Perplexity* as follows:

$$PP(S) = (\prod_{s \in S} P(w_{1:t} = s))^{\frac{-1}{|S|}} \tag{7.2}$$

[2]For example, "a is to a^* as b is to b^*" can be queried with simple vector arithmetic $king - man + woman \approx queen$. More specifically, with a search over vocabulary with cosine similarity: $\underset{b^* \in V/\{a^*,b,a\}}{argmax} \; cos(b^*, a^* - a + b)$

where S is a collection of ground truth sentences. Perplexity is a measure of the difficulty of a generation task which is based on the information theoretic concept of entropy Bahl et al. (1983). It is based on *cross-entropy* which takes into account the probability of a sequence of words in ground truth sentences and the probability of a language model generating that sequence. It is often used for intrinsic evaluation of word- error rates in NLP tasks Chen et al. (1998). However, in this paper we use perplexity as a measure of fit of a pre-trained generative neural language model to a collection of sentences.

Our proposal is as follows. We start with the hypothesis that in spatial descriptions some spatial relations (those that we call functional) are more predictable from the associated word contexts of targets and landmarks than their grounding in the visual features. Hence, this will be reflected in a perplexity of a (text-based) generative language model trained on spatial descriptions. Descriptions with functionally-biased spatial relations will be easier to predict by this language model than geometrically-biased spatial descriptions and will therefore have lower perplexity. If two sequences of words where only the spatial relations differ (but target and landmark contexts as well as other words are the same) have similar perplexity, it means that such spatial relations have similar selectional requirements and are therefore similar in terms of functional and geometric bias. We can exploit this to create vector representations for spatial relations as follows. Using a dictionary of spatial relations, we extract collections of sentences containing a particular spatial relation from a held-out dataset not used in training of the language model. The collection of sentences with a particular spatial relation are our context templates. More specifically, for our list of spatial relations $\{r_1, r_2, ..., r_k\}$, we replace the original relation r_i with a target relation r_j in its collection of sentences, e.g. we replace *to the right of*$_i$ with *in front of*$_j$. The outcome is a collection of artificial sentences $S_{i \to j}$ that are identical to the human-generated sentences except that they contain a substituted spatial relation. The perplexity of the language model on these sentences represents the association between the original spatial relation and the context in which this has been projected:

$$PP(S_{i \to j}) = PP_{i,j} = P(rel_i, c_{rel_j})^{\frac{1}{-N'}} \tag{7.3}$$

where c_{rel_j} is the context of rel_i, and $PP_{i,j}$ is the perplexity of the neural language model on the sentence collection where relation rel_i is artificially placed in the contexts of relation rel_j. If rel_i and rel_j are associated with similar contexts, then we expect low perplexity for $S_{i \to j}$, otherwise the perplexity will be high. Finally, the perplexity of rel_i against each collection c_{rel_j} is computed and normalised within each collection (Equation 7.4) and the resulting vector per rel_i over all contexts is represented as a unit vector (Equation 7.5).

$$m_{i,j} = \frac{PP_{i,j}}{\sum_{i'=1}^{k} PP_{i',j}} \tag{7.4}$$

$$\hat{v_i} = \frac{v_i}{||v_i||} \qquad v_i = (m_{i,1}, ..., m_{i,k})^T \tag{7.5}$$

where $\hat{v_i}$ is the vector representation of the relation rel_i. These vectors create a matrix. In a particular cell of some row and some column, high perplexity means that the spatial relation in that row is less swappable with the context in the column, while a low perplexity means that the spatial relation is highly swappable with that context. This provides a measure similar to mutual information (PPMI) in traditional distributional vectors Church and Hanks (1990).

In conclusion, representing multi-word spatial relations in a perplexity matrix of different contexts allows us to capture their semantics based on the predictions and the discriminatory power of the language model. If all spatial relations are equally predictable from the language model such vector representations will be identical and vector space norms will not be able to discriminate between different spatial relations. In the following sections we report on the practical details how we build the matrix (Section 7.3) and evaluate it on some typical semantic tasks (Section 7.4). The implementation and evaluation code: `https://github.com/GU-CLASP/what_nlm_srels`

7.3 Dataset and models

7.3.1 Corpus and pre-processing

We use Visual Genome region description corpus Krishna et al. (2017). This corpus contains 5.4 million descriptions of 108 thousand images, collected

from different annotators who described specific regions of each image. As stated earlier, the reason why we use a dataset of image descriptions is because we want to have spatial usages of prepositions. Other image captioning datasets such as MSCOCO Lin et al. (2014) and Flickr30k Plummer et al. (2015) could also be used. However, our investigation has shown that since the task in these datasets in not to describe directly the relation between selected regions, common geometric spatial relations are almost missing in them: there are less then 30 examples for "*left of*" and "*right of*" in these datasets.

After word tokenisation with the space operator, we apply pre-processing which removes repeated descriptions per-image and also descriptions that include uncommon words with frequency less than 100[3] Then we split the sentences into 90%-10% portions. The 90% is used for training the language model (Section 7.3.2), and 10% is used for generating the perplexity vectors by extracting sentences with spatial relations that represent our context bins (Section 7.3.3). The context bins are used for generating artificial descriptions $S_{i \to j}$ on which the language model is evaluated for perplexity.

7.3.2 Language model and GloVe embeddings

We train a generative neural language model on the 90% of the extracted corpus (Section 7.3.1) which amounts to 4,537,836 descriptions of maximum length of 29 and 4,985 words in the vocabulary. We implement a recurrent language model with LSTM Hochreiter and Schmidhuber (1997) and a word embeddings layer similar to Gal and Ghahramani (2016) in Keras Chollet et al. (2015) with TensorFlow Abadi et al. (2015) as back-end. The Adam optimiser Kingma and Ba (2014) is used for fitting the parameters. The model is set up with 300 dimensions both for the embedding- and the LSTM units. It is trained for 20 epochs with a batch size of 1024.

In addition to the generative LSTM language model, we also train on the same corpus GloVe (VG) embeddings with 300 dimensions and a context-

[3] The pre-processing leaves 5,042,039 descriptions in the corpus with maximum 31 tokens per sentence. The relatively high threshold of 100 tokens is chosen to insure sufficient support in the 10% of held-out data for bucketing. We did not use OOV tokens because the goal of the evaluation is to capture object-specific properties about spatial relations and OOV tokens would interfere with this.

window of 5 words. Finally, we also use pre-trained GloVe embeddings on the Common Crawl (CC) dataset with 42B tokens[4].

7.3.3 Perplexity vectors

Based on the lists of spatial prepositions in (Landau, 1996) and (Herskovits, 1986), we have created a dictionary of spatial relations which include single word relations as well as all of their possible multi-word variants. This dictionary was applied on the 10% held-out dataset where we found 67 single- and multi-word spatial relation types in total. As their frequency may have fallen below 100 words due to the dataset split, we further remove all relations below this threshold which gives us 57 relations. We also create another list of relations where composite variants such as "to the left of" and "on the left of" are grouped together as "left" which contains 44 broad relations. We group the sentences by the relation they are containing to our context bins using simple pattern matching on strings. Table 7.1 contains some examples of our context bins. The bins are used for artificial sentence generation as explained in the previous section.

Relation (rel_i)	Context bin (c_{rel_i})
above	scissors _____ the pen tall building _____ the bridge ...
below	pen is _____ scissors bench _____ the green trees ...
next to	a ball-pen _____ the scissors car _____ the water ...

Table. 7.1: Examples of context bins based on extracted descriptions from Visual Genome. The images that belong to these descriptions are shown in Appendix B.

For each of the 67 spatial relations extracted from the larger corpus, there are 57 collections of sentences (=the number of relations in the smaller corpus). Hence, there are $3,819 (= 67 \times 57)$ possible projections $S_{i \to j}$, where a relation i is placed in the context j, including the case where there is no swapping of relations when $j = i$. The process is shown in Figure 7.1. The

[4] http://nlp.stanford.edu/data/glove.42B.300d.zip

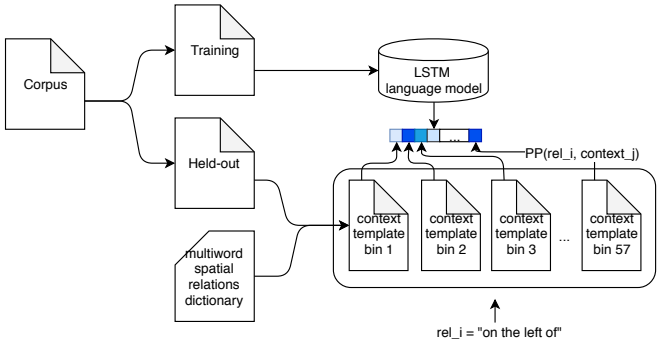


Figure. 7.1: Generating perplexity-based vectors for each spatial relation.

vector of resulting perplexities in different contexts is normalised according to Equation 7.5 which gives us perplexity vectors (P-vectors) as shown in Figure 7.2.

In addition to the P-vectors we also create representations learned by the word embedding layer in the generative language model that we train. For each of the 44 broad single-word spatial relations we extract a 300-dimensional embedding vector from the pre-trained recurrent language model (LM-vectors). In order to produce LM-vectors for the multi-word spatial relations, we simply sum the embeddings of the individual words. For example the embedding vector for "to the left of" is $v_{to} + v_{the} + v_{left} + v_{of}$. The same method is also used for the GloVe embeddings.

7.3.4 Human judgments

In order to evaluate our word representations we compare them to three sources of human judgments. The first one are judgments about the the fit of each spatial relation over different geometric locations of a target object in relation to a landmark which can be represented as spatial templates Logan and Sadler (1996). The second are 88,000 word association judgments by English speakers from De Deyne et al. (2018). In each instance participants were presented a stimulus word and were asked to provide 3 other words. The dataset contains 4 million responses on 12,000 cues. Based on the collective performance of annotators, the dataset provides association strengths between words (which contain any kind of words, not just spatial words) as a measure of their semantic relatedness. Finally, we collected a new dataset

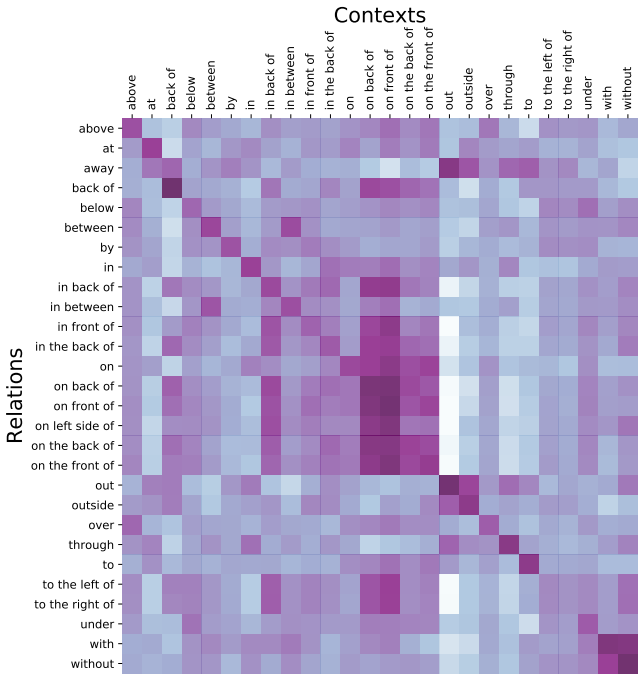


Figure. 7.2: A matrix of perplexity vectors for 28 spatial relations and 26 contexts. For the full 67×57 matrix see Appendix C. The rows represent spatial relations and columns represent the normalised average perplexity of a language model when this relation is swapped in that context.

of word similarity judgments using Amazon Mechanical Turk. Here, the participants were presented with a pair of spatial relations at a time. Their task was to use a slider bar with a numerical indicator to express how similar the pair of words are. The experiment is similar to the one described in Logan and Sadler (1996) except that in our case participants only saw one pair of relations at a time rather than the entire list. The shared vocabulary between these three datasets covers *left*, *right*, *above*, *over*, *below*, *under*, *near*, *next*, *away*.

7.4 Evaluation

As stated in Section 7.2 the P-vectors we have built are intended to capture the discriminatory power of a generative language model to encode and

1. to	18. up; down; off
2. on	19. with; without
3. away	20. together; out
4. here	21. outside; inside
5. into	22. near; beside; by
6. from	23. top; front; bottom
7. during	24. in between; between
8. back of	25. along; at; across; around
9. through	26. beneath; below; under; behind
10. alongside	27. right; back; left; side; there
11. along side	28. to the left of; to the right of; next to
12. underneath	29. in back of; in the back of; on the back of; at the top of
13. in; against	30. on the top of; on side of; on the bottom of; on left side of; on top of; on the front of; on back of; on the side of; on front of; on bottom of
14. in front of	
15. above; over	
16. to the side	
17. onto; toward	

Table. 7.2: K-means clusters of spatial relations based on their P-vectors.

discriminate different spatial relations, their functional bias. In this section we evaluate the P-vectors on several common intrinsic and extrinsic tests for vectors. If successful, this demonstrates that such knowledge has indeed been captured by the language model. We evaluate both single- and multi-word relations.

7.4.1 Clustering

Method Figure 7.2 and its complete version in Appendix C show that different spatial relations have different context fingerprints. To find similar relations in this matrix we can use *K-means clustering*. K-mean is a non-convex problem: different random initialisation may lead to different local minima. We apply the clustering on 67 P-vectors for multi-word spatial relations and qualitatively examine them for various sizes k. The optimal number of clusters is not so relevant here, only that for each k we get reasonable associations that follow our semantic intuitions.

Results As shown in Table 7.2, with $k = 30$, the clustering of perplexity vectors shows acceptable semantics of each cluster. There are clusters with synonymous terms such as (15. *above*, *over*) or (26. *below*, *under*). Some clusters have variants of multi-word antonymous such as (30. *on the top of*, *on the bottom of*). Other clusters have a mixture of such relations, e.g. (27. *right*, *back*, *left*, *side*, and *there*).

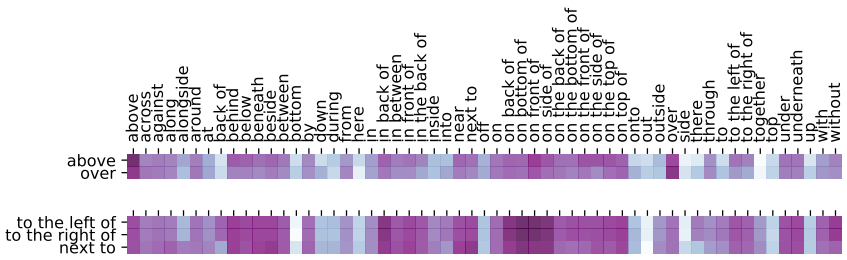


Figure. 7.3: The P-vectors of two clusters.

Discussion The inspection of the perplexities of two of these clusters in Figure 7.3 shows that the language model has learned different selectional properties of spatial relations: *above* and *over* are generally more selective of their own contexts, while *to the left of* and *to the right of* show a higher degree of confusion with a variety of the P-vector contexts. High degree of confusion in *left* and *right* is consistent with the observation in Dobnik and Kelleher (2013) that these relations are less dependent on the functional relation between particular objects and therefore have a higher geometric bias. On the other hand, *above* and *over* seem to be more selective of their contexts. The functional distinction between *above* and *over* is mildly visible: the shades of blue in *above* are slightly darker than *over*.

7.4.2 Analogical reasoning with relations

The intrinsic properties of vector representations (the degree to which they capture functional associations between relations and their objects) can be tested with their performance in analogical reasoning tasks. We compare the performance of the P-vectors (Section 7.3.3), the embeddings of the language model used to create the P-vectors and GloVe embeddings (Section 7.3.2) in two analogical tasks which require both geometric and functional reasoning.

Predicting analogical words

Method The task is similar to the analogy test Mikolov et al. (2013); Levy et al. (2015) where two pairs of words are compared in terms of some relation "a is to a' as b is to b'". We manually grouped spatial relations that are opposite in one geometric dimension to 6 groups. These are: Group

	Single word	Multi-words
GloVe (CC)	0.56	0.36
GloVe (VG)	0.43	0.29
LM	0.86	0.45
P-vectors	0.62	0.47
Random	0.11	0.05

Table. 7.3: The accuracies of different representations on the word analogy test.

1: left, right; Group 2: above, below; Group 3: front, back; Group 4: with, without; Group 5: in, out; and Group 6: up, down. We generate all possible permutations of these words for the analogical reasoning task which gives us 120 permutations. We expand these combinations to include multi-word variants. This dataset has 85,744 possible analogical questions such as (*above* :: *below*, *to the left of* :: ?). We accept all variants of a particular relation (e.g. *to the right side of* and *to the right of*) as the correct answer.

Results As shown in in Table 7.3, on the single-word test suite, the LM-embeddings perform better than other models. On multi-word test suite the P-vectors perform slightly better. On both test suites, GloVe trained on Common Crawl performs better than GloVe trained on Visual Genome. However, its performance on multi-word relations is considerably lower. We simulated random answers as a baseline to estimate the difficulty of the task. Although the multi-word test suite has ~ 700 times more questions than the test suite with single-word relations, it is only approximately 2-times more difficult to predict the correct answer in the multi-word dataset compared to the single-word dataset.

Discussion The perplexity of the language model on complete context phrases (Multi-words) is as good indicator of semantic relatedness as the word embeddings of the underlying language model and much better than GloVe embeddings. The good performance of the P-vectors explains the errors of the language model in generating spatial descriptions. The confusion between *in front of* and *on the back of* is similar to the confusion between *to the left of* and *to the right of* in terms of their distribution over functional contexts. Hence, a similar lack of strong functional associations allows the vectors to make inference about geometrically related word-pairs. This indicates that functional and geometric bias of words are complementary. There are two possible explanations why P-vectors perform better than

LM-embeddings on multi-word vectors: (i) low-dimensions of P-vectors (57D) intensify the contribution of spatial contexts for analogical reasoning compared to high-dimensional LM-embeddings (300D); (ii) summing the vectors of the LM-embeddings for multi-words reduces their discriminatory effect.

Odd-one-out

Method Based on the semantic relatedness of words, the goal of this task is to find the odd member of the three. The ground truth for this test are the following five categories of spatial relations, again primarily based on geometric criteria: X-axis: left, right; Y-axis: above, over, under, below; Z-axis: front, back; Containment: in, out; and Proximity: near, away. Only the Y-axis contains words that are geometrically similar but functionally different, e.g. *above/over*. In total there are 528 possible instances with 3,456 multi-word variations. The difficulty of the task is the same for both single- and multi-word expressions as the choice is always between three words. Hence, the random baseline is 0.33.

Results Table 7.4 shows the accuracy in predicting the odd relation out of the three. We also add a comparison to fully geometric representations captured by spatial templates Logan and Sadler (1996). Ghanimifard and Dobnik (2017) show that spatial templates can be compared with Spearman's rank correlation coefficient $\rho_{X,Y}$ and therefore we also include this similarity measure. Since our groups of relations contain those that are geometric opposites in each dimension, we take the absolute value of $|\rho_{X,Y}|$. Spatial templates are not able to recognise relatedness without the right distance measure, $|\rho_{X,Y}|$. LM-embeddings perform better than other vectors in both tests, but P-vectors follow closely. All models have a low performance on the multi-word test suite. When using $|\rho_{X,Y}|$ all vectors other than P-vectors produce better results. While we do not have an explanation for this, it is interesting to observe that $|\rho_{X,Y}|$ is a better measure of similarity than cosine.

Discussion The results demonstrate that using functional representations based on associations of words can predict considerable information about geometric distinctions between relations, e.g. distinguishing *to the right of* and *above*, and this is also true for P-vectors. As stated earlier, our explanation

	Single word		Multi-words	
	$1-cos$	$\|\rho\|$	$1-cos$	$\|\rho\|$
GloVe (CC)	0.62	0.68	0.52	0.58
GloVe (VG)	0.61	0.61	0.58	0.59
LM	0.87	0.90	0.82	0.88
P-vectors	0.72	0.70	0.64	0.52
Sp Templates	0.22	1.0	-	-

Table. 7.4: The accuracies in odd-one-out tests.

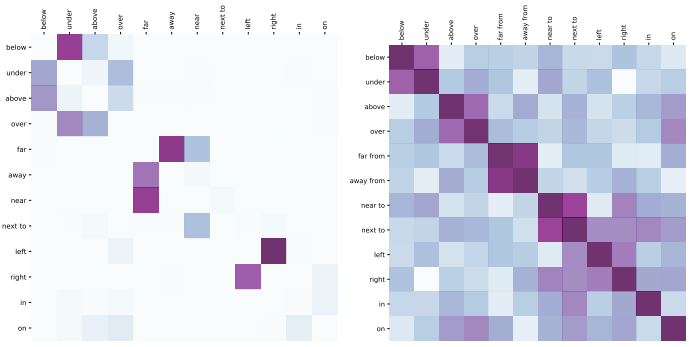
for this is that functional and geometric knowledge is in complementary distribution. This has positive and negative implications for joint vision and language models used in generating spatial descriptions. In the absence of geometric information, language models provide strong discriminative power in terms of functional contexts, but even if geometric latent information is expressed in them, an image captioning system still needs to ground each description in the scene geometry.

7.4.3 Similarity with human judgments

We compare the cosine similarity between words in LM- and P-vector spaces with similarities from (i) word association judgments De Deyne et al. (2018), (ii) our word similarity judgments from AMT, and (iii) spatial templates (Section 7.3.4). We take the maximum subset of shared vocabulary between them, including *on*, *in* only shared between (i) and (ii). Since (i) is an association test, unrelated relations do not have association strengths. There are 55 total possible pairs of 11 words, while only 28 pairs are present in (i) as shown in Figure 7.4.

Method We take the average of the two way association strengths if the association exists and for (i) we assign a zero association for unrelated pairs such as *left* and *above*. Spearman's rank correlation coefficient $\rho_{X,Y}$ is used to compare the calculated similarities.

Results Table 7.5 shows ranked correlations of different similarity measures. Spatial templates do not correlate with (WA) word associations and (WS) word similarities. On 28 pairs there is a weak negative correlation between spatial templates and WS. The correlation of similarities of two



(a) (b)

Figure. 7.4: (i) Word association judgments and (ii) word similarity judgments

different human judgments is positive but weak ($\rho = 0.33$). The similarities predicted by LM-vectors and P-vectors correlate better with WA than WS.

	55 pairs		28 pairs	
	WA	WS	WA	WS
SpTemp	-0.02	-0.08	0.06	-0.35
LM	0.48^{***}	0.15	0.59^{***}	0.08
P	0.48^{***}	0.19	0.40^{**}	-0.08

p-values: $* < 0.01$, $** < 0.01$, $*** < 0.001$

Table. 7.5: Spearman's ρ between pairwise lists of similarities. WA are similarities based on word associations and WS are direct word similarities from human judgments.

Discussion The low correlation between the two similarities from human judgments is surprising. Our explanation is that this is because of different priming to functional and geometric dimension of meaning in the data collection task. In the WA task participants are not primed with the spatial domain but they are providing general word associations, hence functional associations. On the other hand, in the WS task participants are presented with two spatial relations, e.g. *left of* and *right of*, and therefore the geometric dimension of meaning is more explicitly attended. We also notice that judgments are not always unison, the same pair may be judged as similar and dissimilar which further confirms that participants are selecting between two different dimensions of meaning. This observation is consistent with

our argument that LM-vectors and P-vectors encode functional knowledge. Both representations correlate better with WA than with WS. Finally, Logan and Sadler (1996) demonstrate that WS judgments can be decomposed to dimensions that correlate with the dimensions of the spatial templates. We leave this investigation for our future work.

7.5 Conclusion and future work

In the preceding discussion, we have examined what semantic knowledge about spatial relations is captured in representations of a generative neural language model. In particular, we are interested if the language model is able to encode a distinction between functional and geometric bias of spatial relations and how the two dimensions of meaning interact. The idea is based on earlier work that demonstrates that this bias can be recovered from the selectivity of spatial relations for target and landmark objects. In particular, (i) we test the difference between multi-word spatial relations at two levels: the word embeddings which are a form of internal semantic representations in a language model and the perplexity-based P-vectors which are external semantic representations based on the language model performance; (ii) we project spatial relations in the contexts of other relations and we measure the fit of the language model to these contexts using perplexity (P-vectors); (iii) we use these contexts to build a distributional model of multi-word spatial relations; (iv) in the evaluation on standard semantic similarity tasks, we demonstrate that these vectors capture fine semantic distinctions between spatial relations; (v) we also demonstrate that these representations based on word-context associations latently capture geometric knowledge that allows analogical reasoning about space; this suggests that functional and geometric components of meaning are complementary: (vi) doing so we also demonstrated that generation of spatial descriptions is also dependent on textual features, even if the system has no access to the visual features of the scene. This has implications for baselines for image captioning and how we evaluate visual grounding of spatial relations.

Our work could be extended in several ways, including by (i) using the knowledge about the bias of spatial relations to evaluate captioning tasks with spatial word substitutions Shekhar et al. (2017a,b); (ii) examining how functional knowledge is complemented with visual knowledge in language generation Christie et al. (2016); Delecraz et al. (2017) (iii) using different

contextual embeddings such as ELMo Peters et al. (2018) and BERT Devlin et al. (2018) for the embedding layer of the generative language model rather than our specifically-trained word embeddings; note that P-vectors are representations of collections of context based on the performance of the decoder language model while ELMo and BERT are representations of specific context based on the encoder language model; (iv) comparing language models for spatial descriptions from different pragmatic tasks. As the focus of image captioning is to best describe the image and not for example, spatially locate a particular object, the pragmatic context of image descriptions is biased towards the functional sense of spatial relations. Our analysis should be extended to different kinds of corpora, for example those for visual question answering, human-robot interaction, and navigation instructions where we expect that precise geometric locating of objects receives more focus. Therefore, we expect to find a stronger geometric bias across all descriptions and a lower performance of our representations on analogical reasoning.

Acknowledgements

We are grateful to the anonymous reviewers for their helpful comments. The research of the authors was supported by a grant from the Swedish Research Council (VR project 2014-39) to the Centre for Linguistic Theory and Studies in Probability (CLASP) at Department of Philosophy, Linguistics and Theory of Science (FLoV), University of Gothenburg.

7.6 Appendix: Perplexity

The perplexity in the paper is formulated as follows:

$$PP(S) = (\prod_{s \in S} P(w_{1:t} = s))^{\frac{-1}{|S|}}. \tag{7.2}$$

By definition, the perplexity of a model q on a test suit S is defined as follows:

$$PP(S) = 2^{H(p,q)} \tag{7.6}$$

where H is cross entropy, and p is the likelihood of each possible sample in the test suit. The definition of cross entropy is as follows:

$$H(p,q) = -\sum_{x \in S} p(X = x) log_2(q(X = x)) \tag{7.7}$$

where X is a random variable, and x is a possible value of the random variable. In a forward generative language model, the random variable is conditioned on the previous words. With test suite being a sequence of words $S = w_{1:T}$, the likelihood of each word in the sequence is $p(w_t) = \frac{1}{T}$, and the cross entropy of the model on the samples is:

$$H(p,q) = -\sum_{t=1}^{T} p(w_t) log_2(q(w_t|w_{1:t-1})) \tag{7.8}$$

$$= -\frac{1}{T}\sum_{t=1}^{T} log_2(q(w_t|w_{1:t-1})) \tag{7.9}$$

where w_t is a token at a time t, in a sequence with maximum T tokens, $w_{1:t} = w_1, w_2, ..., w_t$. Therefore the perplexity is:

$$PP(S) = 2^{-\frac{1}{T}\sum_{t=1}^{T} log_2(q(w_t|w_{1:t-1}))} \tag{7.10}$$

$$= (\prod_{t=1}^{T} q(w_t|w_{1:t-1}))^{-\frac{1}{T}} \tag{7.11}$$

Equation 7.11 is often used as definition of perplexity in language models Goodman (2001) and Equation 7.10 is its numeric computation to avoid underflow due to adding logits.

There are two ways to extend the definition to the case when perplexity is calculated for a collection of sentences. (i) We can treat the corpus as a long sequence of tokens and use the previous equations. (ii) We can use Equation 7.7 with a change to the model definition, from a token model to a sentence model. The benefit of this method is that it assigns the same likelihood for each sentence regardless of its length. In this case, the chain

rule is used for the sentence model. The likelihood of each sentence is one over the number of sentences in the test suite, $p(s) = \frac{1}{|S|}$:

$$H(p,q) = -\sum_{s \in S} p(s) log_2(\hat{P}(s)) \tag{7.12}$$

$$= -\frac{1}{|S|} \sum_{s \in S} log_2(\hat{P}(s)) \tag{7.13}$$

Based on the chain rule, the sentence model can be calculated as follows:

$$\hat{P}(w_{1:T} = s) = \prod_{t=1}^{T} q(w_t | w_{1:t-1}) \tag{7.14}$$

$$log_2(\hat{P}(w_{1:T} = s)) = \sum_{t=1}^{T} log_2(q(w_t | w_{1:t-1})) \tag{7.15}$$

Perplexity in this case is defined as in Equation 7.2 here repeated as Equation7.16:

$$PP(S) = (\prod_{s \in S} \hat{P}(w_{1:T_s} = s))^{\frac{-1}{|S|}} \tag{7.16}$$

which instead of using the product is computed as a sum of logits from Equation 7.12 and 7.15.

7.7 Appendix: Examples of images from Visual Genome

Figure 7.5 and Figure 7.6 are the examples from VisualGenome which their region descriptions are used in the paper as examples of relation-context substitution table.

Figure. 7.5: image_id = 2367586
tall building above the bridge
bench below the green trees
car next to the water

Figure. 7.6: image_id = 2320485
scissors above the pen
the pen is below scissors
a ball-pen next to the scissorts

7.8 Appendix: Complete P-vectors

Figure 7.7 is the full presentation of P-vectors.

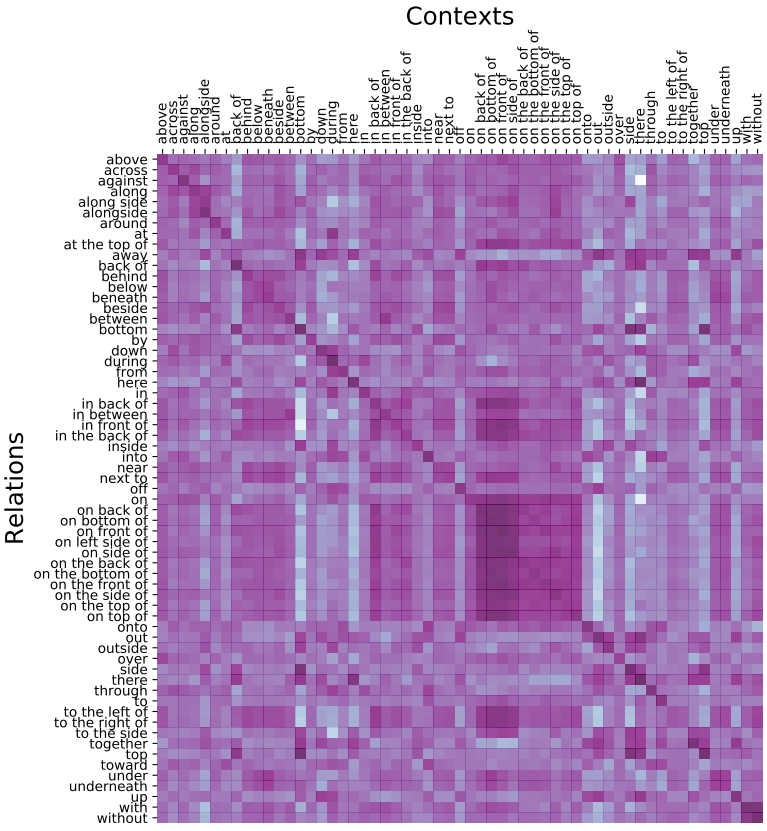


Figure. 7.7: Perplexity vectors for 67 spatial relations on 57 context bins.

7.9 Appendix: Similarity Judgment Dataset

In total 66 worker in Amazon Mechanical Turk annotated the word similarity. For each word pair, we collected 10 judgments. The word pairs vertically in random order were presented to annotators to judge their similarity. The input form was a slider in the web interface which they could freely adjust the indicator position between dissimilar and similar rating (Figure 7.8). In order to identify the bad annotators, we randomly asked the annotators to judge similarity between "*green*" and one of the spatial relations, we also asked similarity judgment between a spatial relation and itself. If the answer to similarity with green was higher than %60, or the answer for self similarity was lower than %90, all contributions of that worker were taken out from

the dataset. This cleaning technique removed 9 workers in total, which left us about 7 annotation on each word pair.

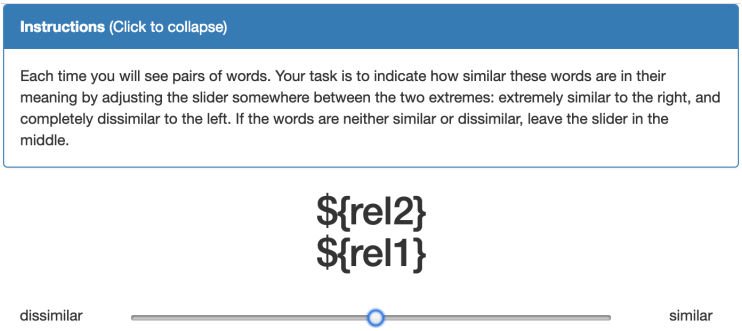


Figure. 7.8: The layout which presents the similarity judgment question.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Lalit R Bahl, Frederick Jelinek, and Robert L Mercer. 1983. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2:179–190.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155.

Stanley F Chen, Douglas Beeferman, and Ronald Rosenfeld. 1998. Evaluation metrics for language models. In *DARPA Broadcast News Transcription and Understanding Workshop*, pages 275–280. Citeseer.

François Chollet et al. 2015. Keras. `https://github.com/keras-team/keras`.

Gordon Christie, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra. 2016. Resolving language and vision ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes. *arXiv preprint arXiv:1604.02125*.

Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

BNC Consortium et al. 2007. The british national corpus, version 3 (bnc xml edition). *Distributed by Oxford University Computing Services on behalf of the BNC Consortium*.

Kenny R Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V Richards. 2004. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In *International Conference on Spatial Cognition*, pages 98–110. Springer.

Kenny R Coventry and Simon C Garrod. 2004. *Saying, seeing, and acting: the psychological semantics of spatial prepositions*. Psychology Press, Hove, East Sussex.

Kenny R Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of memory and language*, 44(3):376–398.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2018. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, pages 1–20.

Sebastien Delecraz, Alexis Nasr, Frédéric Béchet, and Benoit Favre. 2017. Correcting prepositional phrase attachments using multimodal corpora. In *Proceedings of the 15th International Conference on Parsing Technologies*, pages 72–77.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pages 1–6, Berlin, Germany.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

Mehdi Ghanimifard and Simon Dobnik. 2017. Learning to compose spatial relations with grounded neural language models. In *Proceedings of IWCS 2017: 12th International Conference on Computational Semantics*, pages 1–12, Montpellier, France. Association for Computational Linguistics.

Joshua T Goodman. 2001. A bit of progress in language modeling. *Computer Speech & Language*, 15(4):403–434.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Barbara Landau. 1996. Multiple geometric representations of objects in languages and language learners. *Language and space*, pages 317–363.

Barbara Landau. 2016. Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive Science*, 41(2):321–350.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

G.D. Logan and D.D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In M. Bloom, P.and Peterson, L. Nadell, and M. Garrett, editors, *Language and Space*, pages 493–529. MIT Press.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Arvind Neelakantan, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1059–1069.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, Volume 1 (Long Papers), volume 1, pages 2227–2237.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 2641–2649. IEEE.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220.

Angela Schwering. 2007. Evaluation of a semantic similarity measure for natural language spatial relations. In *International Conference on Spatial Information Theory*, pages 116–132. Springer.

Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. Vision and language integration: moving beyond objects. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. Foil it! find one mismatch between image and language caption. *arXiv preprint arXiv:1705.01359*.

Peter D Turney and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37:141–188.

8 Functional/Geometric Spectrum In Bounding Boxes

> **Simon Dobnik and Mehdi Ghanimifard.** Spatial descriptions on a functional-geometric spectrum: the location of objects. *Preprint - Under review 2020.*

Abstract Experimental research on spatial descriptions shows that their semantics are dependent on several modalities, among others (i) a geometric representation of space ("where", geometric knowledge) and (ii) dynamic kinematic routines between objects that are related ("what", functional knowledge). In this paper we examine whether geometric and functional bias of spatial relations is also reflected in large corpora of images and their corresponding descriptions. In particular, we examine whether the variation in object locations in the usage of a relation is a predictor of that relation's functional or geometric bias. Previous experimental psycho-linguistic work has examined the bias of some spatial relations, however our corpus-based computational analysis allows us to examine the bias of spatial relations and verbs beyond those that have been tested experimentally. Our findings have also implications for building computational image descriptions systems as we demonstrate what kind of representational knowledge is required to model spatial relations contained in them.

8.1 Introduction

The work on spatial relations such as "the chair is to the left of the table" and "the bicycle near the door" shows that the semantics of spatial relations is complex, drawing on several different modalities which include among

others (i) scene geometry, (ii) functional interactions between objects, and (iii) dialogue interaction between conversational partners. For example, Landau and Jackendoff (1993) argue that language encodes objects and places differently and this may be a reflection of different cognitive processes in the visual system: "what" and "where". Further, a number of papers Coventry et al. (2001); Coventry and Garrod (2004); Coventry et al. (2005); Hörberg (2008) show experimentally that different spatial relations have different bias in terms of functional ("what") and geometric ("where") knowledge. Similarly, Landau (2016) argues that two classes of spatial relations have different developmental trajectories and may be rooted in different neural representations. Dobnik and Åstbom (2017) argues that the bias to function and geometry of a particular relation is contextual and task-dependent.

For this reason, computational modelling of descriptions of spatial relations is challenging. Firstly, it requires information from each of these modalities to be present in the dataset. For example, it is hard to collect a large enough dataset of functional interactions between objects and represent these interactions as computationally useful representations. Secondly, there is a challenge of information fusion which needs to be attuned for different words in different contexts. Recently, deep neural networks modelling language and vision as perceptually grounded language models have demonstrated a lot of success Xu et al. (2015); Lu et al. (2017). An interesting research question therefore is what information such networks can capture in their representations from the available modalities and whether such representations correspond to the representations that have been argued for in linguistic and psychological literature.

For example, Dobnik and Kelleher (2013, 2014); Dobnik et al. (2018) explore whether functional and geometric bias can be recovered from the information encoded in a language model, the semantic associations encoded in the sequences of words. Language models together with word embeddings Bengio et al. (2003) are widely used to represent linguistic meaning in computational semantics and they are based on the premise known as the *distributional hypothesis* Firth (1957) that words occuring in similar contexts, represented by other words, will have similar meanings Turney et al. (2010). If we relate the distributional hypothesis to grounding in perception, this is because words co-occuring together will refer to identical situations and therefore the contexts of words become proxies for accessing

the underlying situations. It follows that information encoded in language models about spatial descriptions should encode some relevant semantics about dynamic kinematic routines between the objects that are related, albeit very indirectly. Hence, Dobnik and Kelleher (2013, 2014); Dobnik et al. (2018) demonstrate that the functional-geometric bias of expressions that have been tested experimentally in Coventry et al. (2001) is reflected in the degree to which target and landmark objects are associated with a relation in spatial descriptions extracted from a corpus of image descriptions. They start with the idea that while any two (abstract) objects can be related in geometric space, functional relations between the objects and relation are more specific, defined by the possible functional interaction between the objects. They demonstrate that this is expressed in the variability and generality of the target and landmark objects. Since a geometrically-biased spatial relation can relate any kind of objects that can be placed in a particular space, the objects used with such a relation will be more variable than the objects that occur with functionally-biased relations that also encode the nature of object interaction. They also show that usage of descriptions of an image corpus is crucial in this task since in a general corpus, a wider range of situations is reflected in the word contexts that may include metaphoric usages of the spatial words in other domains. We may consider such metaphorical usage of spatial relations in other domains as highly functional.

The experiments based on Coventry et al. (2001) show that spatial relations have functional or geometric *bias* which means that both components are relevant for the semantics of a description, just not the same degree. For example, a functionally-biased relation such as *over* is also sensitive to geometry to some extent, it appears that a presence of a function skews the regions of acceptability for the target object of that relation. The deviation in geometry can be explained by the fact that under a consideration of a functional relation different parts of the target and landmark object will become attended Coventry et al. (2005); Carlson et al. (2006). This results in a situation where the centroids of bounding boxes of target and landmark objects are displaced from the locations where we would expect to find them based on the geometric constraints alone. For example, in the case of a "teapot over a cup" it must be ensured that the spout of the teapot is located in such a way so that the liquid will be poured into a cup. In a scene described by a description "the toothpaste is over a toothbrush" the shape of the bounding boxes will be different from the previous scene as well as the

location of the attended areas. In the case of an "apple in a bowl" the bowl or its contents must constrain the movement of the apple (so that it does not fall out of the bowl) and hence locations of apples that are outside the bounding box of the bowl are also acceptable, for example where an apple is on the top of other apples. These examples suggest that over all contexts of target-landmark objects, the variation in locations of objects represented as bounding boxes will be much higher with functionally-biased spatial relations than geometrically-biased ones which will be closer to the axes of the geometric space. The latter is confirmed by the spatial templates of Logan and Sadler (1996) where in the absence of the functional knowledge, when an abstract shapes are used as targets and landmarks, both geometric and functional relations such as "over" and "above" give very similar axis-centred spatial templates. Hence, in this work, we explore whether we can detect a difference in the variability of the target objects in relation to the landmark objects for spatial relations of either geometric or functional bias in terms of representations of objects as visual features in images from a large corpus of images and descriptions and for relations that go beyond the ones that were tested experimentally. We expect that this variability will be the opposite of the variability that has been previously shown for textual data. Functional information can be recovered from the textual information about *what* objects are interacting, while geometric information can be recovered from *where* the visual features of objects are. Hence, we expect that relations that were experimentally found to have a functional bias will be less variable in their choice of target and landmark objects but more variable in terms of where these objects are in relation to the prototypical axes from the landmark. On the other hand, relations that were experimentally found to have a geometric bias, are expected show a higher variation in terms of the object kinds they relate but these will be geometrically less variable from the axes based on the landmark.

The experimental work on functional and geometric bias of spatial relations focuses on abstract images where the type of objects, their location and the nature of functional interaction is carefully controlled. This gives us accurate judgements about the applicability of descriptions but since the task focuses on abstract scenes this gives us different judgements to those we would have hoped to have obtained in real-life situations simply because of the perceptual and linguistic context is different from real-life situations Dobnik and Åstbom (2017). Ideally, we would need a corpus of interactions

between real objects and their spatial descriptions that on the perceptual side would be represented as 3-dimensional temporal model. Collecting such a corpus on a large scale would be a very challenging endeavour, although important work in this area has recently been done in route instructions in a virtual environments Thomason et al. (2019). On the other hand, there exist several large corpora of image descriptions, e.g. Krishna et al. (2017) which contain spatial descriptions and a large variety of interacting objects in real-life situations. For this reason they are, in our opinion, an attractive test-bed for examining the meaning of geometrically-biased and functionally-biased spatial relations. The down-side of image corpora is that the visual representations scenes are skewed, depending on the angle and the focus/scale at which an image was taken which means that an object such as a chair may have a different shape and size in respect to the image from one image to another. There is also no information about object depth and the dynamic interaction of objects. To counter this variation in objects we will introduce some normalisation steps. Of course, there will also be some noise in the scene representation's we obtain but we hope this noise will be uniform across different images and kinds of descriptions and therefore a relative comparison of descriptions of different bias will still give us a valid result.

Why is identification of functional and geometric bias of spatial relations relevant? Theoretically, the experiments give us more insights into the way spatial cognition is reflected in language. Showing that there is a distinction between these two classes of spatial relations on a large scale dataset of image descriptions gives a further support to the experimental evidence that has been obtained in carefully designed experiments. Knowing that there are different classes of spatial relations can help us in the task of generating image descriptions, for example in a robotic scenario. Following our observation, in an image description task functional relations are more informative than geometric relations as in addition to geometric component they also say something about the relation between the objects.[1] In a given scene a target object can be described and related to the landmark with several spatial relations based on geometric considerations alone. However, these descriptions could be filtered by considering those relations that are functionally more likely. The investigation also has implication for end-to-end

[1] Notice, however, that there are tasks where geometric information may be more informative, for example in locating a named object in a visual scene when answering a question.

image captioning systems build with deep learning architectures. Knowing that different spatial relations have a different bias for visual and textual modality would allow us a better comparison and evaluation of such systems. For example, there is a significant discussion in the vision and language community that end-to-end image captioning systems and visual question answering systems are relying too much on the information from language models Agrawal et al. (2018) rather than grounding words in an image, particularly when it comes to describing relations between objects. Knowing that not all spatial relations are equally geometrically spatial has important implications for evaluating such systems: (i) it shows that provided there is a balanced dataset reliance of a spatial relation on a language model is not necessarily a shortcoming but rather that is in fact the dimension that determines their meaning and there is a gradience in the way a description is grounded in visual vs textual features; (ii) it gives us insights into how we should build such systems in the future so that both (or even more) modalities are appropriately represented.

This paper is organised as follows: in Section 8.2 we describe the dataset of images and descriptions used in our studies; in Section 8.3 we describe how we represent geometric information from image annotations for spatial relations and how such representations can be compared for functional and geometric bias; in Section 8.4 we introduce a more sophisticated comparison in terms of the variation in our feature representations for different spatial relations from a representative representation; and we conclude in Section 8.5.

8.2 Dataset

We base our investigations on the Visual Genome dataset Krishna et al. (2017) which is a crowd-sourced annotations of 108,007 images. The dataset comprises several types of annotations including the region descriptions (phrases and sentences referring to one bounding box), objects (annotated as bounding boxes), attributes for each object annotation, and *relationships* between them (triplet of subject, predicate, object). Most object names, attributes and predicate of relationships are also mapped to WordNet synsets. The predicates in relationships include *spatial relations* such as "*above*", "*under*", "*on*", "*in*" but also verbs describing events such as "*holding*" and "*wearing*", or a combination of both such as "*sitting on*".

Without any data cleaning, the total number of possible forms of relation tokens is 36,550. Since spatial relations are multi-word expressions, we create a dictionary of relations capturing different variations of their syntactic form (e.g. "to the left of", "on the left", "left", etc.) based on the lists of English spatial relation constructions in Landau (1996) and Herskovits (1986). Out of 235 spatial relations, we only found 78 types. Some variation in writing of relationships may be simply due to the annotator shorthand notation, e.g. "to left of". We combine the compound variants of spatial relations to a lower-cased single variant in cases where we can be reasonably sure that this will not affect their semantics in terms of functional and geometric bias. Duplicate descriptions per image which are created by different annotators are removed, as well as those descriptions where the extracted spatial relations are not used in a complete locative description involving a target object, relation and a landmark, e.g. "chair on left". At the end, we only kept those relations which have more than 30 instances in the dataset.

In addition to spatial relations, we also added a few verbal relations which possibly have spatial content. Including the verbs which Collell et al. (2018) showed to have strong predictability of object on the y-axis. The dictionary of all relations examined in this study is given in Table 8.1.

Table. 8.1: The list of spatial relations captured and additional verbs with spatial content.

over, *above*, *below*, *under*, *left of*, *right of*, *on*, *in*, *inside*, *outside*, *far from*, *away from*, *next to*, *near to*, *across*, *at*, *with*, *beneath*, *underneath*, *through*, *alongside*, *against*, *off*, *between*, *from*, *beside*, *to*, *by*, *along*, *around*, *behind*, *bottom*, *top*, *front of*, *back of*, *side of*,
flying, *kicking*, *cutting*, *catching*, *riding*, *seeing*, *looking*, *floating*, *finding*, *pulling*, *removing*, *having*, *wearing*, *containing*, *holding*, *supporting*, *sitting*, *touching*.

8.3 Representing locations as dense geometric vectors

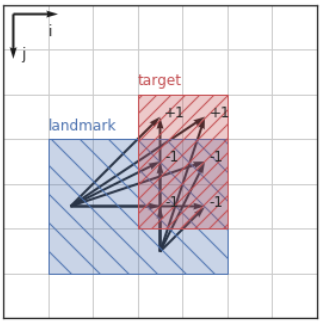
Each bounding box in Visual Genome is represented with 4 numerical values: the x-, y- coordinates relative to the image frame, the bounding box width and height. In order to compare the geometric arrangements of objects

represented as bounding boxes between different spatial relations, as well as to compare this data with the data from spatial templates from Logan and Sadler (1996), we convert both representations to 3-dimensional dense vectors $[x, y, d]$ where x and y represent directions in the 2-dimensional space and d is a Euclidean distance between x and y. Hence, we separate directionality (represented by x and y) from the distance. The intuition behind this comes from a distinction between *directionals* ("to the left of" and "above") and *topological relations* ("close" and "far") where the former are dependent on both directionality and distance but the latter are only dependent on distance. The 3-dimensional vectors (the x and y dimension) are inspired by vectors introduced in the Attentional Vector Sum Model (AVS) Regier and Carlson (2001). However, as we will describe below they are used quite differently. Rather then modelling the attention for a particular pair of bounding boxes in the AVS model we use them to estimate attention between all bounding boxes that are related by a particular spatial relation. In other words, we use them to estimate the likelihood that for a particular spatial relation a particular location is occupied by an object. Therefore, the representations are similar to the notion of spatial templates. Here, other representations of bounding boxes could also be used (see for example Sadeghi et al. (2015); Nikolaus et al. (2019). We opt for low-level features that have been experimentally shown to be directly relevant for the (geometric) semantics of spatial relations and which are also available in spatial templates.

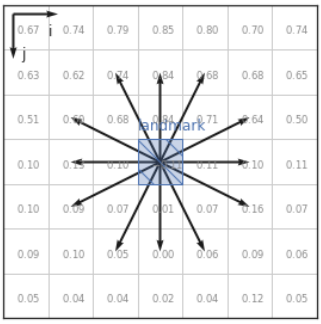
We derive the dense features as follows. First, as shown in Figure 8.1a, we segment images into 7×7 locations. Then, for every pair of points in the locations matrix, we define a dense vector as:

$$\text{for two points on image} \begin{cases} p_1 = \langle i_1, j_1 \rangle \\ p_2 = \langle i_2, j_2 \rangle \end{cases} , \vec{u}_{p_1,p_2} = \begin{bmatrix} x \\ y \\ d \end{bmatrix} = \begin{bmatrix} \frac{i_2 - i_1}{||\overrightarrow{p_1 p_2}||_2} \\ \frac{j_1 - i_2}{||\overrightarrow{p_1 p_2}||_2} \\ \texttt{sgn} \cdot ||\overrightarrow{p_1 p_2}||_2 \end{bmatrix}$$

where $\vec{u}_{p_1,p_2}$ represents the dense geometric relation features between two points, which p_1 is a point on landmark and p_2 is on the target, the Euclidean distance between them is $||\overrightarrow{p_1 p_2}||_2 = \sqrt{(i_2 - i_1)^2 + (j_2 - j_1)^2}$, and $\texttt{sgn}$ is a sign value which is -1 if p_2 is also a point on the landmark bounding box, otherwise $+1$.



(a) Bounding boxes in an image



(b) Spatial template

Figure. 8.1: (a) Images are segmented to a fixed set of locations and relation vectors are calculated for every pair of locations occupied by the bounding boxes of target and landmark. (b) In spatial templates a vector is calculated for every location of the template originating in the location of the landmark.

For each relation REL, this gives us a collection of vectors. For bounding boxes annotated with relations in the images of Visual Genome, we build the collection of dense vectors of all points connecting targets and landmarks related by each particular relation in the dataset ($V_{\text{REL}}^{(vg)}$). Formally, this set is represented as follows:

$$V_{\text{REL}}^{(vg)} = \left\{ \vec{u}_{p_1,p_2} \right\}_{\substack{\langle \text{TRG,REL,LND} \rangle \in \text{Images} \\ p_1 \in \text{bbox}_{\text{LND}} \\ p_2 \in \text{bbox}_{\text{TRG}}}} \tag{8.1}$$

where bbox_{TRG} and bbox_{LND} are the collection of points in bounding boxes of target TRG and landmark LND.[2]

Similarly, we use this method on spatial templates from Logan and Sadler (1996) to build all possible dense vectors. As shown in Figure 8.1b, we create a dense vector originating in the central location of the landmark and ending at every possible location of target in the spatial template. Each vector from a spatial template is associated with the acceptability score of the target location.

$$V^{(st)} = \left\{ \vec{u}_{\langle 3,3 \rangle, \langle i,j \rangle} \right\}_{\substack{i \in \{1,..,7\} \\ j \in \{1,..,7\}}}, S_{\text{REL}} = \left\{ \vec{s}_{i,j} \right\}_{\substack{i \in \{1,..,7\} \\ j \in \{1,..,7\}}} \tag{8.2}$$

[2]For computational convenience, instead of including all possible annotations in this set, we randomly sampled a maximum of 1000 triplets from the relationship dataset.

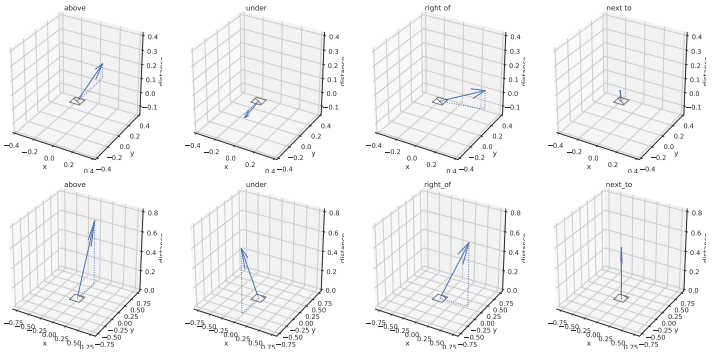


Figure. 8.2: Examples of $\vec{v}_{\text{REL}}^{(vg)}$ and $\vec{v}_{\text{REL}}^{(st)}$: among the examples, x-y features are mostly similar but the scale and sign of distances are different.

where S_{REL} represents the collection of normalised acceptabilities in spatial template of the relation REL.

These vectors in each collection are then projected to a single vector representation using the following methods. For the collection of vectors from a spatial template, the representative vector is the weighted sum of all possible vectors with acceptability scores:

$$\vec{v}_{\text{REL}}^{(st)} = \sum_{\substack{i\in\{1,..,7\}\\ j\in\{1,..,7\}}} s_{i,j} \cdot \vec{u}_{\langle 3,3\rangle,\langle i,j\rangle} \tag{8.3}$$

For the collection of vectors from the Visual Genome bounding boxes, the representative vector is the expected 3-feature vector:

$$\vec{v}_{\text{REL}}^{(vg)} = E[V_{\text{REL}}^{(vg)}] = \frac{1}{|V_{\text{REL}}^{(vg)}|} \sum_{\vec{v}\in V_{\text{REL}}^{(vg)}} \vec{v} \tag{8.4}$$

where $|V_{\text{REL}}^{(vg)}|$ is the number of vectors. Adding vectors with contradicting features will cancel each other and remaining vector points at a direction with least opposite directions. More importantly, $\vec{v}_{\text{REL}}^{(vg)}$ resulted from bounding box annotations in visual genome is similar $\vec{v}_{\text{REL}}^{(st)}$ resulted from compressing the spatial templates into a three dimensional feature vectors.

To compare the projected dense vectors we have obtained from the images with those from the spatial templates we use cosine similarity or distance

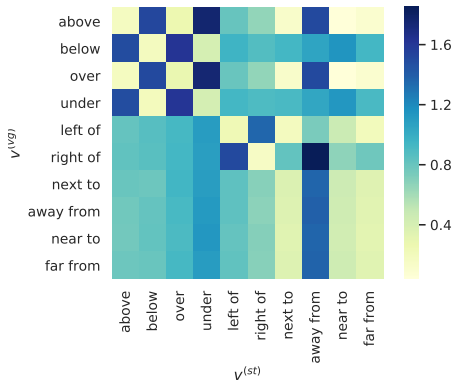


Figure. 8.3: A comparison of dense vector representations from images $\vec{v}^{(vg)}_{\text{REL}}$ and those from spatial templates $\vec{v}^{(st)}_{\text{REL}}$ with the cosine distance: $1 - cosine(\vec{v}^{(vg)}_{\text{REL}}, \vec{v}^{(st)}_{\text{REL}})$.

as shown in Figure 8.3 where the horizontal axis represents the vectors from spatial templates $\vec{v}^{(st)}_{\text{REL}}$ and the vertical axis represents the vectors from images $\vec{v}^{(vg)}_{\text{REL}}$. The results indicate that the 3-dimensional vectors from the two datasets are very similar except in the case of "away from". Except for this case the lowest cosine distance is on the diagonal. The results also indicate that pairs of geometrically or functionally biased spatial relations such as "over" and "above" and "under" and "below" have similar overall directions and distances. Projective relations have clearly defined opposites alongside one axis but topological relations are overlapping with the projective relations. "next to$_{st}$" is similar to "next to$_{vg}$", "away from$_{vg}$", "near to$_{vg}$" and "far from$_{vg}$" and "away from$_{st}$" is dissimilar to all. This has possibly to do with the way distance is represented in images. Humans are able to estimate distance between two focused objects not on their actual size but the size they know from their background knowledge.

The comparison of dense vectors here indicates that similar dense vectors are obtained from both datasets. However, it does not distinguish functional and geometric bias of different relations. For example, "over$_{st}$" is equally similar to "over$_{vg}$" and "above$_{vg}$" while we were expecting that since "over$_{st}$" is used in the geometric context it will more similar to "above$_{vg}$". This is because cosine similarity/distance takes into account all three dimensions x, y and z of the dense vectors. However, we expect that "over$_{st}$" will be

similar to "over$_{vg}$" in y and d dimensions but different in the x dimension which distinguishes its geometric and functional use.

In the following section we examine the 3-dimensional feature space of the dense vectors in terms of the variation in the distribution of features. Therefore, we need to look for a measure that captures variation in distribution of features.

8.4 Variation of features within dense vectors

We argued in Section 8.1 that we expect that functionally-biased relations will be associated with more variable locations of target and landmark objects as these will also be dependent on the functional relations between individual object pairs. In the previous section we represented the locations between targets and landmarks as dense vectors which were then projected to one representative vector for each spatial relation. The degree of divergence from the representative vectors can be considered as an indication for non-geometrical use of spatial relations. In order to test this, for each spatial relation, we calculate a deviation of individual target-landmark vectors $\vec{v}$ from the representative 3-dimensional dense vector $\vec{v}^{(vg)}_{\text{REL}}$. As a metric of deviation we use cosine distance:

$$Distances = \left\{ 1 - cosine(\vec{v}^{(vg)}_{\text{REL}}, \vec{v}) \right\}_{\vec{v} \in V^{(vg)}_{\text{REL}}} \tag{8.5}$$

We expect that on average, cosine distances in geometrically-bias relations are closer to 0 (there is a clearer central tendency), and the overall distribution of cosine distances is positively skewed: the mode of cosine distances is close to zero while the mean and the tail of differences is skewed to the right.[3] In Figure 8.4, we select a set of geometrically- (blue) and functionally-biased (orange) relations as reported in psycho-linguistic experiments and plot (a) their average cosine distances of dense vectors from their representative vector and (b) the skewness of cosine differences. We also include relations the bias of which has not been tested experimentally (grey) but

[3] To calculate skewness we use an implementation of the Fisher-Pearson coefficient (Kokoska and Zwillinger, 2000, s.2.2.24.1) in `scipy.stats.skew`.

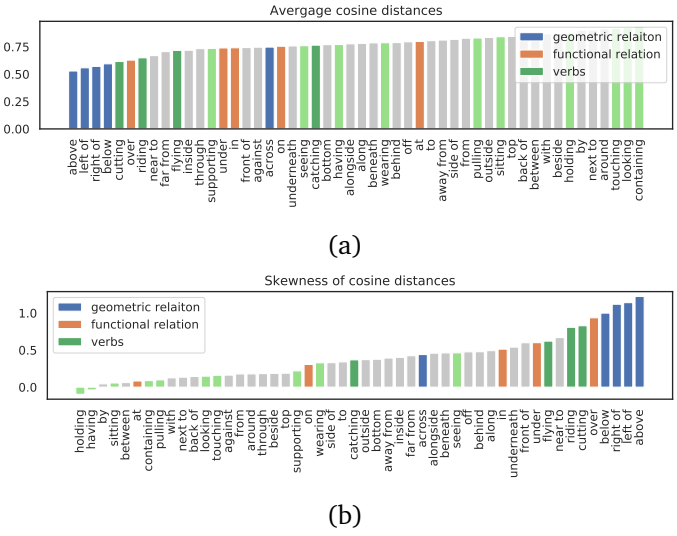


Figure. 8.4: (a) The average cosine distance of dense vectors $[x, y, d]$ from the expected dense vector of each spatial relation. (b) The skewness in distribution of distances.

we expect that this is demonstrated by their position in the graph between the key-points determined experimentally. Finally, we also include some verbs describing events and situations involving interacting objects in space that are also annotated as relationships in the Visual Genome (green), e.g. "boy, feeds, giraffe". We are particularly interested in the verbs that are reported in Collel et al. Collell et al. (2018) for which the location of the (target) object is most strongly predictable from the y dimension ("flying", "kicking", "cutting", "catching" and "riding") (dark green in Figure 8.4) and those for which the y dimensions is the least predictable in respect to the location of the object ("see", "float", "finding", "pulled" and "removes") (light green) listed in their Table 3, p.6770. However, here Collell et al. (2018) do not consider the x-dimension which may be a relevant dimension for the verbs in the picture. A quick comparison of the two lists gives an impression that the former contains descriptions of events involving object relations that more strongly grounded in the image representations (e.g. "riding") and are therefore similar to geometrically-biased spatial relations, while the second list contains descriptions of events that are less strongly grounded in the image representations (e.g. "sees") and would require a simulation of

dynamic kinematic routines between the objects which makes them similar to functionally-biased spatial relations.

Examining the average cosine distances from the representation vector of each spatial relation in Figure 8.4a we can see that relations that have been identified as geometrically-biased (blue) tend to have a lower average cosine distance from the representation's dense vector than those that have been identified as functionally-biased (orange). The same tends also to be the case for verbs identified in Collell et al. (2018) for which the objects are more dependent on the y (dark green) compared to verbs for which the objects are less dependent on the y dimension (light green). Note that in this comparison a deviation of the entire 3-dimensional vector $[x, y, d]$ was taken into account and therefore a deviation can be in any of these dimensions. Examining the skewness of cosine distances from the representation vector of each spatial relation in Figure 8.4b we can see that geometrically-biased verbs and verbs that are more strongly grounded show a tendency towards a higher skewness of distribution, they are more biased towards the representational vectors. Overall, the results indicate support for our hypothesis in Section8.1 that bounding boxes are predictors of the functional and geometric bias as well as they indicate that the same bias is also present in verbal descriptions of scenes.

In Figure 8.5 we examine the histograms of deviations from the representational vectors of "on", "in", "over", "above", "right of" and "left of". To plot these histograms we use Kernel Density Estimation (KDE)[4] Scott (2015) which indicates the density of samples in the range of [0, 2] of the cosine distance (Equation 8.5). We also give examples of target-landmark pairs which have the highest (orange) and the lowest (blue) average distances from the representational vectors. These examples indicate that functionally biased relations ("on", "in" and "over") can be and are used in contexts where the geometric constraint is also satisfied and this is represented in the image while they can also be used in the contexts where there is a deviation from the geometric constraint, just as predicted by experiments in Coventry et al. (2001). Interestingly, among the cases that show high deviation from the representational vectors we also find examples that are typically considered to involve more complex geometric conceptualisation, for example "bracelet on wrist", "woman in dress", "trees over rocks". However, the relations that we consider to be geometrically-biased we also find examples of high

[4]We use an implementation based on `scipy.stats.gaussian_kde`

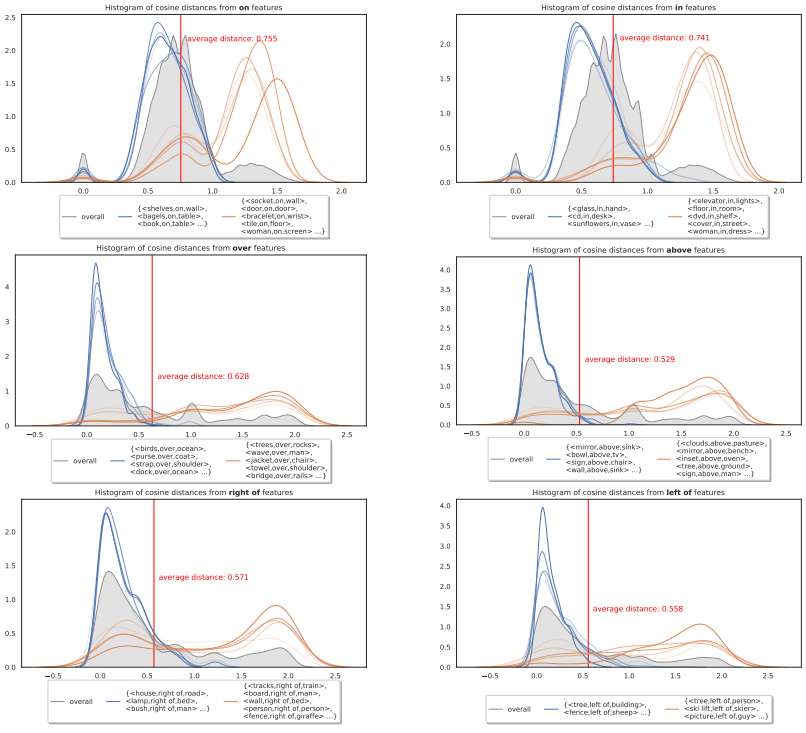


Figure. 8.5: Using the KDE method we plot a histogram of cosine distances of individual examples from the representational vector of each relation which shows skewness to zero for geometrically-biased usages of relations. For the projective relations "right of" and "left of" the examples with landmarks with tendency for enforcing intrinsic frame of reference (animate objects, objects with clearly defined front and back) are negatively skewed which represents maximum cosine distance.

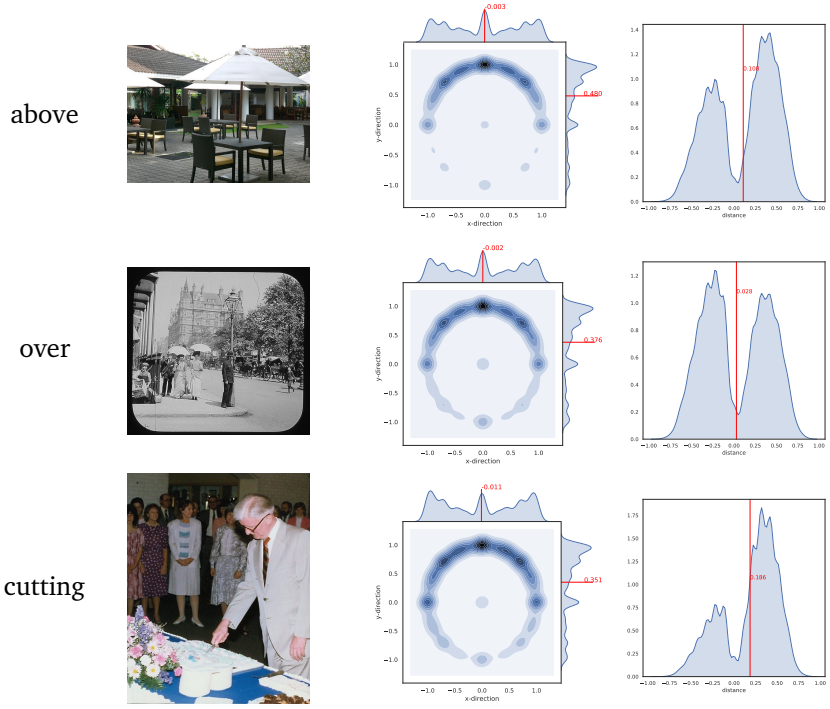


Figure. 8.6: The individual features of dense vectors $[x, y, d]$ have different distributions for each relation.

deviation from the representational vectors. The examples for "above" seem to to correspond to usages where there is an element of covering/protection that has been argued to be the functional component of "over": e.g. "clouds above/over pasture" and "mirror above/over bench" or cases that require complex geometric conceptualisation of the scene "tree above ground".[5] We are intrigued by the examples that deviate from the representational vectors for "left of" and "right of". They frequently contain animate beings (people) or objects with clear orientation. Our assumption is that these examples are a reflection of changes of the perspective from the relative frame of reference of the observer of the image to the intrinsic frame of reference of the landmark.

As stated earlier, the dense vector representations including their cosine distances aggregate three features $[x, y, d]$ and therefore the previous com-

[5] It could be argued that these cases require functional representation since one needs to know how to geometrically conceptualise the scene involving that particular pair of objects in order the geometric relation can be established.

parisons do not take into account the role of each individual feature for spatial relations. In Figure 8.6 we plot the distribution of all features over all vectors of $V_{\text{REL}}^{(vg)}$ for some individual relations.[6] The individual histograms for the x (centre top), y (centre right) and d feature (on the right side) indicate the density of their values and the mixture density graph for x, y (centre) shows how these features interact. This graph demonstrates that "over" and "cutting" have more freedom of variation in the x dimension as well as the negative y dimension (which indicates overlap of objects) than "above". As discussed earlier, there is also considerable overlap between all three graphs which is due to the fact that functionally-biased relations are also used in situations when geometric constraints are satisfied. While "cutting" is more similar to "over" than "above" in terms of the xy dimensions, it has a different distance histogram with far fewer overlapping cases.

8.5 Conclusion

In this paper we have demonstrated and discussed how the functional and geometric bias of spatial relations can be identified from geometric annotations of objects as bounding boxes connected by spatial relations in a corpus of images and associated descriptions. The bounding boxes are converted to 3-dimensional dense vectors that contain information about the x, y and d dimension. These vectors can be then converged to a single representational vector for each spatial relation. Vectors from different relations can then be compared with cosine similarity. To increase the granularity of comparison we examine how individual examples of annotated situations diverge from the representational vectors and what are the distributions of these divergences, also at the level of individual features. Our results indicate that functional and geometric bias of spatial relations can be identified from the geometric spatial information corpus of images and descriptions and also that this distinction can be carried over to verbs describing situations involving objects. In terms of semantics of spatial relations our study shows that to a certain degree information that was previously determined experimentally can be uncovered from a large corpus of image descriptions and for a large number of relations including verbs. Practically, such information is extremely useful for building end-to-end deep neural models of image

[6] These relations were found to be strongly dependent on the y feature in Collell et al. (2018) who did not investigate the contribution of other features.

captioning as it demonstrates what kind of representations are relevant for different kinds of descriptions which has also been the focus of our other studies. Another question that we find relevant to explore in our future work is the observation that the context in which the dataset was created may have a general bias on the degree to which function and geometry is considered to be relevant. For example, is the goal of the image description task to describe *what* is happening with the objects or to locate *where* the objects are. Finally, different classes of verbs would also deserve a more focused study.

Bibliography

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(6):1137–1155.

Laura A. Carlson, Terry Regier, William Lopez, and Bryce Corrigan. 2006. Attention unites form and function in spatial language. *Spatial Cognition & Computation*, 6(4):295–308.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110. Springer Berlin Heidelberg.

Kenny R Coventry and Simon C Garrod. 2004. *Saying, seeing, and acting: the psychological semantics of spatial prepositions*. Psychology Press, Hove, East Sussex.

Kenny R. Coventry, Mercè Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the apprehension of Over, Under, Above and Below. *Journal of Memory and Language*, 44(3):376–398.

Simon Dobnik and Amelie Åstbom. 2017. (Perceptual) grounding as interaction. In *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–26, Saarbrücken, Germany.

Simon Dobnik, Mehdi Ghanimifard, and John Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 1–11, New Orleans. Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2013. Towards an automatic identification of functional and geometric spatial prepositions. In *Proceedings of PRE-CogSsci 2013: Production of referring expressions – bridging the gap between cognitive and computational approaches to reference*, pages 1–6, Berlin, Germany.

Simon Dobnik and John D. Kelleher. 2014. Exploration of functional semantics of prepositions from corpora of descriptions of visual scenes. In *Proceedings of the Third V&L Net Workshop on Vision and Language*, pages 33–37, Dublin, Ireland. Dublin City University and the Association for Computational Linguistics.

John R. Firth. 1957. A synopsis of linguistic theory 1930–1955. *Studies in linguistic analysis*, pages 1–32.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Thomas Hörberg. 2008. Influences of form and function on the acceptability of projective prepositions in Swedish. *Spatial Cognition & Computation*, 8(3):193–218.

Stephen Kokoska and Daniel Zwillinger. 2000. *CRC standard probability and statistics tables and formulae*. Crc Press.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Barbara Landau. 1996. Multiple geometric representations of objects in languages and language learners. *Language and space*, pages 317–363.

Barbara Landau. 2016. Update on "what" and "where" in spatial language: A new division of labor for spatial terms. *Cognitive Science*, 41(2):321–350.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383.

Mitja Nikolaus, Mostafa Abdou, Matthew Lamm, Rahul Aralikatte, and Desmond Elliott. 2019. Compositional generalization in image captioning. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*.

Terry Regier and Laura A. Carlson. 2001. Grounding spatial language in perception: an empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2):273–298.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

David W Scott. 2015. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

9 Evaluating Generation Of Spatial Descriptions With Adaptive Attention

Mehdi Ghanimifard and Simon Dobnik. Evaluating Generation of Spatial Descriptions with Adaptive Attention. *In European Conference on Computer Vision, pp. 153-161. Springer, Cham, 2018.*

Abstract We examine and evaluate adaptive attention Lu et al. (2017) (which balances the focus on visual features and focus on textual features) in generating image captions in end-to-end neural networks, in particular how adaptive attention is informative for generating spatial relations. We show that the model generates spatial relations more on the basis of textual rather than visual features and therefore confirm the previous observations that the learned visual features are missing information about geometric relations between objects.

9.1 Introduction

End-to-end neural networks are commonly used in image description tasks Vinyals et al. (2015); Xu et al. (2015); Lu et al. (2017). Typically, a pre-trained convolutional neural network is used as an encoder which produces visual features, and a neural language model is used as a decoder that generates descriptions of scenes. The underlying idea in this *representation learning* scenario Bengio et al. (2013) is that hidden features are learned from the observable data with minimum engineering effort of background knowledge. For example in word sequence generation only some general properties of a sequence structure Sutskever et al. (2014) are given to the

learner while the learner learns from the observed data what word to choose in a sequence together with a representation of features. Recent models such as Xu et al. (2015); Lu et al. (2017) also add to the neural language model a model of visual attention over visual features which is inspired by the attention mechanism for alignment in neural machine translation Bahdanau et al. (2014). It may be argued that the attention mechanism introduces modularity to representation learning in the sense of *inception modules* Szegedy et al. (2015) and *neural module networks* Andreas et al. (2016). The visual attention is intended to detect the salient features of the image and align them with words predicted by the decoder. In particular, it creates a sum of the weighted final visual features at different regions of an image:

$$c_t = \sum_{i=1}^{k} \alpha_{ti} v_i \tag{9.1}$$

where at time t, c_t represents the pooled visual features, i corresponds to k different regions of image, v_i is the visual representation of a particular region, and α_{ti} represent the amount of attention on the specific region of the image. This representation provides the features for grounding the prediction of next word:

$$logPr(w_{t+1} = y_{t+1} | w_{1:t} = y_{1:t}, I = v_{1:k}) \approx f(y_{1:t}, c_t) \tag{9.2}$$

where f represents the end-to-end neural network for approximating the prediction of the next word in sentence.

However, not all words in natural language descriptions are directly grounded in visual features which leads Lu et al. (2017) to extend the attention model Xu et al. (2015) with an adaptive attention mechanism which learns to balance between the contribution of the visual signal and the language signal when generating a sequence of words.

$$\hat{c}_t = \beta_t s_t + (1 - \beta_t) c_t \tag{9.3}$$

where at time t, $\hat{c}_t$ is a combined representation of language features and visual features in addition to c_t of the visual features from Equation 9.2. s_t is obtained from the memory state of the language model, and β_t ranging between $[0, 1]$ is the adaptive attention balancing the combination of vision and language features.

The performance of the image captioning systems when evaluated on the acceptability of the generated descriptions is impressive. However, in order to evaluate the success of learning we also need to understand better what the system has learned especially because good overall results may be due to the dataset artefacts or the system is simply learning from one modality only, ignoring the other Agrawal et al. (2018). Understanding the representations that have been learned also gives us an insight into building better systems for image captioning, especially since we do not have a clear understanding of the features in the domain. An example of work in this area is Liu et al. (2017) which evaluates visual attention on objects localisation. Shekhar et al. (2017b) developed the FOIL dataset as a diagnostic tool to investigate if models look at images in caption generation. In Shekhar et al. (2017a) they examine the FOIL diagnostic for different parts-of-speech and conclude that the state of the art models can locate objects but their language models do not perform well on other parts-of-speech.

The current paper focuses on generation of spatial descriptions, in particular locative expressions such as "the chair to the left of the sofa" or "people close to the statue in the square". Spatial relations relate a target ("people") and landmark objects ("the statue") with a spatial relation ("close to"). They depend on several contextual sources of information such as scene geometry ("where" objects are in relation to each other), properties or function of objects and their interaction ("what" is related) as well as the interaction between conversational participants Herskovits (1986); Landau and Jackendoff (1993); Regier (1996); Coventry and Garrod (2004); Dobnik and Kelleher (2017). The features that are relevant in computational modelling of spatial language are difficult to determine simply by manually considering individual examples and they are normally identified through experimental work. The representation learning models are therefore particularly suited for their computational modelling.

However, the end-to-end vision and language models with attention are implemented in a way to recognise objects and localise their area in an image Ba et al. (2014); Mnih et al. (2014). To generate spatial relations, Ramisa et al. (2015) propose a combination of visual representations from convolutional neural networks and manually designed geometric representation of targets and landmarks. On quick examination, the representation of attention over images as in Xu et al. (2015) gives an impression that attention captures both "what" and "where", especially because the atten-

tion graphs resemble *spatial templates* Logan and Sadler (1996). However, Kelleher and Dobnik (2017) argue that due to the design properties of image captioning networks, attention does not capture "where" as these models are built to identify objects but not geometric relations between them which they examine at the level of qualitative evaluation of attention on spatial relations.

In this paper we quantitatively evaluate the model of adaptive attention of Lu et al. (2017) in predicting spatial relations in image descriptions. The resources used in our evaluation are described in Section 9.2. In Section 9.3 we examine the grounding of different parts-of-speech in visual and textual part of attention. Furthermore, in Section 9.4 we investigate the attention on spatial relations, targets and landmarks. We conclude by providing the possible directions for future studies and improvements.

9.2 Datasets and Pre-trained Models

As a part of their implementation Lu et al. (2017) provide two different pre-trained image captioning models: Flickr30K Young et al. (2014) and MS-COCO Lin et al. (2014).[1] We base our experiments on spatial descriptions of 40,736 images in the MS-COCO test corpus.

9.3 Visual Attention and Word Categories

Hypothesis Our hypothesis is that visual attention in the end-to-end image captioning systems works as an object detector similar to Ba et al. (2014); Mnih et al. (2014). Therefore, we expect the adaptive attention to prefer to attend to visual features rather than the language model features when predicting categories of words found in noun phrases that refer to objects, in particular head nouns. We expect that both scores will be reversed: more predictable words by the language model in the blind test receive less visual attention.

Method We use the pre-trained model of adaptive attention [2] to generate a description for each of the 40,736 images in the MS-COCO-2014 test. All

[1] https://filebox.ece.vt.edu/~jiasenlu/codeRelease/AdaptiveAttention.
[2] https://filebox.ece.vt.edu/~jiasenlu/codeRelease/AdaptiveAttention/model/COCO/coco_challenge/model_id1_34.t7

the attention values are logged (α, β). We apply universal part-of-speech tagger from NLTK Bird et al. (2009) on the generated sentences and report the average visual attentions on each part-of-speech. We match our results with results on the degree of predictability of each part-of-speech from the language model without looking at the image from the blind test of Shekhar et al. (2017a). Note that we do not investigate the overall quality of the model on the test set (this has already been evaluated by its authors) but what kind of attention this model gives to vision and language features used to generate a word of each category. The evaluaiton code:
`https://github.com/GU-CLASP/eccv18-sivl-attention`

Results Table 9.1 indicates that the highest degree of visual attentions is on numbers (NUM), nouns (NOUN), adjectives (ADJ) and determiners (DET) respectively. Pronouns (PRON) and particles (PRT) receive the lowest degree of visual attention. Verbs (VERB) and adverbs (ADV) are placed in the middle of this sorted list. Spatial relations which are mainly annotated as prepositions/adpositions (ADP) receive the second lowest visual attention, higher only than pronouns (PRON) and particles (PRT). Our results are different from the accuracy scores of detecting mismatch descriptions in the FOIL classification task Shekhar et al. (2017a). For example, the model assigns predicts the mismatch on ADJ easier than mismatch on ADV. As hypothesised, the part-of-speech that make up noun phrases receive the highest visual attention (and the lowest language model attention). The results also indicate that the text is never generated by a single attention alone but a combination of visual and language model attentions. Since some spatial relations are often annotated as adjectives (e.g. "front"), a more detailed comparison on spatial terms is required.

9.4 Visual Attention when Grounding Spatial Descriptions

In generation of a sequence of words that make up a spatial description, which type of features or evidence is taken into consideration by the model as the description unfolds?

Hypothesis In Section 9.3, we argued that the generation of spatial relations (prepositions/adpositions) is less dependent on visual features com-

POS	Count	Mean $\pm$ std	Blind test
NUM	1882	0.81 ± 0.08	-
NOUN	134332	0.78 ± 0.12	0.23
ADJ	23670	0.77 ± 0.14	0.76
DET	96641	0.73 ± 0.12	-
VERB	38381	0.70 ± 0.11	0.57
CONJ	6755	0.70 ± 0.13	-
ADV	184	0.69 ± 0.12	0.18
ADP	64332	0.62 ± 0.15	0.54
PRON	2347	0.53 ± 0.14	-
PRT	6462	0.52 ± 0.21	-

Table. 9.1: The average visual attention $(1-\beta)$ for predicting words on each part-of-speech. The scores from the blind test indicate the accuracy of detecting a mismatch description in the FOIL-classification task Shekhar et al. (2017a).

pared to noun phrases due to the fact that the learned visual features are used for object recognition and not recognition of geometric spatial relations between objects. Moreover, the visual clues that would predict the choice of spatial relation are not in one specific region of an image; this is dependent on the location of the target, the landmark and the configuration of the environment as a whole. Therefore, our hypothesis is that when generating spatial relations the visual attention is more spread over possible regions rather than being focused on a specific object.

Method The corpus tagged with POS from the previous section was used. In order to examine the attention on spatial relations, a list of keywords from Herskovits (1986); Landau and Jackendoff (1993) was used to identify them, provided that they have a sufficient frequency in the corpus. The average adaptive visual attention for each word can be compared with the scores in Table 9.1 for different parts-of-speech. In each sentence, the nouns before the spatial relation and the nouns after the spatial relations are taken as the most likely targets and landmarks respectively. The average adaptive visual attention on targets, landmarks and and spatial relations is recorded.

Results In Table 9.2 we report for each spatial relation and its targets and landmarks the average adaptive visual attention. The adaptive attentions for triplets are comparable with the figures for each part-of-speech in Table 9.1. In the current table, the variance of visual attentions is reported

with the $max - min$ measure which is the difference between maximum and minimum attentions on a 7x7 plane representing the visual regions in the model. Lower values indicate either a low attention or a wider spread of attended area, hence less visual focus. Higher values indicate that there there is more visual focus. For each spatial relation, the triplets must be compared with each other. In all cases, our hypothesis is confirmed: (1) the adaptive visual attention is lower on predicting spatial relations which means that they receive overall less visual attention, (2) with the exception of "under", the difference between maximum and minimum visual attentions are lower with spatial relations which means that the attention is spread more over the 7x7 plane. Figure 9.1 shows a visualisation of these results for "under" and "over". The results also show that landmarks in most cases receive less visual attention in comparison to targets. This indicates that after providing a target and a spatial relation, the landmark is more predictable from the language model (for a similar observation see Dobnik et al. (2018)).

Descriptions **Spatial Relations**	Average $(1 - \beta_t)$ **TRG, REL, LND**	Average $(max(\hat{\alpha_t}) - min(\hat{\alpha_t}))$ **TRG, REL, LND**
under	0.84, **0.73**, 0.79	0.0252, 0.0151, **0.0139**
front	0.83, **0.70**, 0.82	0.0230, **0.0136**, 0.0154
next	0.82, **0.68**, 0.78	0.0224, **0.0136**, 0.0138
back	0.85, **0.68**, 0.84	0.0332, **0.0186**, 0.0272
in	0.82, **0.68**, 0.77	0.0250, **0.0149**, 0.0164
on	0.81, **0.68**, 0.75	0.0249, **0.0154**, 0.0175
near	0.80, **0.67**, 0.76	0.0221, **0.0133**, 0.0169
over	0.77, **0.62**, 0.75	0.0205, **0.0133**, 0.0193
above	0.73, **0.64**, 0.77	0.0167, **0.0134**, 0.0231

Table. 9.2: The average score of adaptive visual attention for target (TRG) relation (REL) landmark (LND) triplets per each relation in the first column and the average difference between the highest and the lowest value of visual attention for the same items in the second column.

9.5 Discussion and Conclusion

In this paper we explored to what degree adaptive attention is grounding spatial relations. We have shown that adaptive visual attention is more important for grounding objects but less important for grounding spatial relations which are not directly represented with visual features. As a result

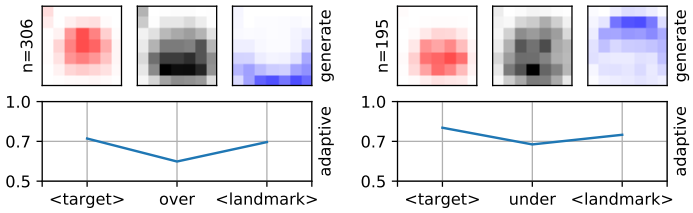


Figure. 9.1: Each square in a box in the first row represents an averaged attention for a location in the 7x7 grid over all n generated samples ($\hat{\alpha}$). The colours fade to white with lower values. The bottom graphs show their average over the entire plane, indicating the degree of adaptive visual attention $(1-\beta)$, also reported in Table 9.2.

the visual attention is diffused over a larger space. The cause for a wider attended area can be due to high degree of noise in visual features or lack of evidence for visual grounding.

This is a clear shortcoming of the image captioning model, as it is not able to discriminate spatial relations on the basis of geometric relations between the objects, for example between relations such as "left" and "right". The future work on generating image descriptions therefore requires models where visual geometry between objects is explicitly represented as in Coventry et al. (2005). The study also shows that when generating spatial relations, a significant part of the information is predicted by the language model. This is not necessarily a disadvantage. The success of distributional semantics shows that language models with word embeddings can learn a surprising amount of semantic information without access to visual grounding. As mentioned in the introduction, spatial relations do not depend only on geometric arrangement of objects but also functional properties of objects. For example, Dobnik et al. (2018) demonstrate that neural language models encode such functional information about objects when predicting spatial relations. Since, each spatial relation has different degree of functional and geometric bias Coventry and Garrod (2004), the adaptive attention considering visual features and textual features is also reflective of this aspect.

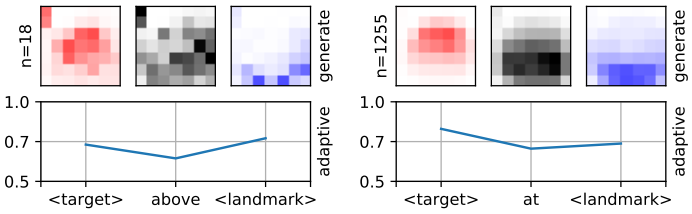
Models for explaining language model predictions such as Park et al. (2016) are also related to this study and its future work.

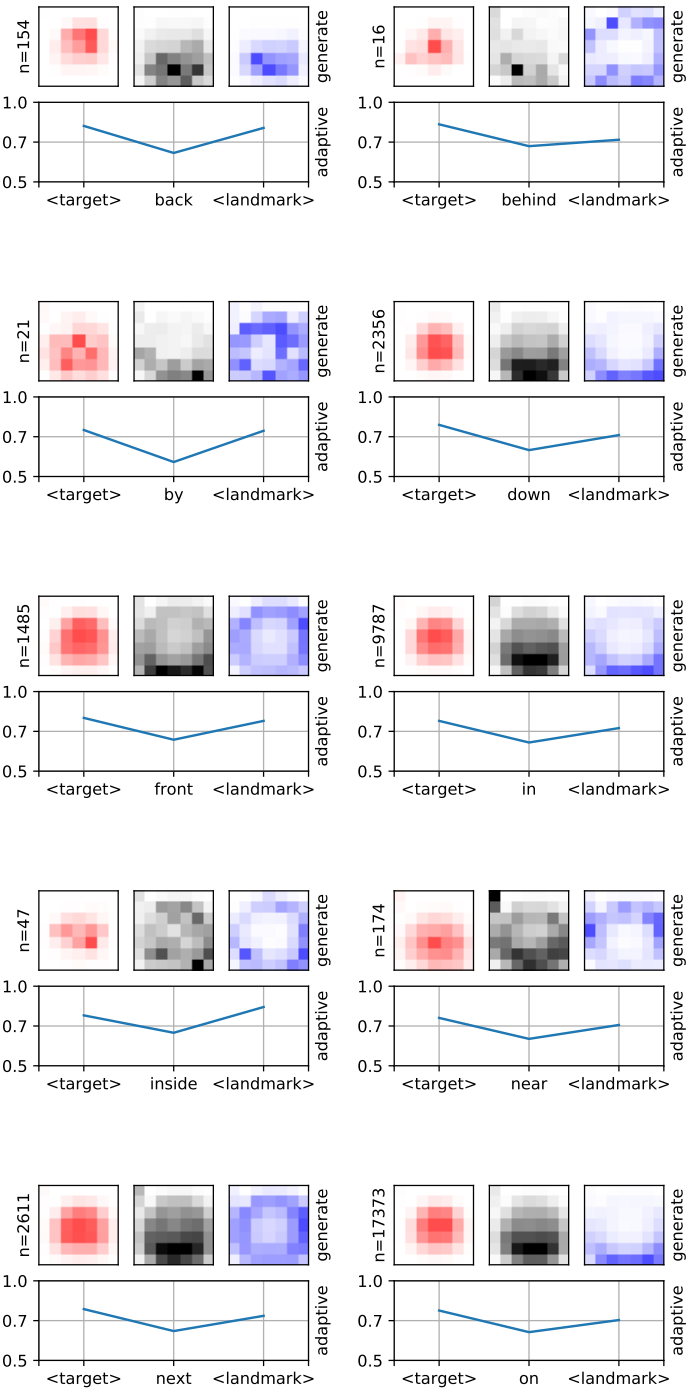
Our study focused on the adaptive attention in Lu et al. (2017) which explicitly models attention as a focus on visual and language features. However, further investigations of other types of models of attention could be made and this will be the focus of our future work. We expect that different models of attention will behave similarly in terms of attending visual features on spatial relations because the way visual features are represented: they favour detection of objects and not their relative geometric arrangement. Our future work we will therefore focus on how to formulate a model to be able to learn such geometric information in an end-to-end fashion. Methodologies such as Ribeiro et al. (2016) and Selvaraju et al. (2017) which investigate the degree of effectiveness of features without attention are also possible directions of the future studies.

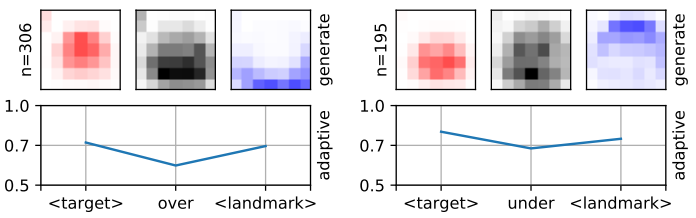
9.6 Acknowledgements

We are also grateful to the anonymous reviewers for their helpful comments on our earlier draft. The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

9.7 Appendix: Supplementary Material




n=154
generate
1.0
0.7
0.5
adaptive
<target>
back
<landmark>
n=16
generate
1.0
0.7
0.5
adaptive
<target>
behind
<landmark>
n=21
generate
1.0
0.7
0.5
adaptive
<target>
by
<landmark>
n=2356
generate
1.0
0.7
0.5
adaptive
<target>
down
<landmark>
n=1485
generate
1.0
0.7
0.5
adaptive
<target>
front
<landmark>
n=9787
generate
1.0
0.7
0.5
adaptive
<target>
in
<landmark>
n=47
generate
1.0
0.7
0.5
adaptive
<target>
inside
<landmark>
n=174
generate
1.0
0.7
0.5
adaptive
<target>
near
<landmark>
n=2611
generate
1.0
0.7
0.5
adaptive
<target>
next
<landmark>
n=17373
generate
1.0
0.7
0.5
adaptive
<target>
on
<landmark>


n=306
generate
1.0
0.7
0.5
adaptive
<target>
over
<landmark>
n=195
generate
1.0
0.7
0.5
adaptive
<target>
under
<landmark>

Bibliography

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48.

Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. 2014. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Kenny R. Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V. Richards. 2005. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In Christian Freksa, Markus Knauff, Bernd Krieg-Brückner, Bernhard Nebel, and Thomas Barkowsky, editors, *Spatial Cognition IV. Reasoning, Action, Interaction*, volume 3343 of *Lecture Notes in Computer Science*, pages 98–110. Springer Berlin Heidelberg.

Kenny R Coventry and Simon C Garrod. 2004. *Saying, seeing, and acting: the psychological semantics of spatial prepositions*. Psychology Press, Hove, East Sussex.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial*

Language Understanding (SpLU 2018) at NAACL-HLT 2018, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik and John D. Kelleher. 2017. Modular networks: An approach to the top-down versus bottom-up dilemma in natural language processing. *Forthcoming in Post-proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML)*, 1(1):1–8.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

John D Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. *CLASP Papers in Computational Linguistics*, page 41.

Barbara Landau and Ray Jackendoff. 1993. "what" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, 16(2):217–238, 255–265.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan L Yuille. 2017. Attention correctness in neural image captioning. In *AAAI*, pages 4176–4182.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2016. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757*.

Arnau Ramisa, Josiah Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220.

Terry Regier. 1996. *The human semantic potential: spatial language and constrained connectionism*. MIT Press, Cambridge, Massachusetts, London, England.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra, et al. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626.

Ravi Shekhar, Sandro Pezzelle, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017a. Vision and language integration: Moving beyond objects. In *IWCS 2017—12th International Conference on Computational Semantics—Short papers*.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017b. FOIL it! find one mismatch between image and language caption". In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 255–265.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew

Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

10 Generating Descriptions With Top-Down Spatial Knowledge

Mehdi Ghanimifard and Simon Dobnik. What Goes Into A Word: Generating Image Descriptions With Top-Down Spatial Knowledge. *In Proceedings of the 12th International Conference on Natural Language Generation. 2019.*

Abstract Generating grounded image descriptions requires associating linguistic units with their corresponding visual clues. A common method is to train a decoder language model with attention mechanism over convolutional visual features. Attention weights align the stratified visual features arranged by their location with tokens, most commonly words, in the target description. However, words such as spatial relations (e.g. *next to* and *under*) are not directly referring to geometric arrangements of pixels but to complex geometric and conceptual representations. The aim of this paper is to evaluate what representations facilitate generating image descriptions with spatial relations and lead to better grounded language generation. In particular, we investigate the contribution of four different representational modalities in generating relational referring expressions: (i) (pre-trained) convolutional visual features, (ii) spatial attention over visual features, (iii) top-down geometric relational knowledge between objects, and (iv) world knowledge captured by contextual embeddings in language models.

10.1 Introduction

Spatial recognition and reasoning are essential bases for visual understanding. Automatically generating descriptions of scenes involves both recognis-

ing objects and their spatial configuration. This project follows up on recent attempts to improve language generation and understanding in terms of using spatial modules in the fusion of vision and language Xu et al. (2015); Johnson et al. (2016); Lu et al. (2017); Hu et al. (2017); Anderson et al. (2018) (see also Section 10.6).

Generating spatial descriptions is an important part of the image description task which requires several types of knowledge obtained from different modalities: (i) invariant visual clues for object identification, (ii) geometric configuration of the scene representing relations between objects relative to the size of the environment (iii) object-specific functional relations that capture interaction between them and are formed by our knowledge of the world for example *an umbrella is over a man* is true if the referring umbrella serves its function, protecting the man from the rain Coventry et al. (2001), and (iv) for projective relations (e.g. "to the left of" and "above") but not topological relations (e.g. "close" and "at"), the frame of reference which can be influenced from other modalities such as scene attention and dialogue interaction Dobnik et al. (2015). Work in cognitive psychology Logan (1994, 1995) argues that while object identification may be pre-attentive, identification of spatial relations is not and is accomplished by a top-down mechanisms of attention after the objects have been identified. It is also the case that we do not identify all possible relations between objects but only those that are attended by such top-down mechanisms considering different kinds of high-level knowledge.

Experiments on training neural recurrent language models in a bottom-up fashion from data[1] demonstrated that spatial relations are frequently not learned to be grounded in visual inputs Lu et al. (2017); Tanti et al. (2018a); Ghanimifard and Dobnik (2018) which has been attributed to the design choices of these models that primarily focus on identification of objects Kelleher and Dobnik (2017). Therefore, targeted integration of different modalities is required to capture the properties from (i) to (iv). We can do this top-down Anderson et al. (2018); Hu et al. (2017); Liu et al. (2017). However, it is not immediately obvious *what* kind of top-down spatial knowledge will benefit the bottom-up models most. Therefore, in this paper we investigate the integration of different kind of top-down

[1] A bottom-up learning acquires higher level representations from examples of local features rather than using an external procedure to extract them. See also Section 10.6.

spatial knowledge beyond object localisation represented as features with the bottom-up neural language model.

The paper is organised as follows. In Section 10.2, we discuss how spatial descriptions are constructed and what components are required to generate descriptions. In Section 10.3, the neural networks' design is explained. In Section 10.4, we explain what dataset is used for this study, what pre-processing was applied on it and how the models are trained. Then the experiments and evaluation results are presented in Section 10.5. The related work in relation to our methods and findings is discussed in Section 10.6. The conclusion is given in Section 10.7.

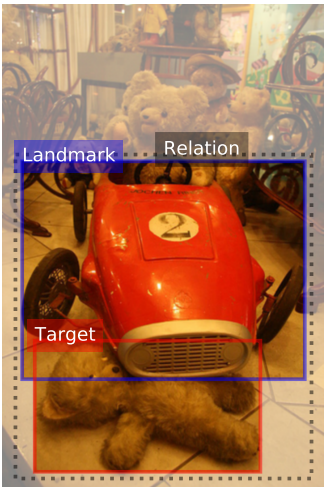
10.2 Generating Spatial Descriptions

When describing a scene, there are several ways to construct spatial descriptions referring to objects and places and their relation with each other. A spatial description has three parts: a TARGET and a LANDMARK referring to objects or places and a RELATION denoting the location of the target in relation to the landmark Logan and Sadler (1996).[2] These are in the example in Figure 10.1 as follows:

There is $\underbrace{a\ teddy\ bear}_{\text{TARGET}}\ \underbrace{partially\ under}_{\text{RELATION}}\ \underbrace{a\ go\ cart}_{\text{LANDMARK}}$.

Therefore generating such description requires (a) identification of objects and their locations: the target is what we want to describe and the landmark is what we will relate the target to; the salience of the landmark is important for the hearer. (b) Grounding of the relation in geometric space: the spatial relation is expressed relative to the landmark which grounds a 3-dimensional coordinate system; furthermore, for projective relations, the coordinate system is aligned with the orientation of the external viewpoint which determines the frame of reference Maillat (2003). (Viewpoint may also be the landmark object itself in which case the coordinate system is oriented in the same way as the landmark). (c) Grounding in function: a spatial relation may be selected also based on the functional properties between target and

[2] Sometimes these are also known as *referent* and *relatum* Miller and Johnson-Laird (1976), *figure* and *ground* Talmy (1983) or *the located object* and *the reference object* Herskovits (1986); Gapp (1994); Dobnik (2009).



⟨ “teddy bear", “partially under", “go cart"⟩

Figure. 10.1: ⟨TARGET, RELATION, LANDMARK⟩ annotation of bounding boxes in VisualGenome 2318741[a]

[a] RaSeLaSeD_Il_Pinguino (2008): CC BY-SA 2.0.

landmark objects, e.g. the difference between “*the teapot is over the cup*" and “*the teapot is above the cup*" Coventry et al. (2001).

Generating spatial descriptions requires knowing the intended target object and how we want to convey its location to the listener. The bottom-up approach in image captioning is focused on learning the salience of objects and events to generate captions expressed in the dataset (e.g. Xu et al. (2015)). The combination of bottom-up and top-down approaches for generating descriptions use modularisation in order to improve the generation of descriptions of different kind (e.g. You et al. (2016)). However, as we have seen in the preceding discussion, the generation of spatial descriptions requires a highly specific geometric knowledge. How is this knowledge approximated by the bottom-up models? To what degree can we integrate this knowledge with the top-down models? In this paper, we investigate these questions in a language generation task by comparing different variations of included top-down spatial knowledge. More specifically, for each image, we generate a description for every pair of objects that are localised in the image. We consider a variety of top-down spatial knowledge representations about objects as inputs to the model: (a) explicit object localisation and extrac-

tion of visual features; (b) explicit identification of the target-landmark by specifying their order in the feature vector; and (c) explicit geometric representation of objects in a 2D image. We investigate the contribution of each of these sets of features to generation of image descriptions.

10.3 Neural Network Design

Our method is to add step-by-step modules and configurations to the network providing different kind of top-down knowledge in Section 10.2 and investigating the performance of such configurations. There are several design choices with small effects on the performance but costly in terms of parameter size Tanti et al. (2018b). Therefore, if there is no research question related to that choice, we take the simplest choice as reported in the previous work such as Lu et al. (2017); Anderson et al. (2018). We use the following configurations:

1. Simple bottom-up encoder-decoder;
2. Bottom-up object localisation with attention;
3. Top-down object annotated localisation;
4. Top-down target and landmark assignment;
5. Two methods of top-down representation of geometric features (s-features).

These five configurations give us 10 variations of the model design as shown in Table 10.1. A detailed definition of each module is given in the Appendix 10.8 in the supplementary material.

Model name	Regions Of Interest	TARGET-LANDMARK	s-features	Architecture
$simple$	-	-	-	Figure 10.3a
$bu49$	Bottom-up (7x7 grid)	Bottom-up attention	-	Figure 10.3b
$bu49 + mask$	Bottom-up (7x7 grid)	Bottom-up attention	Multi-hot 98	Figure 10.3c
$bu49 + VisKE$	Bottom-up (7x7 grid)	Bottom-up attention	Dense 11	Figure 10.3c
td	Top-down (2 bbox)	Bottom-up attention	-	Figure 10.3d
$td + mask$	Top-down (2 bbox)	Bottom-up attention	Multi-hot 98	Figure 10.3e
$td + VisKE$	Top-down (2 bbox)	Bottom-up attention	Dense 11	Figure 10.3e
$td\ order$	Top-down (2 bbox)	Top-down assign.	-	Figure 10.3d
$td\ order + mask$	Top-down (2 bbox)	Top-down assign.	Multi-hot 98	Figure 10.3e
$td\ order + VisKE$	Top-down (2 bbox)	Top-down assign.	Dense 11	Figure 10.3e

Table. 10.1: The 10 variations of the neural network model after incrementally adding modules and features.

Generative language model We use a simple forward recurrent neural model with cross-entropy loss in all model configurations.

Simple encoder-decoder An encoder-decoder architecture without spatial attention shown in Figure 10.3a and similar to Vinyals et al. (2015) is the simplest baseline for fusing vision and language. The input to the model is an image and the start symbol $< s >$ of a description and the output is produced by the language model decoder. The embeddings are randomly initialised and learned as a parameter set of the model. The visual vectors are produced by a pre-trained ResNet50 He et al. (2016). A multi-layer perceptron module (F_v in Figure 10.2) is used to fine-tune the visual features.

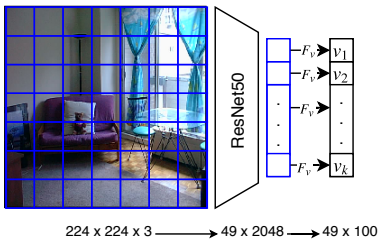


Figure. 10.2: Visual features are obtained from the pre-trained ResNet50, then translated to a low dimensional vector with a dense layer F_v.

Bottom-up localisation With visual feature representing all regions of the image as in Figure 10.2, the attention mechanism is used as a localisation module. We generalised the adaptive attention introduced in Lu et al. (2017) to be able to fuse the modalities. As shown in Figure 10.3b, the interaction between the attention mechanism and the language model is more similar to Anderson et al. (2018): two layers of stacked LSTM, the first stack (LSTM_a) to produce the features for the attention model and the second stack (LSTM_l) to produce contextualised linguistic features which are fused with the attended visual features. This design is easier to extend with additional top-down vectors.

Top-down localisation Unlike the bottom-up unsupervised localisation, the top-down method includes a provision of a list regions of interest (ROI) from external procedures. For example, the region proposals can come from another bottom-up task as in Anderson et al. (2018); Johnson et al. (2016) which use a Faster R-CNN Ren et al. (2015) to extract possible regions of interest from the ConvNets regions in Figure 10.2. Here, as shown in

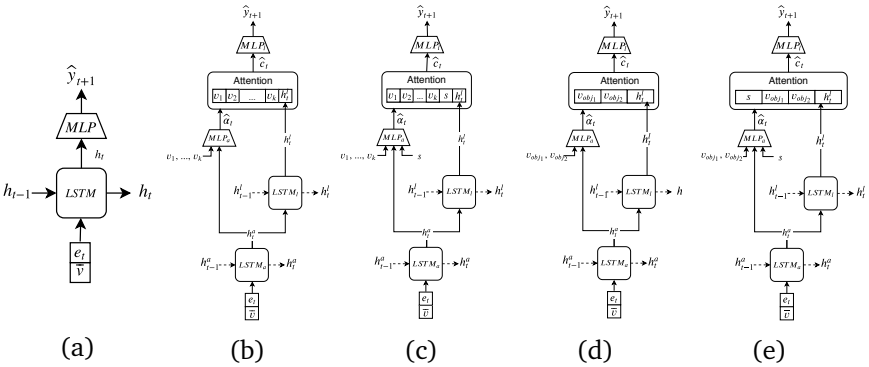


Figure. 10.3: Five architectures: (a) simple encoder-decoder ($simple$). (b) bottom-up localisation with adaptive attention on 49 regions ($bu49$). (c) bottom-up localisation with explicit spatial vectors of the bounding boxes $bu49+mask/bu49+VisKE$. (d) top-down localisation with attentions on two bounding boxes (td). (e) top-down localisation augmented with explicit spatial vectors of the bounding boxes ($td+mask/td+VisKE$).

Figure 10.4 we use the bounding box annotations of objects in images as the top-down localisation knowledge and then extract ResNet50 visual features from these regions. In the first stage the top-down visual representation only proposes visual vectors of the two objects in a random order without their spatial role as targets and landmarks in the descriptions. The model is shown in Figure 10.3d.

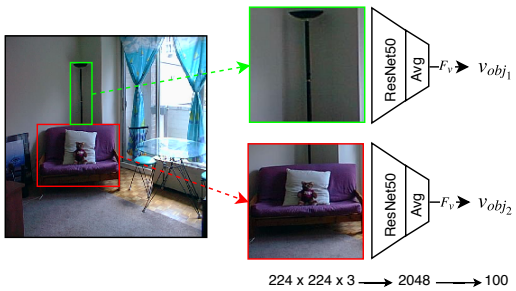


Figure. 10.4: Top-down localisation of objects with bounding boxes whose visual features are extracted and translated to lower dimensions with F_v.

Top-down target-landmark assignment In the second iteration of the top-down localisation module we assign semantic roles to regions as targets and landmarks. This is directly related to localisation as spatial relations are

asymmetric. We encode this top-down knowledge by fixing the order of the regions in the feature vector. The first object is the target and the second object is the landmark. Otherwise, the model is the same as in the previous iteration shown in Figure 10.3d.

Top-down geometric features The localisation procedure of objects discussed previously does not provide any geometric information about the relation between the two regions. However, top-down geometric features are required for grounding spatial relations where the location of the target object is expressed relative to the landmark. For example, a simple (but by no means sufficient) geometric relation between two bounding boxes can be represented by an arrow from the centre of one bounding box to the centre of the other and by ordering the information about bounding boxes in the feature vector as in the previous model to encode target-landmark asymmetry. The network architecture of the model with top-down geometric features expressing relations between the objects is shown in Figure 10.3e. We consider two different representations of the top-down geometric features shown in Figure 10.5: Multi-hot mask over 49 vectors independently for target and landmark ($Mask$) over 49 locations (Figure 10.5a) and $VisKE$ Sadeghi et al. (2015) dense representations with 11 geometric features (Figure 10.5b) where dx, dy are changes in the coordinates of the centres, ov, ov_1, ov_2 the overlapping areas (total, relative to the first, and the second bounding box), h_1, h_2 heights, w_1, w_2 widths and a_1, a_2 areas. Note that $Mask$ features provide geometric information about the size and the location of objects relative to the picture frame and $VisKE$ feature provide more detailed geometric information that expresses the relation between the objects. The latter therefore more closely match the features that were identified in spatial cognitive models. A feed-forward network with two layers (F_s) is used to project geometric features into a vector with the same dimensionality as the F_v outputs so that different modalities are comparable in weighted sum model of attention.

10.4 Dataset and Training

We use the relationship dataset in Visual Genome Krishna et al. (2017) which is a collection of referring expressions represented as triplets $\langle \text{subject}, \text{predicate}, \text{object} \rangle$ on 108K images. Unlike image captioning datasets such as MSCOCO Chen

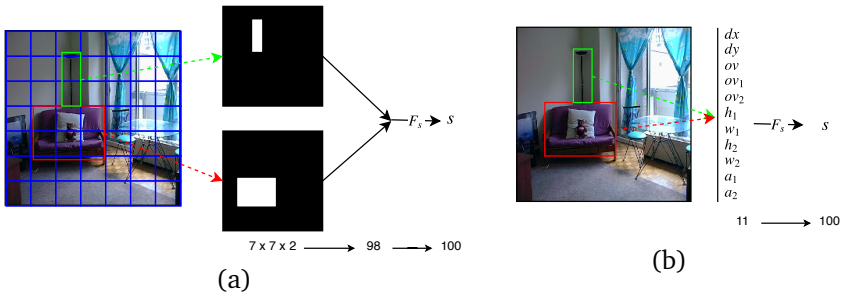


Figure. 10.5: (a) Each bounding box is converted to a mask of multi-hot vector on 49 regions. (b) The geometric relation between the two bounding boxes are represented with features from Sadeghi et al. (2015).

et al. (2015) and Flickr30K Plummer et al. (2015) where only 5 captions are given for each image, each image in this dataset is annotated with 50 phrases. The annotators were asked to annotate relations between two given bounding boxes of subject and object by freely writing the text for each of the three parts of the annotation. The bounding boxes produced by another annotation procedure which detected objects in the images. In total, there are $2,316,104$ annotations of $664,805$ unique triplets, $35,744$ unique labels of subjects and $21,299$ unique labels of objects most of which consist of multiple tokens. We omit all repetitions of triplets on each image, this leaves total $1,614,055$ annotations.[3]

Spatial relations Based on the lists of spatial prepositions in (Landau, 1996) and (Herskovits, 1986), we have created a dictionary of spatial relations and their possible multi-word variants including their composite forms. This dictionary contains $7,122$ entries of 235 relations (e.g. *right* to represent both *on the right hand side of* and *to the right of*). Of these only 202 are found in Visual Genome dataset covering 79 spatial relations. $328,966$ unique triplets in Visual Genome are based on exactly one of these terms which covers 49.4% of all possible relationships.[4]

[3] The repetitions include reflexive expressions (e.g. *horse next to horse*), annotations of several objects of the same type (e.g. *cup on table*), and repetitions due to several bounding box annotations of the same objects with different sizes.

[4] Other triplets in Visual Genome also have spatial content. Some of them include modifiers such as *partially under* as in Figure 10.1 and some of them are descriptions of an event or an action such as *sitting on* and *jumping over*. Some annotated relationships are verbs such as *flying* with less obvious spatial denotation. The spatial bias in the dataset was studied in Collell et al. (2018). The most frequent spatial relation in the dataset is "*on*" (over 450K instances), the second place is "*in*" (150K instances), then "*with*", variations of "*behind*", "*near*", "*top*", "*next*", "*under*", "*front*", and "*by*" (less than 10K instances each).

Bounding boxes Each bounding box is a tuple of 4 numbers (x, y, w, h). We normalise the numbers to the range of $(0, 1)$ relative to the image size to create geometric feature vectors (Section 10.3). The image is split into a grid with 7x7 cells to which bounding boxes are mapped, one bounding box potentially covering more than one cell. With this bounding box granularity, there are exactly $308,330$ possible bounding boxes. However, only $151,974$ are observed in the relationships dataset. The spatial distribution of paired objects reflects how natural pictures are framed and how related objects are understood by annotators.

Pre-processing We first removed duplicate triplets describing the same image. Then we converted each triplet into a word sequence by concatenating the strings and de-tokenising them with the white space separator. This produced a corpus with a vocabulary of $26,530$ types with a maximum sequence length of 16 tokens and on average 15 referring expressions per image. We use 95% of the descriptions for training and 5% for validation and testing (5,230 images with 80,231 triplets).

Training We use Keras Chollet et al. (2015) with TensorFlow backend Abadi et al. (2015) to implement and train all of the neural network architectures in Section 10.3. The models are trained with the Adam optimiser Kingma and Ba (2014) ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) with a batch size of 128 and 15 epochs.

10.5 Evaluation

All implementations are available online[5].

10.5.1 Qualitative Examples

Figure 10.6 shows generated descriptions for two examples of unseen pictures from the test dataset by five models. The generated word sequence is that with the lowest loss using beam search with $k = 5$. The first example shows exactly how top-down localisation of objects is important especially if the goal is to refer to specific objects in the scene. In the second example, the visual features inside the bounding box are confusing for all 5 models. More examples are in Figure 10.13 in the Appendix.

[5] https://gu-clasp.github.io/generate_spatial_descriptions/

⟨ “bat", “over", “shoulder"⟩

simple	player
bu49	man wearing shirt
td	bat in hand
td order	bat in hand
td order + *VisKE*	bat in hand

⟨ “hood", “above", “oven"⟩

simple	window
bu49	pot on stove
td	oven has door
td order	vent above sink
td order + *VisKE*	cabinet has door

Figure. 10.6: From VisualGenome: 2412051[6] 2413282[7]

10.5.2 Overall Model Performance

Hypothesis Top-down spatial knowledge improves the model performance. We consider three categories of top-down spatial knowledge: (i) top-down localisation of regions of interest; (ii) top-down assignment of semantic roles to regions; and (iii) two kinds of geometric feature vectors.

Method After training the models we evaluate them by calculating the average word level cross-entropy loss on held out instances in the test set[8]. We also calculate the loss on descriptions containing specific spatial relations for qualitative understanding of the effects of each type of top-down knowledge.

Results The overall loss of each model on the unseen descriptions of images is shown in Figure 10.7. The fully bottom-up model with no spatial attention ($simple$) has the highest loss. The loss in the variations of the model with bottom-up localisation in $bu49$ is higher than the one in the models with top-down localisation. The models with the top-down assignment of TARGET-LANDMARK achieves the best results. The effect of top-down geometric features is not significant.

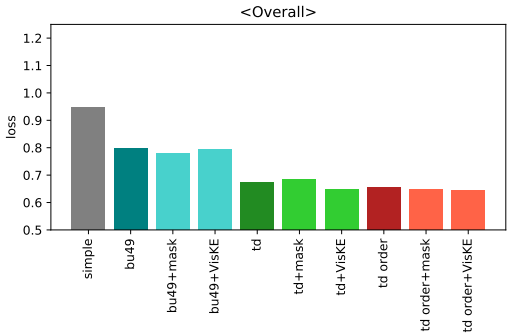


Figure. 10.7: Cross-entropy loss of different model configurations on evaluation data.

Figure 10.8 shows the performance of the models on a selection spatial relations.

Discussion The top-down localisation (td) certainly improves the performance of the language models compared to purely bottom-up representations. However, additional top-down assignment of TARGET-LANDMARK ($td\ order$) and their additional geometric arrangement of bounding box features ($mask$ and $VisKe$) has a small positive effect on overall performance. The overall performance is not a representative of how these configurations effect the grounding of spatial relations. More specifically, the imbalance of certain groups of relations (especially a generally lower proportion of

[8]Equivalent to log-perplexity of the language model.

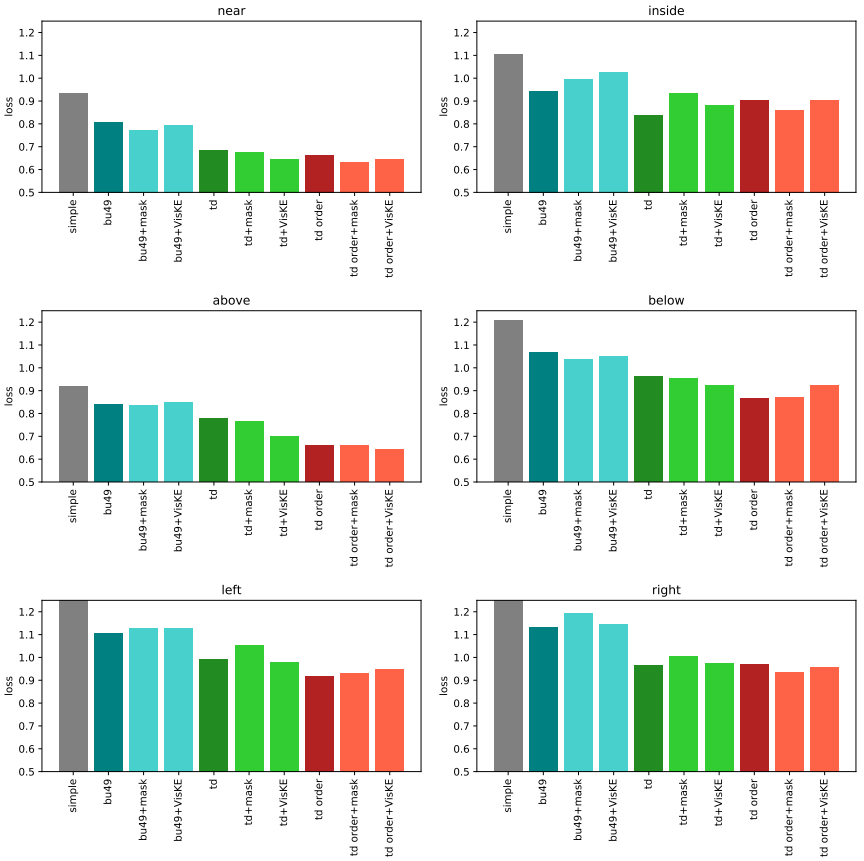


Figure. 10.8: Cross-entropy loss of different model configurations on 40 descriptions for each relation: *near*, *inside*, *above* and *below*.

geometrically biased relations such as "left" and "right" in this dataset and the presence of relations with a minimum spatial content such as *has*, *wearing*) makes it harder to make conclusions about overall performance of the models. We further examine two groups of some frequent spatial relations. The relations such as *inside* and *near* represent one group and *above* and *below* represent the other. Some top-down knowledge (as represented by our features) is less informative for the first group but is informative for the second group. For example *near* does not require the assignment of TARGET-LANDMARK roles. We observe that $td\ order$ is not performing better than td. On the other hand, *inside* is sensitive to TARGET-LANDMARK assignment. However, since the relation is also restricted by a choice of objects (only

certain objects can be inside others) their TARGET-LANDMARK assignment may already be inferred without such top-down knowledge from a language model. For the second group, the top-down knowledge about the semantic role of objects is important. However, *left* and *right* are among the least frequent relations in the dataset which is demonstrated by the fact that their descriptions have a higher loss than *above* and *below*. For these relations the loss of the $simple$ model is much higher than other configurations. It can be seen that td is performing better than bu and $td\ order$ is contributing over td but geometric features have a lesser effect than identification of semantic roles ($td\ order$).

10.5.3 Grounding in features

Hypotheses With the aim to evaluate *what* top-down information contributed to grounding of words we examine the following hypotheses:

- H1 s-features contribute to predicting spatial relation words.
- H2 Without top-down TARGET-LANDMARK role assignments to each region, attention is uniformly distributed over region choices at the beginning of a sequence generation.

Method In order to check the contribution of each feature from different modalities in prediction of each word, we look at the adaptive attention on each feature at the point of predicting the word[9]. Since feature vectors are not normalised against the number of features of each modality, we first multiply each attention measure with the magnitude of the feature vector, and then we normalised it to sum to 1 again:

$$\beta_{t,f_i} = \frac{\alpha_{t,f_i}||f_i||}{\sum_j \alpha_{t,f_j}||f_j||} \tag{10.1}$$

where t refers to the time in the word sequence, and f_i is the feature the attention of which α_{t,f_i} is applied to it. We report the average β_{t,f_i} over the instances in the validation dataset.

Figure 10.9 shows β on two examples in three models. For each word, the bar chart is divided between four features (in Figure 10.3e): (1) target v_{obj_1}

[9] In this experiment, we do not check if the estimated likelihood for the correct word is the highest predicted score. The generated descriptions may still be acceptable with an alternative spatial relation. Furthermore, in the following analysis we report the attention over semantic roles and not individual words.

(2) landmark v_{obj_2} (3) s-features for bounding boxes (4) contextualized embeddings h^l.

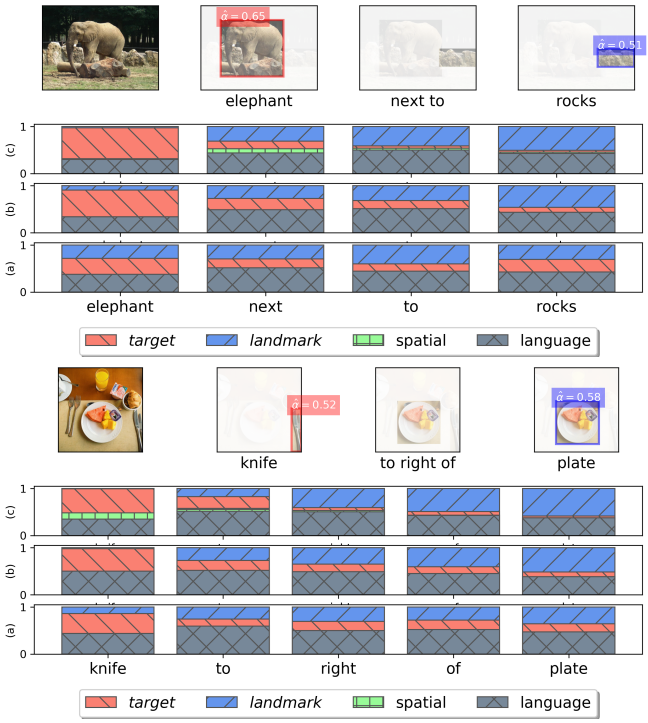


Figure. 10.9: β is plotted in bar charts for each word. (a) $td\ order + VisKE$ (b) $td + VisKE$ (c) td. The values of β for each word that constitute description referring to each bounding box region is given in images.

After measuring the normalised attention on each feature according to Equation 10.1, we report the average of attentions on each token at that time step of the word sequence. We also group the tokens based on their semantic role in the triplets and report the average β on these tokens for a given role.

Results The average of attentions over triplets of tokens is plotted in Figure 10.10. The behaviour of attentions on word sequences in the four models in given in Figure 10.11.

Discussion The comparison of 6 models in Figure 10.10 shows that geometric $mask$ s-features are not contributing as well as dense $VisKE$ s-features. In the models without top-down semantic role assignment only

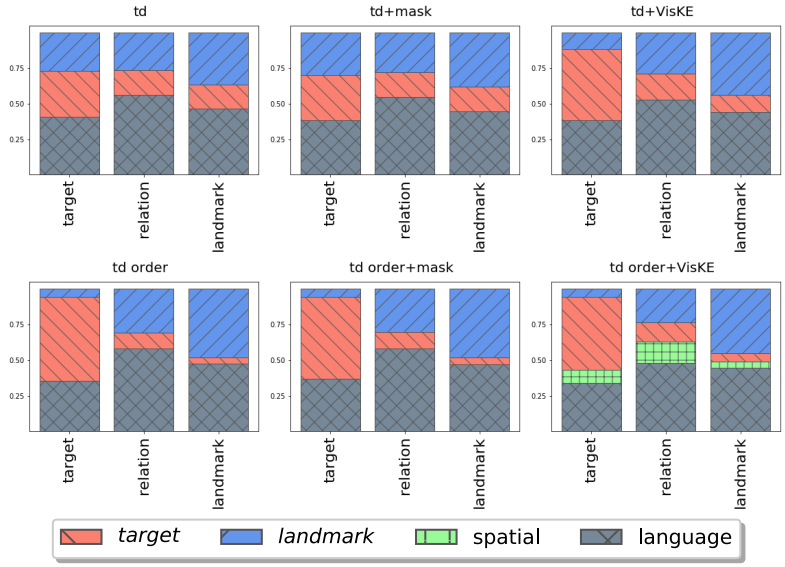
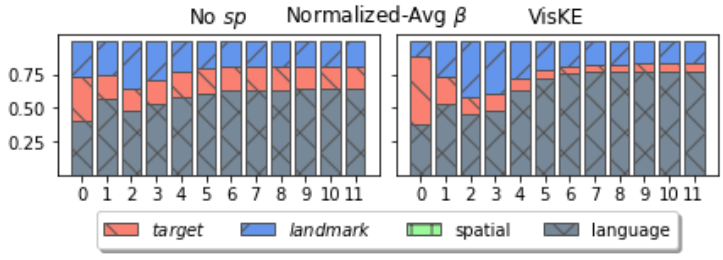


Figure. 10.10: The overall average of β on tokens of each semantic role (target, relation, landmark) on all examples of the test dataset, for 6 variations of the top-down knowledge about regions of interest (ROI): location of objects and their order as target and landmark.

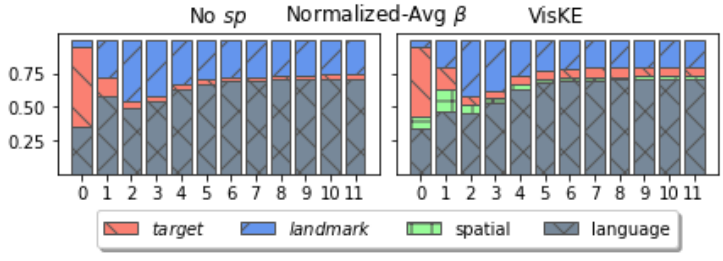
the model with $+VisKE$ features has the expected attention on target and landmark, but there is no attention on the s-features. In the models with top-down semantic role assignment, the model with $VisKE$ s-features has higher attention on s-features when predicting a relation word (H1). A similar situation is observable over word sequences in Figure 10.11. Without prior semantic role assignment the model is more confused how to attend target or landmark (H2). Finally, note that geometric $VisKE$ s-features help predicting the TARGET-LANDMARK roles when these are not assigned top-down.

10.6 Related Work

Generating referring expressions Generating locative expressions is part of the general field of generating referring expressions Dale and Reiter (1995); Krahmer and van Deemter (2011) with applications such as describing scenes Viethen and Dale (2008) and images Mitchell et al. (2012). The research on describing visible objects Mitchell et al. (2013) and human-



(a)



(b)

Figure. 10.11: The average of β attentions of top-down models over sequences of words $1 \ldots 11$ (a) comparing td and $td+VisKE$ and (b) comparing $td\ order$ and $td\ order+VisKE$.

robot dialogue Kelleher and Kruijff (2006) raised question about grounding relations in hierarchical representation of context. Application of neural language models and using convolutional neural networks for encoding visual features is an open question in interactive GRE tasks.

Encoder-decoder models with attention Recently several methods focused on finding better neural architectures for generating image descriptions based on pre-trained convolutional neural networks have been introduced. Karpathy and Fei-Fei (2015) align descriptions with images. Vinyals et al. (2015) introduce an encoder-decoder framework. Xu et al. (2015) improve this approach with spatial attention. Lu et al. (2017) introduce adaptive attention that balances language and visual embeddings. The attention measure provides an explanation of encoder-decoder architectures on how each modality contributes to language generation. Based on the attended features the performance of these models can be examined Liu et al. (2017); Ghanimifard and Dobnik (2018). In our paper, we develop a model similar to the adaptive attention which exploits its expressive aspects as a degree of grounding in different features.

Outputs of external models as top-down features In another line of work, the output of the bottom-up visual understanding is used as top-down features for language generation. For example, an object detection pipeline is combined explicitly with language generation. This procedure was previously used in template-based language generation Elliott and Keller (2013); Elliott and de Vries (2015). There have been attempts to combine this process with neural language models with attention. For example, You et al. (2016) extract candidate semantic attributes from images (e.g. a list of objects in the scene), then the attention mechanism is used to learn to attend on them when generating tokens of image descriptions. Instead of semantic attributes, Anderson et al. (2018) use a region proposal network from a pre-trained object detection model to extract the generated bounding box regions as possible locations of visual clues. Then, the attention model learns to attend on the visual features associated with these regions. The idea of using an object detection module is also used in Johnson et al. (2016) where Faster R-CNN Ren et al. (2015) is used to find regions of interest. Instead of assigning one object class to each region, a full description is generated for each proposed region. In all of these models, an image understanding module extracts some proposed representations and then this knowledge is used as a top-down representation of the scene to generate an image description. In this paper, we investigate the extent to which different spatial information is facilitating as a top-down knowledge to generate descriptions of scenes with neural language models.

Modular design Our paper examines strategies that can demonstrate language grounding within a neural architecture. The studies of neural architectures such as Tanti et al. (2018b) provide analytical insight on differences between multimodal architectures for language generation. The modular design is mostly used in language parsing tasks such as Hu et al. (2017) where object recognition, localisation and relation recognition are separate modules for grounding different parts of image descriptions in images in order to solve tasks such as visual question answering. In our paper, the modularity of the neural architecture is not focused on parsing text but used to incrementally demonstrate the contribution of each introduced modality to language generation.

Multimodal embeddings There are related studies on learning multimodal embeddings Kiros et al. (2014); Lazaridou et al. (2015) to represent

vision and language in the same semantic space. The focus of our paper is to investigate how these different modalities complement each other in neural language generation. In our models, the semantic representations of spatial relations are considered as a separate modality extending both the language and visual embeddings. There are related studies on encoding spatial knowledge in feature space in order to predict spatial prepositions Ramisa et al. (2015) or on prepositional embeddings which can predict regions in space Collell and Moens (2018). In our paper, we investigate the degree in which each embedding contributes to language generation within the neural language model.

10.7 Conclusions

We explored the effects of encoding top-down spatial knowledge in a bottom-up trained generative neural language model for the image description task. The findings of the experiments in this paper are as follows:

(1) Overall, integration of top-down knowledge has a positive effect on grounded neural language models for this task. (2) When combining bottom-up language grounding with top-down knowledge representation as different features, different types of top-down knowledge have different contribution to grounded language models. The general picture is further complicated by the fact that different spatial relations have different bias to different knowledge. (3) The performance gain from the geometric features extracted from bounding boxes (s-features) is smaller than initially expected, with two possible explanations related to the nature of the corpora of image descriptions: (i) The corpus contains images of typical scenes where the relation of objects with each other is predictable from the description and therefore is captured in the language model; (ii) As annotators are focused on describing "what is in the image" rather "where things are spatially in relation to each other", descriptions of geometric spatial relations which refer to the locational information are rare in the corpus. (4) The majority of attention is placed on the language model which demonstrates that this provides significant information when generating spatial descriptions. While this may be a confounding factor if the visual features are ignored, the language model also encodes useful information about spatial information as discussed in Kulkarni et al. (2011); Dobnik et al. (2018).

The results open several questions about grounded language models. Firstly, the degree to which the system is using each modality can be affected by dataset biases and this should be taken into account in the forthcoming work. Given this bias, learning a single common language model for descriptions of spatial scenes is insufficient as different kinds of knowledge may come to focus in different interactional scenarios. This further supports the idea that top-down integration of knowledge is required where we hope that the models will learn to attend to the appropriate features. Secondly, our investigation leaves open the question whether the representations both visual and geometric that we use are good representations for learning spatial relations. Further work will include a focused investigation of what kind of geometric relations they encode.

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

10.8 Appendix: Model Details

Generative language model We use a simple forward recurrent neural model with cross-entropy loss in all model variations:

$$P(w_{t+1}|w_{0:t}, c) = \hat{y}_{t+1} = F(w_{0:t}, c; \theta) \tag{10.2}$$

$$loss(k_{1:t}, \theta) = -\sum_{t=0}^{T} \log(\delta_{k_t} \cdot \hat{y}_t) \tag{10.3}$$

where F represents the neural network function with parameters θ, inputs $w_{0:t}$ the sequence of words with w_0 the sentence marker '⟨s⟩', and c to represent the image or with additional top-down knowledge. $\hat{y}_{t+1} \in [0,1]^{|V|}$ is a categorical distribution over the choices in vocabulary V for the conditional probability of the next word. The loss is calculated for each sample of word sequence $[v_{k_0}, v_{k_1}, ..., v_{k_T}]$, which $k_t \in \{1, ...|V|\}$ refers to the word index in the vocabulary, and δ_{k_t} is its one-hot encoding.

Simple encoder-decoder An encoder-decoder architecture without spatial attention, similar to Vinyals et al. (2015), is the most simple baseline for setting up the experiments and designing the foundation for fusing vision and language. The input to the model is an image and the start symbol $< s >$ for the language model decoder. The word embeddings e_t are concatenated with the scene visual features ($\bar{v}$). The embeddings are randomly initialised and learned as a parameter set for the model. The visual vectors are produced by a pre-trained ResNet50 He et al. (2016). Then, $\bar{v}$ is made by a dense layer translating the visual vector to a unified tensor size for computational convenience. This layers also helps fine-tuning the visual features.

$$F_v(x) = \text{ReLU}(W_v \cdot x + b_v)$$
$$\bar{v} = \frac{\sum_{i=1}^{k} F_v(v'_i)}{k}$$

where F_v the function in Figure 10.2, $v'_i \in \mathbb{R}^{2048}$ with ResNet50 dimensions, $W_v \in \mathbb{R}^{100\times2048}$ and $b_v \in \mathbb{R}^{100}$ are parameters to be learned as fine-tuning. The resulting vector is concatenated to a word embedding and fed to the Long-Short Term Memory (LSTM) network Hochreiter and Schmidhuber (1997) and its output to a multi-layer perceptron (MLP) with a softmax layer which predicts the next word, as it was described earlier in Equation 10.2. This function would be:

$$\hat{y}_{t+1} = \text{softmax}(\text{MLP}(\text{LSTM}([e_t; \bar{v}], h_{t-1}))) \tag{10.4}$$

where e_t and h_t respectively represent the word embedding and the hidden unit in recurrent cell at time t of the word sequence (Figure 10.3a). Ideally, the spatial features must be learned bottom-up in $\bar{v}$ as other visual features in the deep layers of convolutions in ResNet.

Adaptive attention The simple encoder-decoder architecture relies on bottom-up learning of visual features and geometric arrangement of objects. However, it has been shown in recent image captioning models that a spatial attention mechanism to localise each word improves the language generation Xu et al. (2015). Moreover, the attentions can be learned as an adaptation of modalities. Based on this assumption we will use the adaptive attention similar to Lu et al. (2017). In generalisation of adaptive attention, the feature vectors including visual features from different locations as well as the

contextual language features and other modalities $\hat{f} = [f_1, f_2, ..., f_n]$ are fused with weighted sum according to their attention weight $\hat{\alpha}$.

$$\hat{c}_t = \sum_{i=1}^{n} \alpha_i f_i \tag{10.5}$$

where $\hat{c}_t$ represents the fused vector after applying adaptive attention on n feature vectors. Knowing which features in what degree contribute to prediction of the next word is decided in a multi-layer perceptron (MLP_a) with softmax as $\hat{\alpha}$ in Figure 10.12. This module is formalised in a sequential process as follows:

$$z_t = W_a^2 \tanh(W_a^1 \cdot \hat{f}_t)$$
$$\hat{\alpha}_t = \text{softmax}(z_t).$$

where $\hat{\alpha}_t = [\alpha_{t,1}, \alpha_{t,2}, ..., \alpha_{t,n}]$ is the output of the module in time t, and W_a^1,W_a^2 are the parameters of the module which will be trained in the model.

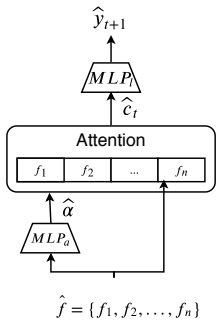


Figure. 10.12: The generalised adaptive attention module.

Bottom-up localisation With visual feature representing each region of the image as in Figure 10.2, attention mechanism is going to work as localisation model. We designed the interaction between the attention mechanism and the language model more similar to Anderson et al. (2018): two layers of stacked LSTM, the first stack (LSTM_a) to produce features for attention model, then the second stack (LSTM_l) to produce contextualised linguistic features to be fused with attended visual features (Fig-

ure 10.3b). This design makes it easier to be extended with top-down visual vectors.

$$\hat{c}_t = \sum_{i=1}^{49} \alpha_{t,i} v_i + \alpha_{t,50} h_t^l \tag{10.6}$$

where each v_i is a visual feature referring to one of the 49 locations in Figure 10.2, and h_t^l is the contextualised language feature from LSTM_l.

Top-down localisation Unlike the bottom-up localisation, the top-down method has a list of regions of interest pre-processed from other procedures. The process of region proposals can be part of a bottom-up process as in Anderson et al. (2018) or Johnson et al. (2016) which instead of the grids of regions in ConvNets in Figure 10.2 a Faster R-CNN Ren et al. (2015) is used to extract all possible regions of interest. In this paper, we use the bounding box annotations on images as the top-down localisation knowledge, then we use ResNet50 to extract visual features from these regions Figure 10.4. At this stage the top-down visual representation only proposes visual vectors of two objects in random order without their spatial role in intended descriptions shown in Figure 10.3d.

$$\hat{c}_t = \alpha_{t,1} v_{obj_1} + \alpha_{t,2} v_{obj_2} + \alpha_{t,3} h_t^l \tag{10.7}$$

where each v_{obj_1} and v_{obj_2} are the visual features referring to two regions in Figure 10.4, and h_t^l is the contextualised language feature from LSTM_l.

Top-down target-landmark assignment Another top-down information is the assignment of one region as the target and another region as the landmark. This top-down knowledge is encoded as the order in the list of two object, first object is the target and the second object is the landmark in Equation 10.7.

$$\hat{c}_t = \alpha_{t,1} v_{\text{TARGET}} + \alpha_{t,2} v_{\text{LANDMARK}} + \alpha_{t,3} h_t^l \tag{10.8}$$

where each v_{TARGET} and v_{LANDMARK} are the visual features referring to two regions in Figure 10.4 and their semantic role is defined top-down.

Top-down geometric features With top-down localisation we may lose the relative location of two objects since they are processed separately in two disconnected convolutional neural networks. Therefore, the top-down

geometric features are required for grounding of denotation of the locational words. Additionally, representing geometric knowledge can encode the frame of reference. For example, a simple geometric relation between two bounding boxes can be an arrow from the centre of one bounding box to the centre of the other, however the choice between the order of objects depends on the frame of reference (i.e. $obj_1 \rightarrow obj_2$ or $obj_1 \leftarrow obj_2$). We represent the geometric features by considering the top-down target-landmark assignment (i.e. TARGET $\rightarrow$ LANDMARK). Therefore with these feature vectors we encode the top-down frame of reference as well. This creates different variations of feature fusions (Table 10.2).

Model name	Visual features	Attention
$bu49$	$[v_1, ..., v_{49}]$	$\hat{c}_t = \sum_{i=1}^{49} \alpha_{t,i} v_i + \alpha_{t,50} h_t^l$
$bu49 + mask$	$[v_1, ..., v_{49}]$	$\hat{c}_t = \sum_{i=1}^{49} \alpha_{t,i} v_i + \alpha_{t,50} h_t^l + \alpha_{t,51} s$
$bu49 + VisKE$	$[v_1, ..., v_{49}]$	$\hat{c}_t = \sum_{i=1}^{49} \alpha_{t,i} v_i + \alpha_{t,50} h_t^l + \alpha_{t,51} s$
td	$[v_{obj_1}, v_{obj_2}]$	$\hat{c}_t = \alpha_{t,1} v_{\text{TARGET}} + \alpha_{t,2} v_{\text{LANDMARK}} + \alpha_{t,3} h_t^l$
$td + mask$	$[v_{obj_1}, v_{obj_2}]$	$\hat{c}_t = \alpha_{t,1} v_{obj_1} + \alpha_{t,2} v_{obj_2} + \alpha_{t,3} h_t^l + \alpha_{t,4} s$
$td + VisKE$	$[v_{obj_1}, v_{obj_2}]$	$\hat{c}_t = \alpha_{t,1} v_{obj_1} + \alpha_{t,2} v_{obj_2} + \alpha_{t,3} h_t^l + \alpha_{t,4} s$
$td\ (order)$	$[v_{\text{TARGET}}, v_{\text{LANDMARK}}]$	$\hat{c}_t = \alpha_{t,1} v_{\text{TARGET}} + \alpha_{t,2} v_{\text{LANDMARK}} + \alpha_{t,3} h_t^l$
$td + mask\ (order)$	$[v_{\text{TARGET}}, v_{\text{LANDMARK}}]$	$\hat{c}_t = \alpha_{t,1} v_{\text{TARGET}} + \alpha_{t,2} v_{\text{LANDMARK}} + \alpha_{t,3} h_t^l + \alpha_{t,4} s$
$td + VisKE\ (order)$	$[v_{\text{TARGET}}, v_{\text{LANDMARK}}]$	$\hat{c}_t = \alpha_{t,1} v_{\text{TARGET}} + \alpha_{t,2} v_{\text{LANDMARK}} + \alpha_{t,3} h_t^l + \alpha_{t,4} s$

Table. 10.2: The visual features and their attention

In order to find the best encoding of top-down geometric features, we considered two different vectorisation strategies to represent relation between two bounding boxes Figure 10.5.

- $(mask)$ a concatenation of two mask vectors in 49 locations (Figure 10.5a).
- $(VisKE)$ a dense representation with 11 geometric features according to Sadeghi et al. (2015) (Figure 10.5b): where dx, dy are changes in coordinates of the centres, ov, ov_1, ov_2 the overlapping areas (total, relative to the first, and the second bounding box), h_1, h_2 heights, w_1, w_2 widths and a_1, a_2 areas.

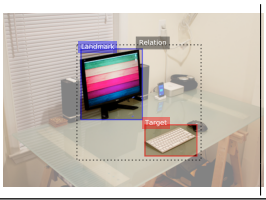
Then, a feed-forward network with two layers (F_s) is used to project geometric features into a 100-dimension vector to become comparable with other modalities.

$$F_s(x) = W_s^2 \tanh(W_s^1 \cdot x + b_s^1)$$
$$s = F_s(s')$$

where s represents the transformed geometric spatial features, and $W_s^2 \in \mathbb{R}^{100\times100}, W_s^1 \in \mathbb{R}^{100\times11} (or\ \mathbb{R}^{100\times98})$ are the set parameters regarding this module to be learned in the model.

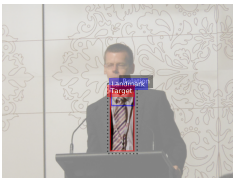
10.9 Examples of generated descriptions

More examples of generated descriptions with beam search of depth 5 are shown in Figure 10.13.



⟨ "keyboard", "in front of", "computer"⟩	
simple	computer
bu49	keyboard on desk
td	computer on top of desk
td order	keyboard on computer
td order + *VisKE*	keyboard on computer

⟨ "mirror", "in side of", "semi"⟩	
simple	truck
bu49	truck has door
td	door on truck
td order	light on road
td order + *VisKE*	mirror on truck



⟨ "lanyard", "around", "neck"⟩	
simple	tie
bu49	man has hair
td	tie around neck
td order	tie around neck
td order + *VisKE*	tie around neck

Figure. 10.13: From VisualGenome: 2413204[a] 2417890[b] 2413371[c]

[a] Schmidt (2010): CC BY-NC-SA 2.0.
[b] Yap (2008): CC BY-NC 2.0.
[c] Coghlan (2011): CC BY-SA 2.0.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *CVPR*, 3(5):6.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

François Chollet et al. 2015. Keras. `https://keras.io`.

Michael Coghlan. 2011. Tony cook. VisualGenome image id 2413371.

Guillem Collell and Marie-Francine Moens. 2018. Learning representations specialized in spatial knowledge: Leveraging language and vision. *Transactions of the Association of Computational Linguistics*, 6:133–144.

Guillem Collell, Luc Van Gool, and Marie-Francine Moens. 2018. Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Kenny R Coventry, Merce Prat-Sala, and Lynn Richards. 2001. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of memory and language*, 44(3):376–398.

Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik, Mehdi Ghanimifard, and John D. Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding (SpLU 2018) at NAACL-HLT 2018*, pages 1–11, New Orleans, Louisiana, USA. Association for Computational Linguistics.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL – Semdial 2015: The 19th Workshop on the Semantics and Pragmatics of Dialogue*, pages 24–32, Gothenburg, Sweden.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1292–1302.

Desmond Elliott and Arjen de Vries. 2015. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 42–52.

Klaus-Peter Gapp. 1994. Basic meanings of spatial relations: computation and evaluation in 3D space. In *Proceedings of the twelfth national conference on Artificial Intelligence (AAAI'94)*, volume 2, pages 1393–1398, Menlo Park, CA, USA. American Association for Artificial Intelligence, AAAI Press/MIT Press.

Mehdi Ghanimifard and Simon Dobnik. 2018. Knowing when to look for what and where: Evaluating generation of spatial descriptions with adaptive attention. In *Proceedings of the 1st Workshop on Shortcomings in Vision and Language (SiVL'18), ECCV, 2018*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Dave Herholz. 2005. Wide stance. VisualGenome image id 2412051.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1115–1124.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574.

juanjogasp. 2013. Baltic trip. VisualGenome image id 2413282.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

John D. Kelleher and Simon Dobnik. 2017. What is not where: the challenge of integrating spatial representations into deep learning architectures. In *CLASP Papers in Computational Linguistics: Proceedings of the Conference on Logic and Machine Learning in Natural Language*.

John D. Kelleher and Geert-Jan M. Kruijff. 2006. Incremental generation of spatial referring expressions in situated dialog. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1041–1048, Sydney, Australia. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *International Conference on Machine Learning*, pages 595–603.

Emiel Krahmer and Kees van Deemter. 2011. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2011. Baby talk: Understanding and generating image descriptions. In *Proceedings of the 24th CVPR*. Citeseer.

Barbara Landau. 1996. Multiple geometric representations of objects in languages and language learners. *Language and space*, pages 317–363.

Angeliki Lazaridou, Marco Baroni, et al. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017. Attention correctness in neural image captioning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Gordon D Logan. 1994. Spatial attention and the apprehension of spatial relations. *Journal of Experimental Psychology: Human Perception and Performance*, 20(5):1015.

Gordon D Logan. 1995. Linguistic and conceptual control of visual spatial attention. *Cognitive psychology*, 28(2):103–174.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6.

Didier Maillat. 2003. *The semantics and pragmatics of directionals: a case study in English and French*. Ph.D. thesis, University of Oxford: Committee for Comparative Philology and General Linguistics, Oxford, United Kingdom.

George A. Miller and Philip N. Johnson-Laird. 1976. *Language and perception*. Cambridge University Press, Cambridge.

Margaret Mitchell, Kees van Deemter, and Ehud Reiter. 2013. Generating expressions that refer to visible objects. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1184, Atlanta, Georgia. Association for Computational Linguistics.

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Arnau Ramisa, JK Wang, Ying Lu, Emmanuel Dellandrea, Francesc Moreno-Noguer, and Robert Gaizauskas. 2015. Combining geometric, textual and visual features for predicting prepositions in image descriptions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 214–220. Association for Computational Linguistics.

RaSeLaSeD_Il_Pinguino. 2008. Killer bear. VisualGenome image id 2318741.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464.

Jared Schmidt. 2010. Desk 2010-08-08. VisualGenome image id 2413204.

Leonard Talmy. 1983. How language structures space. In Herbert L. Pick Jr. and Linda P. Acredolo, editors, *Spatial orientation: theory, research, and application*, pages 225–282. Plenum Press, New York.

Marc Tanti, Albert Gatt, and Kenneth P Camilleri. 2018a. Quantifying the amount of visual information used by neural caption generators. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0.

Marc Tanti, Albert Gatt, and Kenneth P Camilleri. 2018b. Where to put the image in an image caption generator. *Natural Language Engineering*, 24(3):467–489.

Jette Viethen and Robert Dale. 2008. The use of spatial relations in referring expression generation. In *Proceedings of the Fifth International Natural Language Generation Conference*, pages 59–67. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3156–3164. IEEE.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.

Brian Yap. 2008. New england highway. VisualGenome image id 2417890.

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.

11 Learning To Compose Grounded Spatial Relations

Mehdi Ghanimifard and Simon Dobnik. Learning to Compose Spatial Relations with Grounded Neural Language Models. *In Proceedings of 12th International Conference on Computational Semantics (IWCS)-Long papers. 2017.*

Abstract Language is compositional: we can generate and interpret novel sentences by having a notion of meaning of their individual parts. Spatial descriptions are grounded in perceptional representations but their meaning is also defined by what neighbouring words they co-occur with. In this paper we examine how language models conditioned on perceptual features can capture the semantics of composed phrases as well as of individual words. We generate a synthetic dataset of spatial descriptions referring to perceptual scenes and examine how grounded language models built with deep neural networks can account for compositionality of descriptions – by evaluating how the learned language models can deal with novel grounded composed descriptions and novel grounded decomposed descriptions, constituents previously not seen in isolation.

11.1 Introduction

Representing and reasoning with linguistic meaning is a central task in computational linguistics. Here two kinds of meaning representations are used: (i) *probabilistic language models* and (ii) *meaning representations grounded* in other, typically perceptual information. Recently, there have been several approaches in deep learning that deal with both, either independently or together.

The main goal of *probabilistic language models* is to estimate a probability distribution of sequences of words based on observable samples from language production, typically by estimating conditional probabilities of words with a categorical distribution. This gives language models means for representing words as sequences with a measure of likelihood for each sequence. Neural language models perform this objective by parametrising a probability density function with parametric representations of words and functions which compose words into phrases Bengio et al. (2003); Mnih and Hinton (2007); Mikolov et al. (2010). The gradient based learning in neural networks turns the modelling problem into an optimisation problem, minimising the error or distance between a model prediction and an observable data over a list of parameters:

1. parameters representing words with feature vectors known as *word embeddings*;
2. parameters of functions composing word features into a structure;
3. parameters of projections from final composed representations to categorical probabilities which in sequential models are the next word predictions.

There have been many attempts to show that the learned word embeddings in vector spaces are good representations of meaning. Basing the argument on the distributional hypothesis, if a probabilistic model of words is conditioned on their context words (i.e. skip-grams or bag-of-words), the word embeddings must encode semantic information by having learned distances in vector spaces which correspond to semantic similarity scores obtained through relatedness tests performed by native speakers. These representations were extended to word compositions by considering different compositional functions as vector manipulations Mitchell and Lapata (2010); Coecke et al. (2010); Baroni et al. (2014). Our notion of composition in a language model is broader than this: it involves (1) distributional models of words estimated from word sequences as well as (2) their grounding into representations of physical space. This extends the Montague's notion of compositionality. Lexical representations and their compositions are not dependent on meaning postulates and lexicalised constraints but rather perceptual evidence which is (probabilistically) associated with them.

Harnad (1990); Roy (2005) define language grounding as a process of relating words with an agent's perception. The ambiguity and vagueness of grounded meanings as well as of syntactic structures suggest that the connection between language and perception is gradient and therefore probabilistic. The main approaches to probabilistic models of grounded language are probabilistic learning of grounded language and grammar Roy and Mukherjee (2005); Matuszek et al. (2012), classifiers Dobnik (2009), and feature representations in perceptual space such as colour McMahan and Stone (2015). Our proposal is in line with all three approaches.

A *grounded language model* is a language model conditioned by perceptual representations that it refers to. Ideally, the model should capture how each constituent in the composed phrase relates to some perceptual representations. For example, in an image captioning task, a grounded language model estimates a conditional probability of a word sequence $w_{1:T}$ given some image feature c that the words refers to. A general way to model word sequences is to use the chain rule as follows. The model can generate phrases and sentences step-by-step by predicting the next word in a sequence:

$$Pr(w_{1:T}|c) = \prod_{t=1}^{T} Pr(w_t|w_{1:t}, c) \tag{11.1}$$

The parametrisation of vision and language is often done by combining word-embeddings with multimodal embeddings Kiros et al. (2014); Socher et al. (2014). In the state of the art models for image captioning with encoder-decoder architecture, the encoder module is trained under the assumption that grounded words only denote features in subareas of an image, e.g bounding boxes Karpathy and Fei-Fei (2015) and pixel-wise mapping with attention models Xu et al. (2015); Lu et al. (2017). Another example of a visually grounded language model is a model that is used to demonstrate the compositionality of colour descriptions in Monroe et al. (2016) where linguistic descriptions are associated with areas of the colour space. Similar to McMahan and Stone (2015), each observed instance is a colour term paired with a colour code but instead of considering each description as a lexical entry, phrases are captured by a grounded language model as in Equation 11.1. The qualitative human evaluation of how newly composed colour words by this model refer to the colour space suggest that language models can capture compositionality through gradient learning used with neural networks.

In this paper, we follow up and extend the work of Monroe et al. (2016). We focus on recurrent neural language models of sequences of words conditioned by encoded locations that these words refer to in visual scenes. Hence, we are interested in grounded semantic composition that is not only captured by probabilistic models of words given their context words, but also by models of their relatedness to perceptual representations. An important and novel question we investigate is *what these models are learning*: to what degree the representations of meaning (both collocational from vector spaces and grounded in perception) are interpretable and therefore *compositional* in the sense of Montague (1974). We focus on one domain of grounded meaning: spatial descriptions of various length and their grounding in spatial templates of Logan and Sadler (1996). In particular we try to answer the following questions: *(1)* To what extent are the language models that have been learned grounded in spatial representations? *(2)* Is it possible to generate new, previously unseen grounded composed spatial descriptions from observing their words only in other grounded composed phrases?

This paper is organised as follows. In Section 11.2 we describe the creation of an artificial dataset of composed spatial templates and the associated descriptions based on the experimental work of Logan and Sadler (1996). In Section 11.3 we describe our neural network model which we use for training our grounded language model. Section 11.4 describes an evaluation of the learned representations compared to the original representations the system was learning from. Finally, Section 11.5 points to conclusions and further work. The code and results are available at `https://github.com/GU-CLASP/spatial-composition`.

11.2 The dataset

In order to train a grounded language model we require samples of language use paired with locations they are referring to. Considering the rationality of speakers and their observers Grice (1975), the frequency of each co-occurring utterance–location corresponds to the appropriateness of such utterance as a description of that location. One complication of judging the appropriateness of spatial terms this way is that they are not only depended on the location they describe but also on other properties of the situation such as the agreed frame of reference, object shape, and the function of the

landmark and the target objects involved, etc. Herskovits (1986); Dobnik and Cooper (2017). However, these properties will not be considered in the present study.

Logan and Sadler (1996) performed several psychological experiments related to the geometric apprehension of spatial relations. For example, they collected acceptability ratings (1–9) for a set of spatial relations per different locations of the target object in a 7×7 grid relative to the landmark object in the centre $(3, 3)$. The acceptability scores were collected from 32 informants through random presentation and then averaged per location. The matrix of average acceptability scores per description is called a *spatial template* and represents the appropriateness of each location in the process of interpreting that spatial relation Logan and Sadler (1996). They collect spatial templates for the following spatial relations: *right_of*, *left_of*, *below*, *under*, *over*, *above*, *near_to*, *next_to*, *far_from*, and *away_from* which we also apply in our work. Furthermore, in order to be able to explore the limits of the language models for learning compositions, we extend this vocabulary with a few additional words. We describe how we used them to synthesise the composed spatial templates for our training data in the following section.

11.2.1 Spatial templates as probabilities

As stated earlier, the spatial templates of Logan and Sadler (1996) give us the average acceptability scores on the scale 1–9 for each of $7 \times 7 - 1$ locations. In the process of grounding a description ($w_{1:T} = w_1 w_2 \dots w_T$), a vector of scores representing its spatial template is used to rank the description's acceptability across all possible locations:

$$T_{w_{1:T}} = \{Score(w_{1:T}, l)\}_{l \in L} \tag{11.2}$$

Our goal is to find such representation for any composed phrase $w_{1:T}$. We introduce the following assumption to convert the acceptability scores to probabilities. The acceptability scores are an indicator of a degree of belief Ramsey (1931) that a rational speaker would use a particular description ($w_{1:T}$) to describe the landmark object at a certain location ($c \in L$). We therefore expect:

$$Score(w_{1:T}, c) \propto Pr(w_{1:T}, c) \tag{11.3}$$

where the $Pr(w_{1:T}, c)$ is the probability of observing a co-occurrence of a phrase $w_{1:T}$ and a location c. In order to be able to compare spatial templates generated by the learned neural language models and the original acceptability scores which were used to generate the training data, we assume that all locations are equally accessible, then:

$$\begin{aligned} & Pr(w_{1:T}, c) = Pr(w_{1:T}|c)Pr(c) \\ \implies & Score_(w_{1:T}, c) \propto Pr(w_{1:T}|c) \end{aligned} \tag{11.4}$$

We compare the generated probability scores by our neural language model, a vector of probabilities over all locations, for a particular description with its expected spatial template. We use a correlation coefficient to quantify the difference between a predicted and the "real" spatial template. A spatial template gives us information about the applicability of each location. When choosing a location given a description we would consider the ranking of locations by their applicability score. Hence, since we are not interested in the actual scores but their ranking, Spearman's rank correlation coefficient is a suitable measure for comparing spatial templates.

$$\begin{aligned} T_{w_{1:T}} &= \{Score_{w_{1:T},l}\}_{l \in L} \\ \hat{T}_{w_{1:T}} &= \{Pr(w_{1:T}|c)\}_{l \in L} \\ \rho(T_{w_{1:T}}, \hat{T}_{w_{1:T}}) & \quad \text{Spearman's rank correlation coefficient} \end{aligned} \tag{11.5}$$

11.2.2 Synthesised data

Considering the assumptions from the previous section, using a simple min-max normalisation, the list of scores in a spatial template can be translated to a Bernoulli probability of events:

$$Pr(w_{0:T}, c) \approx s_{w_{1:T},c} = \frac{Score(w_{1:T}, c) - 1}{9 - 1} \tag{11.6}$$

Using these probabilities, we synthesise instance events of locations and descriptions that make our training dataset using the same method as Coventry et al. (2004). Having normalised acceptability ratings as probabilities, we can generate samples with a frequency corresponding to these probabilities.

$$freq(w_{0:T}, c) = n \times Pr(w_{0:T}, c) \tag{11.7}$$

For example, by choosing $n = 5$, for a location with normalised scores 0.58 for $right\_of$, 0.15 for $left\_of$ and 0.91 for $next\_to$, we generate 2, 0, 4 instances for each respective description.

Logan and Sadler (1996) present acceptability scores for spatial descriptions obtained experimentally only for single-word spatial descriptions such as $left$ and $above$. However, in our task we need their composed representations. We take the assumption that all spatial templates compose with some known function. For example for two spatial descriptions conjoined with an intersective and "{spatial_term1} and {spatial_term2}", Gapp (1994) discusses (but not experimentally evaluates) five compositional functions for grounding spatial templates. More recently, Dobnik and Åstbom (2017) show that taking a *geometric mean* over acceptability scores per location give highly correlated compositions with spatial templates of composed descriptions obtained experimentally. Another study on representing binary beliefs with beta distributions Jøsang and McAnally (2005), shows that the product of scores has the best approximation for conjoined opinions. We also take this as our compositional function to generate spatial templates for composite descriptions as in Figure 11.1, here further defined as:

$$\begin{aligned} g_{\wedge} &: (v_i, v_j) \rightarrow [v_i, \text{“and”}, v_j] \\ \hat{s}_{g_{\wedge}(v_i,v_j),c} &= s_{v_i,c} \times s_{v_j,c} \end{aligned} \tag{11.8}$$

Where $g_{\wedge}$ is a grammar rule for conjoined composition. Similarly, following Jøsang and McAnally (2005), logical OR-composition can be defined with co-multiplication:

$$\begin{aligned} g_{\vee} &: (v_i, v_j) \rightarrow [\text{“}either\text{”}, v_i, \text{“or”}, v_j] \\ \hat{s}_{g_{\vee}(v_i,v_j),c} &= s_{v_i,c} + s_{v_j,c} - s_{v_i,c} \times s_{v_j,c} \end{aligned} \tag{11.9}$$

For negation "not {spatial_term}" we take a complement of the acceptability scores as shown in Figure 11.1.

$$\begin{aligned} g_{\neg} &: v \rightarrow [\text{“not”}, v] \\ \hat{s}_{g_{\neg}(v),c} &= 1 - s_{v,c} \end{aligned} \tag{11.10}$$

The resulting compositions are shown in Figure 11.1. One might object to the usage of such synthetic data. It is important to note that the prime goal of this work is not to learn grounded models of spatial language that would best approximate human intuitions but to test to what degree grounded

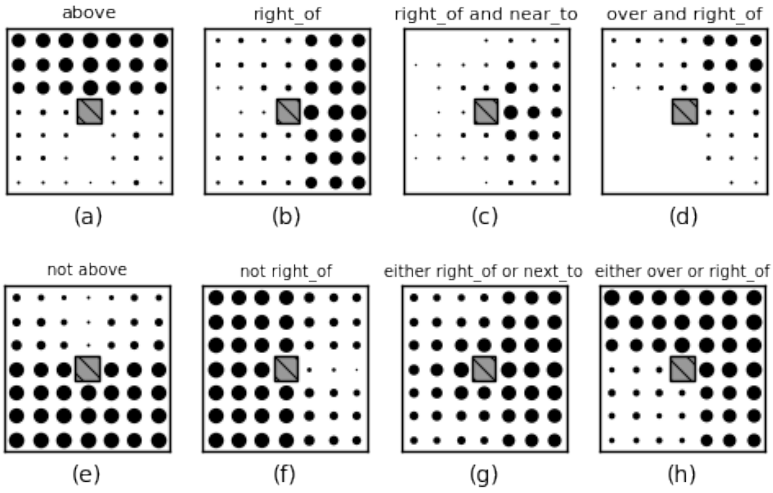


Figure. 11.1: Spatial templates in a 7×7 grid: (a) and (b) are spatial templates for "above" and "right" from Logan and Sadler (1996) collected from human judgements. (c-h) are their synthetic compositions. (c) and (d) are intersective-AND compositions of two spatial templates using point-wise multiplication. (e) and (f) represent the negation of (a) and (b) using a complement operation. (g) and (h) are logical-OR compositions of two spatial templates using a point-wise co-multiplication.

neural language models are capable of capturing grounded compositionality expressed as compositional functions of various complexities which have been confirmed in the previous literature to work well. Hence, we are interested in testing to what extent new machine learning models are capable of learning these functions.

We create two datasets. In the first dataset all descriptions are grounded in spatial templates as described above. In the second dataset additional words were added which we assume have no grounding in perception to test if the neural language model is able to distinguish them from the words sensitive to grounding. For example: "{the object | it | the ball} is {spatial_phrase}

{the object | it | the box}". The following additional grammar rules were applied during the generation of the second dataset:

$$
\begin{array}{lcl}
g_1 & : & (v*) \rightarrow [v*] \\
g_2 & : & (v*) \rightarrow [\text{“}it\text{”}, \text{“}is\text{”}, v*] \\
g_3 & : & (v*) \rightarrow [\text{“}it\text{”}, \text{“}is\text{”}, v*, \text{“}the''\text{”}, \text{“}box''\text{”}] \\
g_4 & : & (v*) \rightarrow [\text{“}the\text{”}, \text{“}ball\text{”}, \text{“}is\text{”}, v*, \text{“}the''\text{”}, \text{“}box''\text{”}] \\
g_5 & : & (v*) \rightarrow [\text{“}the\text{”}, \text{“}object\text{”}, \text{“}is\text{”}, v*, \text{“}the''\text{”}, \text{“}box''\text{”}]
\end{array}
\tag{11.11}
$$

In the generated descriptions, words such as *and, not, the, box, ball, it, object,*

Algorithm 1 Synthetic generator

$n = 5$
$g_{compositional} = \{g_1, g_\neg, g_\wedge, g_\vee\}$
$g_{textual} = \{g_1, g_2, g_3, g_4, g_5\}$
procedure SYNTHETICGENERATOR($v*, c, g$)
    $freq \leftarrow n \times \hat{s}_{g(v*),c}$
    for 1 **to** $freq$ **do**
        $syntax \leftarrow$ choose_random($g_{textual}$)
        $text \leftarrow syntax(g(v*))$
        Generate($text, c$)

and *is* are not grounded in locations individually but the phrases they occur in refer to locations on the map.

11.3 Neural network architecture

We use the Recurrent Neural Network (RNN) architecture for a language model Graves (2013) with Long-Short Term Memory (LSTM) Hochreiter and Schmidhuber (1997) and a decoder architecture from Cho et al. (2014) which concatenates word-embeddings of each input word with an encoded location:

$$
\begin{array}{lcl}
\mathbf{y}_t & = & Pr(w_t | w_{1:t-1}, c) \\
\mathbf{h}_t & = & f_\theta(\mathbf{e}_{w_{t-1}}; \mathbf{c}, \mathbf{h}_{t-1}) \\
\hat{\mathbf{y}}_t & = & \text{softmax}(\mathbf{W}\mathbf{h}_t + \mathbf{b})
\end{array}
\tag{11.12}
$$

where $\hat{\mathbf{y}}_t$ is the expected categorical probability at time t, f is a recurrent cell with parameters θ, $\mathbf{e}_w$ is an embedding vector for a word w, and $\mathbf{c}$ is an encoded location as a one-hot vector as shown in Figure 11.2.

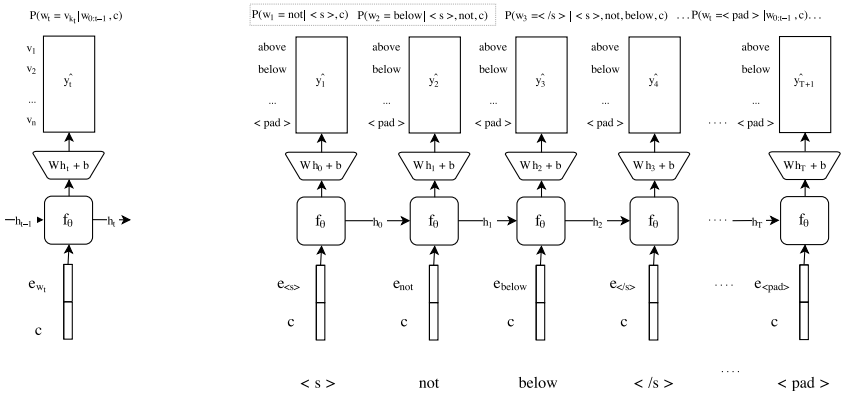


Figure. 11.2: The diagram on the left illustrates the architecture of the model at word/time-step t using a vocabulary size n. On the right, there is an unfolded example how a phrase like "not below" is paired with a location c as in $(w_{1:T}, c)$ and fed as input to the LSTM decoder. In this setup, similar to Graves (2013), we train the model to predict the next word in a sequence and the chain of output probabilities is taken to estimate the final probability. The sequence can be cut before reaching the end tag $< /s >$.

The training set in a batch are pairs of word sequences and their corresponding location codes: $\{(w^{(i)}{}_{1:T}, c^{(i)})\}_{i \in D}$ where D is our training dataset. The loss function used is the cross entropy distance between predicted distribution and targeted distribution or *log-loss*. The observed true output $\mathbf{y}_t^{(i)}$ is represented with one-hot encodings. The training process can be summarised as follows:

$$\begin{aligned} \mathbf{y}_t^{(i)} &= \delta_{w_t^{(i)}} \\ L^{(i)}(\Theta) &= -\sum_{t=1}^{T+1} \mathbf{y}^{\mathrm{T}(i)}_t \log(\hat{\mathbf{y}}_t^{(i)}) \\ &= -\sum_{t=1}^{T+1} \log(\hat{\mathbf{y}}_t^{(i)}(w_t^{(i)})) \end{aligned} \tag{11.13}$$

We train the network parameters with *Adam stochastic gradient descent* Kingma and Ba (2014) with batch normalisation implemented as an optimiser in Keras Chollet (2015). On each mini-batch update as ($\Theta_b \leftarrow$

$\Theta_{b-1} + AdamSGD(\nabla_\Theta \mathcal{L})$) the following parameters of the model (Θ) are updated:

$\{\mathbf{e}_w\}_{w \in V}$	Embedding vectors for all words
θ	Parameters of the RNN cell, composed feature vectors
$\mathbf{W}, \mathbf{b}$	Parameters of the final dense layer

(11.14)

11.3.1 Implementation

We implemented our model in Keras Chollet (2015) with TensorFlow Abadi et al. (2015) as a back-end. All parameters were initialised randomly with Keras recommendations. In the current implementation, the size of the $\mathbf{h}_t$, the hidden unit of LSTM, is 15, and the parameters of the RNN cell have a dropout of 0.1. The dropout on embeddings is set to 0.3.

We left-padded descriptions $w_{1:T'}$ with a starting token $w_0 =< s >$ and right-padded them with a finishing token $w_{T'+1} =< /s >$ while the rest was padded with $< pad >$ up to the maximum description length of $T+1$ as illustrated in Figure 11.2. The final y_{T+1} can be either $< pad >$ or $< /s >$. The length of the RNN chain has to be of the fixed size $T+1$, the length of the longest possible sentence, in order to be used with Keras and its implementation on graphic cards.

During each experiment, we trained the model until it reached an over-fitting point with equal training and validation loss.

11.3.2 From the outputs of the RNN to probabilities of composed descriptions

The decoder architecture of RNNs is normally used as a generator which produces sequences of words or characters from an encoded sequence, e.g. Cho et al. (2014); Graves (2013). This can be achieved by applying Equation 11.1. The decoder predicts the most likely next word in a chain of softmax productions $\hat{\mathbf{y}}_t$. The unfolded RNN in Figure 11.2 shows how for a sequence of words as input vectors, $\hat{\mathbf{y}}_t$ are predicted which represent categorical probabilities for all possible following words at a time step t.

For a given sequence, $w_{1:T} = v_{k_1:k_t}$, we estimate the probabilities using Equation 11.1 as follows:

$$\begin{aligned} Pr(w_t = v_{k_t} | w_{1:t-1} = v_{k_1:k_t}, c) &= \hat{\mathbf{y}}_t(v_{k_t}) \\ Pr(w_{1:T} = v_{k_1:k_T} | c) &= \prod_{t=1}^{T} \hat{\mathbf{y}}_t(v_{k_t}) \end{aligned} \tag{11.15}$$

The estimated probability is then used to generate spatial templates as in Equation 11.5. The probabilities over all possible locations on the map L for a given composition of words can be aggregated as follows:

$$\hat{T}_{v_{k_1:k_{T'}}} = \{Pr(w_{1:T'} = v_{k_1:k_{T'}} | c)\}_{c \in L} \tag{11.16}$$

11.4 Evaluation

We evaluate the learning of composed grounded phrases by examining to what degree the spatial templates produced by the learned model correspond to the original spatial templates that were used in generating the training data, how successful is the learning with different kinds of compositions, and what is the effect of adding distractor words. We ran two experiments, (1) on a simple synthetic dataset containing short phrases where all words are grounded in locations, and (2) on a synthetic dataset generated with five additional grammar rules from Equation 11.11, introducing words without spatial grounding or distractor words. We test the learning of compositional phrases by training a language model on phrases produced by individual composition types as well as all composition types in both synthetic datasets. A comparison of the predicted spatial templates with the original spatial templates with Spearman's rank correlation coefficient (Equation 11.5) in Table 11.1 shows that there is high correlation between them. We report the average Spearman's ρ and their median p-values for statistical significance.

For both Experiment 1 and 2 we created two variations: (1) learning of novel grounded compositions, where different proportions of AND-phrases and OR-phrases are omitted from the dataset and therefore hidden from the learner; (2) learning of novel single words from grounded compositions, where proportions of single-word instances are omitted from the dataset and their representations can only be learned from their occurrence in composed phrases with other words.

	Simple phrases	With distractors	Untrained
AND-phrases	0.87	0.85	-0.00
NEG-phrases	0.72	0.82	0.03
OR-phrases	0.79	0.80	-0.03
SINGLE-word	0.92	0.91	-0.05
All previous	0.83	0.83	-0.01
All previous + distractors	NaN	0.84	-0.03

Table. 11.1: For each type of compositional phrases we calculate the average Spearman's rank correlation coefficient (ρ) between the predicted spatial templates and the templates used to generate the training data. The median p-value of ρ of all trained models is < 0.001. The column *Untrained* indicates the performance of the model with a random initialisation of weights.

In all experiments we hold out 10% of the dataset for validation. In Experiment 1 we iterated the training over 64 epochs using a batch size 8. In Experiment 2, using a batch size 256, we stopped learning iterations before 1024 epochs if the validation loss became equal to the training loss.

11.4.1 Experiment 1: Learning composition of short phrases

In this experiment the training data is generated for single spatial words, AND-compositions, OR-compositions, and negated phrases without additional distractor words save "and", "either", "or", and "not".

Learning of novel grounded compositions

The training data contains synthesised samples of all single words and their negations. However, different proportions of AND-phrases and OR-phrases are removed from the training set to test if the model can learn unseen composed phrases. Table 11.2 shows the average of Spearman's ρ correlation coefficient for different portions of held-out phrases. Figure 11.3 illustrates some predicted novel grounded compositions where 50% of complex phrases were held out. The ρ scores lower than 0.6 may not be trustworthy, e.g. "above and left_of" with $\rho = 0.5$ in Figure 11.3.

Proportions of 90 combinations	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
AND-phrases	0.84	0.8	0.78	0.76	0.71	0.67	0.64	0.53	0.45	0.29
OR-phrases	0.74	0.73	0.69	0.67	0.56	0.57	0.54	0.38	0.23	-0.23

Table. 11.2: Spearman's ρ for held-out proportions of phrases. up to 80% have a median p-value < 0.001 and p-value > 0.05 for higher proportions.

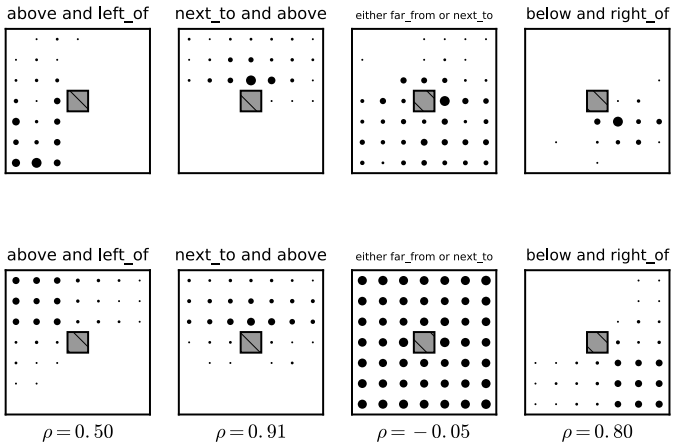


Figure. 11.3: The predicted spatial templates are shown on the top and the original spatial templates in the bottom.

The results indicate that the model can produce spatial templates for novel compositions. However, the learning of composed phrases is dependent on the size and the variety of training instances. Some phrases are more difficult to train than others. For example, OR phrases correspond to regions that are more spread out across the 48 locations which makes them more difficult to learn, e.g. an extreme case such as "either far from or next to".

Learning of novel single words from grounded compositions

In this experiment we omit identical proportions of all description types, thus also single word descriptions and negated descriptions. In this case, the predicted novel spatial templates are learned solely based on observing these words in combination with other words. As before, we conduct the test with different sizes of held-out data. The results are shown in Table 11.3. When omitting up to 4 single descriptions (*right_of*, *over*, *far_from* and *under*) the average ρ on grounded SINGLE-word descriptions decreases only by 0.05 (from 0.92, Table 11.1). This means that their grounding is successfully learned from grounded composed expressions. Figure 11.4 shows a novel learned spatial template for "above".

	10%	20%	30%	40%
AND-phrases	0.86	0.8	0.77	0.81
NEG-phrases	0.83	0.64	0.59	0.43
OR-phrases	0.73	0.78	0.68	0.69
SINGLE-word	0.9	0.9	0.84	0.87

Table. 11.3: The average Spearman's ρ for different proportions of unseen examples.

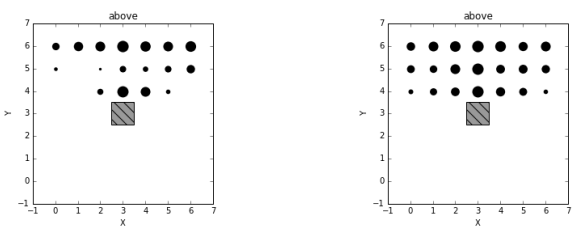


Figure. 11.4: The predicted and the original spatial template.

Qualitative observations

A qualitative examination of the predicted spatial templates shows that spatial templates with the lowest ρ are those with no points in space ("right_of and left_of") or those with a uniform spread of points across space ("either far_from or next_to") which in our scenario includes a number of training instances as rules from Section 11.2.2 were applied to all combinations of spatial templates. We get the highest ρ with compositions such as "over and above", possibly because the two spatial templates overlap and result in a simplified composed representation.

11.4.2 Experiment 2: Adding distractor words with no spatial grounding

In Experiment 2 we train and measure the performance of the model on grounded descriptions which also include non-grounded distractor words, for example: "the ball is not left_of the box" or "it is above and right_of the object". The words such as "ball", "object", "box", "it" and "is" provide no contribution to the grounded meaning (location). In this dataset the number of possible composed phrases increases from 200 to 1,000. Algorithm 1 in Section 11.2.2 ensures that in the 1,000 possible phrases the same number of instances is generated as before, now per each of the five permutation rules introducing distractors. The held-out proportions of spatial descriptions are created before Algorithm 1 is applied so permutations including these are not generated.

Learning of novel grounded compositions

Although now the training data includes longer sequences and several distractors which make these compositions harder to learn, the results are only slightly weaker than in Experiment 1 as shown by a comparison of Table 11.4 with Table 11.2.

Learning of novel single words from grounded compositions

The results of this task on the dataset from Experiment 2 are shown in Table 11.5. The ρ are nearly identical or only slightly lower for SINGLE-words

Proportions of 90 combinations	10%	20%	30%	40%	50%	60%	70%	80%
AND-phrases	0.82	0.79	0.75	0.78	0.73	0.69	0.66	0.45
OR-phrases	0.78	0.69	0.67	0.66	0.59	0.59	0.44	0.33

Table. 11.4: The average Spearman's ρ with the median p-value of < 0.001. After 80% of held-out phrase types the ρ values are not statistically significant.

	10%	20%	30%	40%
AND-phrases	0.82	0.60	0.71	0.81
NEG-phrases	0.75	0.66	0.45	0.30
OR-phrases	0.76	0.76	0.71	0.64
SINGLE-word	0.88	0.43	0.73	0.84

Table. 11.5: The average Spearman's correlations decomposition task Experiment 2.

compared to Experiment 1 (Table 11.3). There is an unusual drop in ρ at 20% of held-out descriptions which requires further investigation. Overall, we can conclude that the system successfully learned omitted single words from their grounded compositions even with distractor words.

11.4.3 Experiment 3: How much grounding?

In Experiment 3, we examine how the amount of training corresponds to the groundedness of expressions in spatial templates. In particular, we examine the learning curve across several epochs at which more of the same data is presented incrementally to the learner and how well does the currently

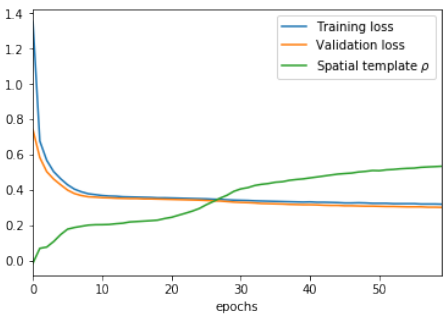


Figure. 11.5: The learning curve for Experiment 3.

learned model corresponds to the target spatial templates. Typically, the performance of the learner at each epoch is estimated by a loss function, here the cross-entropy (log-loss). We compare the loss at each epoch with the average Spearman's ρ between the predicted templates and the original templates for 110 possible combinations of descriptions from Experiment 1 (excluding OR-phrases). Here, we only run the experiment with 20% omission of the dataset. Figure 11.5 shows how average ρ corresponds to the learning progress. The figure shows that even after the training and the validation loss are only slightly decreasing between epochs the groundedness is increasing at a higher rate. This can be explained by the fact that the network is not only predicting locations but also sequences of descriptions which adds a further complexity to learning which is reflected in the loss.

11.5 Conclusion and future work

We have presented a grounded language model with recurrent deep neural networks. The objective of our task was to examine to what extent our neural network architecture can learn a grounded language model that generated the training data and whether a word that is grounded as a part of a phrase can "carry over" its grounding to another phrase not observed in the training data. In our view this is the ultimate test that grounding is compositional. We conduct two learning experiments. In the first experiment we learn a grounded language model where all descriptions in a sequence are grounded. In the subsequent sub-experiments we test the success of the grounded language models where some word compositions are omitted from training. We show that the model is capable of grounding novel compositions and also predicting grounding of single words while only learning from compositions. However, the degree of success, while on overall high, is dependent on the amount of the absent information and the coverage of the training instances. In the second experiment, we add words to our grounded language model that have no grounding and test whether the system is able to learn different grounding sensitivities of different words. We show that our language model is capable of recognising the contribution of each constituent to the meaning of the entire grounded composition. Finally, in the third experiment we examine grounding related to the log-loss success rate of learning. Overall, we conclude that our deep neural architecture successfully learns grounded

spatial descriptions in a way that the learned functions are similar to the ones that generated the data. This is a useful result which points towards the fact that language is compositional both at the level of word sequences and the portions of scenes that they refer to, thus confirming the result in Dobnik and Åstbom (2017). In the future work we will focus on the effects of the varying dataset sizes on the rate of learning and test the learning setup on more complex perceptual representations (in terms of the expected irregularities) such as images.

Bibliography

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A neural probabilistic language model. *journal of machine learning research*, 3(Feb):1137–1155.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

François Chollet. 2015. Keras. `https://github.com/fchollet/keras`.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Kenny R Coventry, Angelo Cangelosi, Rohanna Rajapakse, Alison Bacon, Stephen Newstead, Dan Joyce, and Lynn V Richards. 2004. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In *International Conference on Spatial Cognition*, pages 98–110. Springer.

Simon Dobnik. 2009. *Teaching mobile robots to use spatial words*. Ph.D. thesis, University of Oxford: Faculty of Linguistics, Philology and Phonetics and The Queen's College, Oxford, United Kingdom.

Simon Dobnik and Amelie Åstbom. 2017. (Perceptual) grounding as interaction. In *Proceedings of Saardial – Semdial 2017: The 21st Workshop on the Semantics and Pragmatics of Dialogue*, pages 17–26, Saarbrücken, Germany.

Simon Dobnik and Robin Cooper. 2017. Interfacing language, spatial perception and cognition in Type Theory with Records. *Accepted for Journal of Language Modelling*, n(n):1–30.

Klaus-Peter Gapp. 1994. A computational model of the basic meanings of graded composite spatial relations in 3D space. In *Advanced geographic data modelling. Spatial data modelling and query languages for 2D and 3D applications (Proceedings of the AGDM'94)*, Publications on Geodesy 40, pages 66–79. Netherlands Geodetic Commission.

Alex Graves. 2013. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Herbert Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 41–58. Academic Press.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.

Annette Herskovits. 1986. *Language and spatial cognition: an interdisciplinary study of the prepositions in English*. Cambridge University Press, Cambridge.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Audun Jøsang and David McAnally. 2005. Multiplication and comultiplication of beliefs. *International Journal of Approximate Reasoning*, 38(1):19–51.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 6.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. *arXiv preprint arXiv:1206.6423*.

Brian McMahan and Matthew Stone. 2015. A bayesian model of grounded color semantics. *Transactions of the Association for Computational Linguistics*, 3:103–115.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, page 3.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning*, pages 641–648. ACM.

Will Monroe, Noah D Goodman, and Christopher Potts. 2016. Learning to generate compositional color descriptions. *arXiv preprint arXiv:1606.03821*.

Richard Montague. 1974. *Formal Philosophy: Selected Papers of Richard Montague*. Yale University Press, New Haven.

Frank P Ramsey. 1931. Truth and probability (1926). *The foundations of mathematics and other logical essays*, pages 156–198.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205.

Deb Roy and Niloy Mukherjee. 2005. Towards situated speech understanding: Visual context priming of language models. *Computer Speech & Language*, 19(2):227–248.

Richard Socher, Andrej Karpathy, Quoc V Le, Christopher D Manning, and Andrew Y Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81.

12 Metaphoricity Of Compositions With Distributional Representations

Yuri Bizzoni, Stergios Chatzikyriakidis and Mehdi Ghanimifard. "Deep" Learning: Detecting Metaphoricity in Adjective-Noun Pairs. *In Proceedings of the Workshop on Stylistic Variation, pp. 43-52. 2017.*

Abstract Metaphor is one of the most studied and widespread figures of speech and an essential element of individual style. In this paper we look at metaphor identification in Adjective-Noun pairs. We show that using a single neural network combined with pre-trained vector embeddings can outperform the state of the art in terms of accuracy. In specific, the approach presented in this paper is based on two ideas: a) transfer learning via using pre-trained vectors representing adjective noun pairs, and b) a neural network as a model of composition that predicts a metaphoricity score as output. We present several different architectures for our system and evaluate their performances. Variations on dataset size and on the kinds of embeddings are also investigated. We show considerable improvement over the previous approaches both in terms of accuracy and w.r.t the size of annotated training data.

12.1 Introduction

The importance of metaphor to characterize both individual and genre-related style has been underlined in several works (Leech and Short, 2007; Simpson, 2004; Goodman, 1975). Studying the kinds of metaphors used in a text can contribute to differentiate between poetic and prosaic style,

etc. In literary studies, metaphor analysis is often undertaken on a stylistic perspective: "after all, metaphor in literature is a stylistic device and its forms, meanings and use all fall within the remit of stylistics" Steen (2014). Metaphor is thus often taken into consideration qualitative stylistic analyses (Fahnestock, 2009). Nonetheless, it is still very difficult to take metaphors into account in computational stylistics due to the complexity of automatic metaphor identification (Neuman et al., 2013; Beigman Klebanov et al., 2015), which is the task of identifying metaphorical usages of text, sentences or subsentential fragments.

This paper's focus of interest is the automatic detection of adjective-noun (AN) pairs like the following:

- Clean floor / clean performance
- Bright painting / bright idea
- Heavy table / heavy feeling

The above examples illustrate that adjectives "normally" used to describe physical characteristics, e.g. a feature that can be perceived through senses like size or weight, are reused to describe more abstract properties. Thus, both a painting and an idea can be bright, both a table and a feeling can be heavy. We will not provide a mean to retrieve AN metaphors in unconstrained texts (e.g. we won't focus on segmentation) but we will study ways to detect metaphoricity in given pairs. Theoretical work on metaphor in the linguistics literature goes back a long way and spans different theoretical paradigms. One of the earliest and most influential works is Conceptual Metaphor Theory (CMT) Lakoff and Johnson (2008) (originally published in 1981) and subsequently elaborated in a couple of papers Lakoff (1989, 1993). According to CMT, metaphors in natural language can be seen as instances of conceptual metaphors. A conceptual metaphor roughly corresponds to understanding a concept or an idea via association or relation with another idea or concept. Other influential linguistic approaches to metaphor include pragmatic approaches cast within frameworks like relevance theory Romero and Soria (2014); Wilson (2011), and also approaches where some sort of formal semantics is used Vogel (2001). The common denominator in all these approaches is the recognition that there is systematicity in the way metaphorical meanings arise and also that the process of metaphor construction is extremely productive. Thus, given these properties, one would expect metaphors to be quite common in Natural Language

(NL). Evidence from corpus linguistics seems to support this claim Cameron (2003).

Metaphor detection in statistical NLP has been attempted through several different frames, such as topic modeling Li and Sporleder (2010b), semantic similarity graphs (Li and Sporleder, 2010a), distributional clustering (Shutova et al., 2010), vector space based learning Gutiérrez et al. (2016) and, most of all, feature-based classifiers Tsvetkov et al. (2014). In the latter case, the challenge consists in selecting the right features to annotate the training data with, and to review their "importance" or weight based on machine learning results.

In this paper we show how using a single-layered neural network combined with pre-trained distributional embeddings can outperform the state of the art in an AN metaphor detection task.

More specifically, this paper's contributions are the following:

- We introduce a system to predict AN metaphoricity and test it on the corpus introduced by Gutiérrez et al. (2016), showing a significant improvement in accuracy.
- We explore different variations of this model based on ideas found in the literature for composing distributional meaning and we evaluate them under different constraints.

The paper is structured as follows: in Section 2 we present the background on AN metaphor detection and we detail the dataset we use to train our model. In Section 3 we describe our approach, giving a general overview and further describing three alternative architectures on the same model. In Section 4 we present several evaluations of our model. Table 12.1 and Table 12.2 synthesize some of our findings. In Section 5 we discuss our findings and possible future applications of the work described in this paper.

12.2 Background

In the specific task of detecting metaphoricity for AN pairs we find four relevant works that seem to represent the main stages in figurative language detection until now.

	Accuracy	Feature engineering	Annotated dataset	Embedding
Turney et al. (2011)	0.79	Yes	100	LSA
Tsvetkov et al. (2014)	0.85	Yes	200	-
Gutiérrez et al. (2016)	0.81	No	8592	DSM
Our model	**0.91**	No	8592	Word2Vec

Table. 12.1: The reported accuracy from previous words on AN metaphor detection. The first two studies used different datasets. We are using larger pre-trained vectors than Gutiérrez et al. (2016); at the same time, we don't need a parsed corpus to build our vectors and we don't use adjectival matrices. Given these differences, this comparison should not be considered a "competition".

The oldest work of the series, Krishnakumaran and Zhu (2007), strongly relies on external resources. They adopt a WordNet based approach to recognize Noun-Noun (NN), Noun-Verb (NV) and AN metaphors. Their work is mainly based on qualitative analyses of specific examples and shows that, while they can be useful in such a task, hyponym/hypernym relations are not enough to distinguish metaphors from literal expressions.

More recently, Turney et al. (2011) adopt a two-stage machine learning approach. They first try to learn the words' degree of concreteness and then use this knowledge to detect whether an AN couple is metaphorical or not. They measure their performance on 100 phrases involving 5 adjectives and reach an accuracy of 0.79. It is worth noting that this choice is not random: the authors select the abstract/concrete polarity based on psycholinguistic findings that seem to validate the hypothesis that some kinds of metaphorical expressions are processed as abstract elements.[1]

These results were outperformed by Tsvetkov et al. (2014) through a random forest classifier using DSM vectors, WordNet senses and several accurately selected features, such as abstractness. They also introduce a new set of 200 phrases, on which they declare an F-score of 0.85.

[1] For a more recent study on this issue see Forgács et al. (2015).

	Random W	**Trained** W
cat-linear	0.8973	0.9153
cat-relu	0.8763	0.9228
sum-linear	0.8815	0.9068
sum-relu	0.8597	0.9150
mul-linear	0.7858	0.8066
mul-relu	0.7795	0.8186

Table. 12.2: The accuracy results after training the model based on each architecture. In all setups, we trained on 500 samples in 20 epochs. Using a random W is equivalent to preventing our network from learning any form of compositionality (we could consider it as a baseline for models with trained W). As we discuss in the paper, the difference in accuracies with the “baseline" (not training W) shows that training W is helpful.

Finally, Gutiérrez et al. (2016) train a distributional model on a corpus of 4.58 billion tokens and test it on an annotated dataset they introduce consisting of 8592 AN phrases. This is the same dataset we are using in this paper and the largest available to date.

They first train distributional vectors for the words in the dataset using positive pointwise mutual information. Then, for each adjective present in the dataset, they divide the literal phrases the adjective occurs in from the metaphorical phrases the same adjective appears in. Then, three different adjective matrices are trained: one to model the adjective’s literal sense, one to model its metaphorical sense, and one trained on all the phrases containing this adjective, both literal and metaphorical. They then develop a system to “decide" whether a particular occurrence of an adjective is more likely to relate to the “literal matrix" or the “metaphorical matrix". It is shown that, although such matrices are trained on relatively few examples, they can reach an accuracy of over 0.78.

12.2.1 Corpus/Experimental Data

The dataset we are using comes from Gutiérrez et al. (2016). [2] It contains 8592 annotated AN pairs, 3991 being literal and 4601 being metaphorical. The dataset focuses on a set of 23 adjectives that: a) can potentially have both metaphorical and literal meanings, and b) are fairly productive.

[2]The dataset is publicly available here: http://bit.ly/1TQ5czN

The choice of adjectives was based on the test set of Tsvetkov et al. (2014) and focuses on 23 adjectives.

In details, all adjectives belong to one of the following categories:

1. temperature adjectives (e.g. cold)
2. light adjectives (e.g. bright)
3. texture adjectives (e.g. rough)
4. substance adjectives (e.g. dense)
5. clarity adjectives (e.g. clean)
6. taste adjectives (e.g. bitter)
7. strength adjectives (e.g. strong)
8. depth adjectives (e.g. deep)

The corpus was carefully built in order to avoid non-ambiguous elements: all the AN phrases present in this dataset were extracted from large corpora and all phrases that seemed to require a larger context for their interpretation were filtered out in order to eliminate potentially ambiguous idiomatic expressions such as *bright side*.

In other terms, the corpus was designed to contain elements whose metaphoricity could be deduced by a human annotator without the need of a larger context.

More details about the construction of the dataset and annotation methodology can be found in Gutiérrez et al. (2016).

12.3 Describing our approach

12.3.1 The model framework

Our objective is to build a classifier that disambiguates between metaphoric and literal AN compositions by providing a probability measure between 0 and 1. We based the framework of the model on the following ideas:

1. Transfer learning: we use pre-trained word-vectors to represent AN pairs as input.

2. A neural network as a model of composition for the AN phrase: our model represents phrases with vectors, then based on this representation predicts a metaphoricity score as output. Although we are going to present several variations of this framework, it's important to remember that the basic model is always a standard NN with a single fully connected hidden layer we will call $\mathbf{p}$.

Our approach is thus based on the idea that well-trained distributional vectors contain more valuable information than their reciprocal similarity and, furthermore, that it is possible to treasure such information through machine learning in different tasks. We use 300-dimensional word vectors trained on different corpora (see Evaluation for more details) . Our approach can be considered as a way of transferring the learned representation from one task to another. Although it is not possible to point out an explicit mapping between the word-vector learning task (e.g. Word2Vec model) and our metaphoricity task, as it is pointed out by Torrey and Shavlik 2009, we use neural networks which automatically learn how to adapt the feature representations between two tasks Bengio et al. (2013). In this way we stretch the original embeddings, trained in order to learn lexical similarity, to identify AN metaphors.

Our neural network, being a parameterized function, follows the generalized architecture of word-vector composition similar to Mitchell and Lapata (2010):

$$\mathbf{p} = f(\mathbf{u}, \mathbf{v}; \theta) \tag{12.1}$$

where $\mathbf{u}$ and $\mathbf{v}$ are two word vector representations to be composed, while $\mathbf{p}$ is the vector representation of their composition with the same dimensions. The function f in our model is parameterized by θ, a list of parameters to be learned as part of our neural network architecture.

Based on the argument by Mitchell and Lapata (2010), parameters such as θ are encoded knowledge required by the compositional process. In our case, the gradient based learning in neural networks will find these parameters as an optimization problem where $\mathbf{p}$ is just an intermediate representation in the pipeline of the neural network, which ends with a prediction of a metaphoricity score.

In other words, in order to predict the degree of metaphoricity, we end up learning a specific semantic space for phrase representations $\mathbf{p}$ and a vector

$\mathbf{q}$ which actually does not represent a phrase itself, but rather the maximal possible level of metaphoricity given our training set.

The degree of metaphoricity of a phrase can thus be directly computed as cosine similarity between this vector and the phrase vector. However, in the network we used a sigmoid function to produce the measure:

$$\hat{y} = \sigma(\mathbf{p} \cdot \mathbf{q} + b_1) = \frac{1}{1 + e^{-\mathbf{p} \cdot \mathbf{q} + b_1}} \tag{12.2}$$

where $\mathbf{q}$ and b_1 are parameters of the final layer and work as metaphoricity indicators, while $\hat{y}$ is the predicted score (*metaphoric* or *literal*) for the composition $\mathbf{p}$. Given a dataset of $D = \{(x_t, y_t)\}_{t \in \{1,\ldots,T\}}$, the composition $\mathbf{p}$ can be formalized as a model for Bernoulli distribution:

$$\begin{array}{rcll} y_t & = & Pr(x_t \text{ being metaphorical} | D) & \in \{0,1\} \\ \hat{y}_t & = & \sigma(\mathbf{p}_t \cdot \mathbf{q} + b_1) & \\ & \approx & Pr(x_t \text{ being metaphorical}) & \in (0,1) \end{array} \tag{12.3}$$

where each x_t is an AN pair in the training dataset labeled with a binary value y_t (0 or 1). Given the labels in D, we interpret y_t as a categorical probability score: the probability of a given phrase being metaphorical. Then, for each pair of words in x_t, we use pre-trained word-vector representations such as $\mathbf{u_t}$ and $\mathbf{v_t}$ in the Equation 12.1 to produce $\mathbf{p}_t$ and, consequently, the score $\hat{y}_t$.

In this formulation, the objective is to minimize the binary cross entropy distance between the estimated $\hat{y}_t$ and the given annotation y_t. Adding $\mathbf{q}$ and b_1 in the list of parameters Θ, we fit all parameters with a small annotated data size T:

$$\begin{array}{rcl} \mathbf{x} & = & (x_1, \ldots x_T) \\ \mathbf{y} & = & (y_1, \ldots y_T) \\ \Theta & = & (\theta, \mathbf{q}, b_1) \end{array} \tag{12.4}$$

$$\begin{array}{rcl} \mathcal{L}(\Theta; \mathbf{x}, \mathbf{y}) & = & -\sum_{t=1}^{T} (y_t \log(\hat{y}_t) + \\ & & (1 - y_t) \log(1 - \hat{y}_t)) \end{array} \tag{12.5}$$

where, on each iteration, we update the parameters in Θ using Adam stochastic gradient descent Kingma and Ba (2014), with a fixed number of iterations over $\mathbf{x}$ and $\mathbf{y}$ to minimize $\mathcal{L}$.

In this paper, we describe three alternative architectures to implement this framework. All three, with small variations, show a robust ability to generalize on the dataset and perform correct predictions.

12.3.2 First Architecture

One possible formulation of this frame is similar to additive composition as described in Mitchell and Lapata (2010), but instead of performing a scalar modification of each vector, a weight matrix modifies all feature dimensions at once:

$$\mathbf{p} = W_{adj}^T \mathbf{u} + W_{noun}^T \mathbf{v} + b \tag{12.6}$$

$$W = \begin{bmatrix} W_{adj} \\ W_{noun} \end{bmatrix} \tag{12.7}$$

where the composition function in equation (12.1) now has $\theta = (W, b)$.

This formulation is very similar to the composition model in Socher et al. (2011) without the syntactic tree parametrization. As such, instead of the non-linearity function we have linear identity:

$$\mathbf{p} = f_\theta(\mathbf{u}, \mathbf{v}) = W^T \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} + b \tag{12.8}$$

In practice, this approach represents a simple merging through concatenation: given two words' vectors, we concatenate them before feeding them to a single-layered, fully connected Neural Network.

As a consequence, the network learns a weight matrix that represents linearly the AN combination. To visualize this concept, we could say that, since our pairs always hold the same internal structure (adjective in first position and noun in second position), the first half of the weight matrix is trained on adjectives and the second half of the weight matrix is trained on nouns.

By using 300 dimension pre-trained word vectors, the parameter space for this composition function will be as following: $W \in \mathbb{R}^{300 \times 600}$ and $b \in \mathbb{R}^{300}$.

12.3.3 Second architecture

The second architecture we describe has the advantage of training a smaller set of parameters with respect to the first. In this model, the weight matrix is shared between the noun and the adjective:

$$\mathbf{p} = f_\theta(\mathbf{u}, \mathbf{v}) = W^T\mathbf{u} + W^T\mathbf{v} + b \tag{12.9}$$

Notice that in the case of comparing the vector representations of two different AN phrases, b will be essentially redundant. An advantage of this model is that the learned composition function f can also map all words' vectors, regardless of the part of speech these words belong to, in the new vector space without losing accuracy in the original task. In this new vector space, a simple addition operator composes two vectors:

$$\mathbf{u}' = W^T\mathbf{u} \tag{12.10}$$

$$\mathbf{v}' = W^T\mathbf{v} \tag{12.11}$$

$$\mathbf{p} = \mathbf{u}' + \mathbf{v}' \tag{12.12}$$

Compared to the first architecture, in this architecture we don't assume the need of distinguishing the weight matrix for the adjectives from the weight matrix for the nouns.

It is rather interesting, then, that this architecture doesn't present significant differences in performance with respect to the first one. The number of parameters, however, is smaller: $W \in \mathbb{R}^{300\times300}$ and $b \in \mathbb{R}^{300}$.

12.3.4 Third Architecture

The third architecture, similarly to the second, features a shared composition matrix of weights between the noun and the adjective, but we perform elementwise multiplication between the two vectors:

$$\mathbf{p} = f_\theta(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \times \mathbf{v})W + b \tag{12.13}$$

The number of parameters in this case is similar to previous architecture: $W \in \mathbb{R}^{300\times300}$ and $b \in \mathbb{R}^{300}$.

12.3.5 Other Architectures

In all three previous architectures we saw that a weight matrix W can be learned as part of the composing function. Throughout our exploration, we found that W can be a random and a constant uniform matrix (not trained in the network) and still being able to learn $\mathbf{q}$ unless we use a non-linear activation functions over the AN compositions.

$$\mathbf{p} = g(f_\theta(\mathbf{u}, \mathbf{v})) \tag{12.14}$$

An intuition is to take W as an identity matrix in Second architecture, the network will take the sum of pre-trained vectors to as features and learn how to predict metaphoricity. A fixed uniform W basically keeps the information in input vectors. For a short overview of all these alternative architectures see Table 12.2.

12.4 Evaluation

Our classifier achieved 91.5% accuracy trained on 500 labeled AN-phrases out of 8592 in the corpus and tested on the rest. Training on 8000 and testing on the rest gave us accuracy of 98.5%.[3]

We tested several combinations of the architectures we described in the paper. For each of the three architectures, we also tested the Rectified linear unit (ReLU) as the non-linearity mentioned in Section 3.5. Our test also shows that a random constant matrix W is enough to train the rest of the parameters (reported in Table 12.2). In general, the best performing combinations involve the use of concatenation (the first architecture), while multiplication led to the lowest results. In any case, all experiments returned accuracies above 75% [4].

To test the robustness of our approach, we have evaluated our model's performance under several constraints:

[3] These results are based on the first architecture, the performance of other architectures are not very different in this simple test. The sample code is available on `https://gu-clasp.github.io/anvec-metaphor/`

[4] The number of parameters in case of using concatenation (as in first architecture) is 180 601 and other compositions, including addition and multiplication, number of parameters is almost the half: 90 601.

- Total separation of vocabulary in train and test sets (Table 12.3) in case of out of vocabulary words.
- Use of different pretrained word embeddings (Figure 12.3).
- Cross validation (Figure 12.1).
- Qualitative selection of the training data based on the semantic categories of adjectives (Figure 12.2).

Finally, we will provide some qualitative insights on how the model works.

Our model is based on the idea of transfer learning: using the learned representation for a new task, in this case the metaphor detection. Our model should generalize very fast with a small set of samples as training data. In order to test this matter, we have to train and test on totally different samples so vocabulary doesn't overlap. The splitting of the 8592 labeled phrases based on vocabulary gives us uneven sizes of training and test phrases[5]. In Table 12.3 using the pretrained Word2Vec embeddings trained on Google News Mikolov et al. (2013) we examined the accuracy, precision and recall of the our trained classifier.

We have used three different word embeddings: Word2Vec embeddings trained on Google News Mikolov et al. (2013), GloVe embeddings Pennington et al. (2014) and Levy-Goldberg embeddings Levy and Goldberg (2014).

These embeddings are not up-dated during the training process. Thus, the classification task is always performed by learning weights for the pre-existing vectors.

The results of our experiment can be seen in Figure 12.3. All these embeddings have returned similar accuracies both when trained on scarce data (100 phrases) and when trained on half of the dataset (4000 phrases).

Training on 100 phrases indicates the ability of our model to learn from scarce data. One way of checking the consistency of our model under data scarcity is to perform *flipped* cross-validation: this is a cross-validation where, instead of training our model on 90% of the data and testing it on the remaining 10%, we flipped the sizes train it on 10% of the data and test it on the remaining 90%. Results for both classic cross-validation and flipped cross-validation can be seen in Figure 12.1. Training on 10% of the data

[5] We chose the vocabulary splitting points for every 10% from 10% to 90%, then we applied the splitting separately on nouns and adjective

proved to consistently achieve accuracies not much lower than 90%. In other terms, a model trained on 90% of the data does not do much better than a model trained on 10%.

Finally, we tried training our model on only one of the semantic categories we introduced at the beginning of the paper and testing it on the rest of the dataset. Results can be seen in Figure 12.2.

We can wonder “why" our system is working: with respect to more traditional machine learning approaches, there is no direct way to evaluate which features mostly contribute to the success of our system. One way to have an idea of what is happening in the model is to use the “metaphoricity vector" we discussed in Section 3. Such vector represents what is learned by our model and can help making it less opaque for us.

If we compute the cosine similarity between all the nouns in our dataset and this learned vector, we can see that nouns tend to polarize on an abstract/concrete axis: abstract nouns tend to be more similar to the learned vector than concrete nouns.

It is likely that our model is learning nouns' level of abstractness as a mean to determine phrase metaphoricity. In Table 4 we show the 10 most similar and the 10 least similar nouns obtained with this approach. As can be seen, a concrete-abstract polarity is apparently learned in training.

This factor was amply noted and even used in some feature-based metaphor classifiers, as we discussed in the beginning: the advantage of using continuous semantic spaces probably relies on the possibility of having a more nuanced and complex polarization of nouns along the concrete/abstract axes than using hand-annotated resources.

12.5 Discussion and future work

In this paper we have presented an approach for detecting metaphoricity in AN pairs that out-performs the state of the art without using human annotated data or external resources beyond pre-trained word embeddings. We treasured the information captured by Word2Vec vectors through a fully connected neural network able to filter out the "noise" of the original semantic space. We have presented a series of alternative variations of this approach

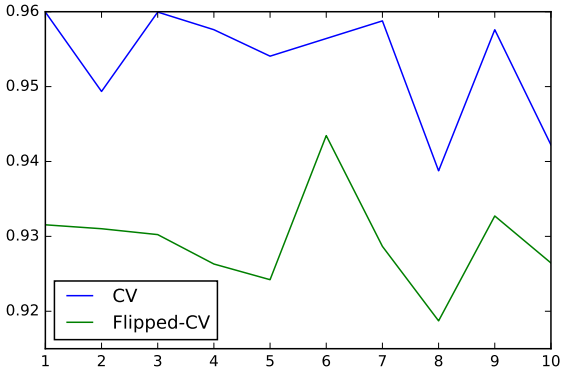


Figure. 12.1: Accuracies for each fold over two complementary approaches: cross-validation (CV) and *flipped* cross-validation ("flipped-CV"). *Flipped* cross-validation takes 90% of our dataset for training. The graph shows that both methods yield good results: in other words training on just 10% of the dataset yields results that are just few points lower than normal cross-validation.

and evaluated its performance under several conditions - different word embeddings, different training data and different training sizes - showing that our model can generalize efficiently and obtain solid results over scarce training data. We think that this is one of the central findings in this paper, since many semantic phenomena similar to metaphor (for example other figures of speech) are under-represented in current NLP resources and their study through supervised classifiers would require systems able to work on small datasets.

The possibility of detecting metaphors and assigning a degree of "metaphoricity" to a snippet of text is essential to automatic stylistic programs designed to go beyond "shallow features" such as sentence length, functional word counting etc. While such metrics have already allowed powerful studies, the lack of tools to quantify more complex stylistic phenomena is evident (Hughes et al., 2012; Gibbs Jr, 2017). Naturally, this work is intended as a first step: the "metaphoricity" degree our system is learning would mirror the kinds of combination present in this specific dataset, which represents a very specific type of metaphor.

It can be argued that we are not really learning the defining ambiguities of an adjective (e.g. the double meaning of "bright") but that we are probably

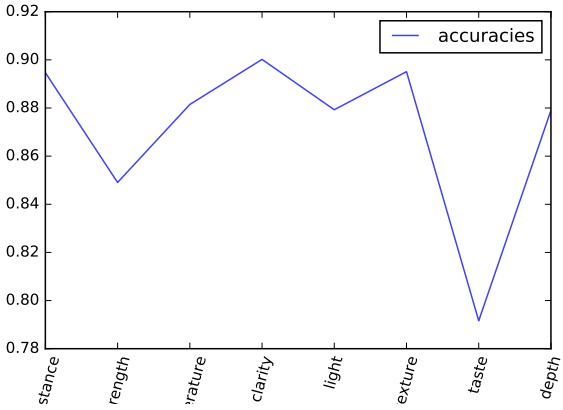


Figure. 12.2: Accuracy training on different categories of adjectives. In this experiment, we train on just one category of the dataset and test on all the others. In general, training on just one category (e.g.*temperature*) and testing on all other categories still yields high accuracy. While the power of generalization of our model is still unclear, we can see that it can detect similar semantic mechanisms even without any vocabulary overlap. The category *taste* is a partial exception: this category seems to be a relative "outlier".

side-learning nouns' degree of abstraction. This would be in harmony with psycholinguistic findings, since detecting nouns' abstraction seems to be one of the main mechanisms we recur to, when we have to judge the metaphoricity of an expression Forgács et al. (2015) and is used as a main feature in traditional Machine Learning approaches to this problem. In other terms, our system seems to detect when the same adjective is used with different categories of words (abstract or concrete) and generalize over this distinction; a behavior that might not be too far from the way a human learns to distinguish different senses of a word.

An issue that we would like to further test in the future is metaphoricity detection on different datasets, to explore the ability of generalization of our models. Researching on different datasets could also help us gaining a better insight about the model's learning.

An obvious option is to test verb-adverb pairs (VA, e.g. *think deeply*) using the same approach discussed in this paper. It would then be interesting to see whether having a common training set for both the AN and the VA pairs will allow the model to generalize for both cases or different

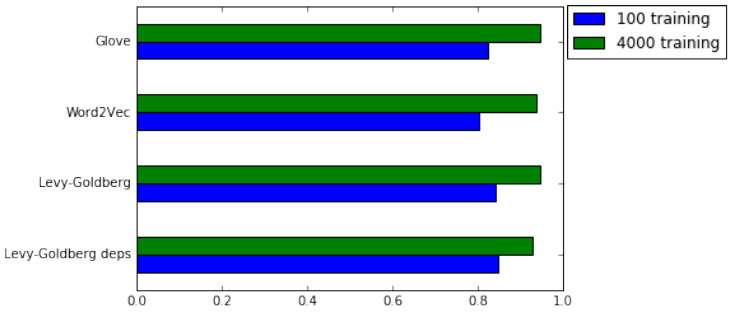


Figure. 12.3: Accuracy on different kinds of embeddings, both training on 100 phrases and 4000 phrases.

Test	Train	Accuracy	Precision	Recall
6929	72	0.83	0.89	0.77
5561	299	0.89	0.86	0.93
4406	643	0.91	0.92	0.90
3239	1203	0.90	0.91	0.88
2253	1961	0.91	0.92	0.92
1568	2763	0.89	0.90	0.90
707	4291	0.91	0.94	0.91
313	5494	0.93	0.92	0.95
148	6282	0.93	0.94	0.92

Table. 12.3: This table shows consistent results in accuracy, precision and recall of the classifier trained with different split points of vocabulary instead of phrases. Splitting the vocabulary creates different sizes of training phrases and test phrases.

training on two training sets, one for AN and one for VA, will be needed. Other cases to test include N-N compounds or proposition/sentence level pairs.

Another way such an approach can be extended, is to investigate whether reasoning tasks typically associated with different classes of adjectives can be performed. One task might be to distinguish adjectives that are intersective, subsective or none of the two. In the first case, from *A N x* one should infer that *x* is both an *A* and an *N* (something that is a black table is both black and a table), in the second case one should infer that *x* is *N* only (for example someone who is a skillful surgeon is only a surgeon but we do not know if s/he is skillful in general), and in the third case neither of the two should be inferred. However, this task is not as simple as giving a training

Top ten	reluctance, reprisal, resignation, response, rivalry, satisfaction, storytelling, supporter, surveillance, vigilance
Bottom ten	saucepan, flour, skillet, chimney, jar, tub, fuselage, pellet, pouch, cupboard

Table. 12.4: 10 most similar and 10 least similar terms with respect to the "metaphoricity vector", concatenated using an all-zeros vector for the adjective. In practice, this is a way to explore which semantic dimensions are particularly useful to the classifier. A concrete/abstract polarity on the nouns was apparently derived

set with instances of AN pairs, to recognize where novel instances of AN pairs belong to. Going beyond logical approaches by having the ability to recognize different uses of an adjective requires a richer notion of context which extends way beyond the AN-pairs.

A further idea we want to pursue in the future is the development of more fine grained datasets, where metaphoricity is not represented as a binary feature but as a gradient property. This means that a classifier should have the ability to predict a degree of metaphoricity and thus allow more fine-grained distinctions to be captured. This is a theoretically interesting side and definitely something that has to be tested since not much literature is available (if at all) on gradient metaphoricity. It seems to us that similar approaches, quantifying a text's metaphoricity and framing it as a supervised learning task, could help having a clear view on the influence of metaphor on style.

Bibliography

Beata Beigman Klebanov, Chee Wee Leong, and Michael Flor. 2015. Supervised word-level metaphor detection: Experiments with concreteness and reweighting of examples. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 11–20.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Lynne Cameron. 2003. *Metaphor in educational discourse*. A&C Black.

Jeanne Fahnestock. 2009. Quid pro nobis. rhetorical stylistics for argument analysis. *Examining argumentation in context. Fifteen studies on strategic maneuvering*, pages 131–152.

Balint Forgács, Megan D. Bardolph, Amsel B.D., DeLong K.A., and M. Kutas. 2015. Metaphors are physical and abstract: Erps to metaphorically modified nouns resemble erps to abstract language. *Front. Hum. Neurosci.*, 9(28).

Raymond W Gibbs Jr. 2017. *Metaphor Wars*. Cambridge University Press.

Nelson Goodman. 1975. The status of style. *Critical Inquiry*, 1(4):799–811.

E Darıo Gutiérrez, Ekaterina Shutova, Tyler Marghetis, and Benjamin K Bergen. 2016. Literal and metaphorical senses in compositional distributional semantic models. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics*, pages 160–170.

James M Hughes, Nicholas J Foti, David C Krakauer, and Daniel N Rockmore. 2012. Quantitative patterns of stylistic influence in the evolution of literature. *Proceedings of the National Academy of Sciences*, 109(20):7682–7686.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Saisuresh Krishnakumaran and Xiaojin Zhu. 2007. Hunting elusive metaphors using lexical resources. In *Proceedings of the Workshop on Computational Approaches to Figurative Language*, FigLanguages '07, pages 13–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

George Lakoff. 1989. Some empirical results about the nature of concepts. *Mind & Language*, 4(1-2):103–129.

George Lakoff. 1993. The contemporary theory of metaphor. In Andrew Ortony, editor, *Metaphor and thought*, page 202–251. Cambridge University Press.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Geoffrey N Leech and Mick Short. 2007. *Style in fiction: A linguistic introduction to English fictional prose*. 13. Pearson Education.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Linlin Li and Caroline Sporleder. 2010a. Linguistic cues for distinguishing literal and non-literal usages. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 683–691. Association for Computational Linguistics.

Linlin Li and Caroline Sporleder. 2010b. Using gaussian mixture models to detect figurative language in context. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 297–300, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jeff Mitchell and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive science*, 34(8):1388–1429.

Yair Neuman, Dan Assaf, Yohai Cohen, Mark Last, Shlomo Argamon, Newton Howard, and Ophir Frieder. 2013. Metaphor identification in large texts corpora. *PloS one*, 8(4):e62343.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Esther Romero and Belén Soria. 2014. Relevance theory and metaphor. *Linguagem em (Dis) curso*, 14(3):489–509.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010. Association for Computational Linguistics.

Paul Simpson. 2004. *Stylistics: A resource book for students*. Psychology Press.

Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the conference on empirical methods in natural language processing*, pages 151–161. Association for Computational Linguistics.

Gerard Steen. 2014. Metaphor and style. *The Cambridge handbook of Stylistics*, pages 315–328.

Lisa Torrey and Jude Shavlik. 2009. Transfer learning. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, 1:242.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 680–690, Stroudsburg, PA, USA. Association for Computational Linguistics.

Carl Vogel. 2001. Dynamic semantics for metaphor. *Metaphor and Symbol*, 16(1-2):59–74.

Deirdre Wilson. 2011. Parallels and differences in the treatment of metaphor in relevance theory and cognitive linguistics. *Intercultural Pragmatics*, 8(2):177–196.

When we describe the location of objects in an image, we relate them by their physical location and by the nature of their interaction. This thesis examines how artificial neural networks learn what information is relevant to spatial descriptions. Favouring "the frog is outside the pond" rather than "the pond is outside the frog" is possible by considering the knowledge about the world and human interactions in language models. The findings of this thesis benefit the design of systems that automatically generate image descriptions and search engines and lead to a more natural human-robot interaction.

ISBN 978-91-7833-916-7 (PRINT)
ISBN 978-91-7833-917-4 (PDF)