Learning visually grounded and multilingual representations

Ákos Kádár

Learning visually grounded and multilingual representations

Ákos Kádár
PhD Thesis
Tilburg University, 2019

TiCC PhD Series No. 64

Financial Support was received from NWO

Cover design: Reka Kadar
Design:
Print:

Learning visually grounded and multilingual representations

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus,
prof.dr. E.H.L. Aarts,
in het openbaar te verdedigen ten overstaan van
een door het college voor promoties aangewezen commissie
in de aula de Universiteit
op woensdag 19 september 2018 om 14.00 uur

door
Ákos Kádár
geboren op 30 december 1989 te Budapest, Hungary

Promotores
Prof. Dr. E. Postma
Dr. A. Alishahi
Dr. G. Chrupała

Commissieleden
Prof. Dr. T. Mitchell
Dr. A. Bisazza
Dr. R. Bernardi
Dr. W.H. Zuidema

"My principal motive is the belief that we can still make admirable sense of our lives even if we cease to have ... an ambition of transcendence"

– Richard Rorty

"Outside is pure energy and colorless substance. All of the rest happens through the mechanism of our senses. Our eyes see just a small fraction of the light in the world. It is a trick to make a colored world, which does not exist outside of human beings."

– Albert Hoffman

No book can ever be finished. While working on it we learn just enough to find it immature the moment we turn away from it."

– Karl Popper

Acknowledgements

I would like to thank my main supervisor Afra Alishahi for providing the PhD opportunity and NWO for the funding. Thanks also go to my second supervisor Grzegorz Chrupała for joining the project and putting effort into improving the thesis. They started me off on my path as a researcher and will have a lasting impact on my future development. Eric Postma, my promotor, oversaw the process and gave crucial input especially at the finishing stages for which I'm grateful. Thank goes to the Integrating Vision and Language COST Action unit for providing me with the "Short Term Scientific Mission grant" and for Tartu University for hosting. Tambet Matiisen was an exceptional host and I hope that we have an entangled future ahead of us. Further thanks go to the organizers of the "Integrating Vision and Language Summer School" for awarding me the trainee grant twice in a row and the same goes to the organizers of the "LxMLS Summer School". I had a wonderful time working for Microsoft Research Montreal during my five months internship and I am very grateful for the opportunity and for the mentorship my colleagues provided. I would also like to thank my collaborators Desmond Elliot and Marc-Alexandre Côté for their contributions to my thesis. It was such a pleasure working together and I hope we will have the opportunity to continue in the future despite the current dystopia created around "intellectual property rights". The future's uncertain and the end is always near. Let it roll, baby, roll!

Contents

1 Introduction

The ability to understand natural language plays a central role in humans' conception of intelligent machines. Alan Turing already in his now famous Imitation Game (Turing, 1950) gives natural language understanding a key role in tricking humans into thinking that they are interacting with their fellow specimen rather than a machine. One of the goals of Natural Language Processing (NLP) is to develop algorithms and build systems to help machines understand what humans are talking about; to understand the *meaning* of natural language utterances.

In the first part of the thesis I explore computational techniques to *learn the meaning of words and sentences* considering the *visual world* as a naturally occurring *meaning representation*. Furthermore, I consider images as a means to *bridge* between languages and present methods seeking to find relationships between *images* and *natural* *utterances* in *multiple languages*.

The Chapters of the thesis follow a progression starting with a single language at the word-level and arriving to multilingual visually grounded sentence representations:

Chapter 1 introduces the topic and contributions of the thesis.

Chapter 2 discusses the related work and technical background in detail.

Chapter 3 presents a cognitive model of language learning that learns *visually grounded word representations*.

Chapter 4 focuses on *visually grounded sentence representations* and their interpretations from a linguistic perspective using the architecture that is the basis for the chapters to follow: combination of a Convolutional Neural Network to extract visual features and a Recurrent Neural Network to learn sentence embeddings.

Chapter 5 applies visually grounded representation learning approach that forms the basis of Chapter 4 to *improve machine translation* in the domain of visually descriptive language.

Chapter 6 shows the clear benefits of learning visually grounded representations for multiple languages jointly.

Chapter 7 extends the investigations of Chapter 6 to the cross-domain setup, breaking the assumption that for each language the same images are annotated with different languages.

1.1 Learning representations

The foundational methodology applied in all chapters is *statistical learning*. The early days of NLP were characterized by rule-based systems building on such foundations as Chomskyan theories of grammar (Chomsky, 1957) or Montague Semantics (Montague, 1970). Since the 1980s partly due to such theories falling out of fashion, but also due to the increase in the amount of available computational power Machine Learning (ML) approaches revolutionized the field. *Learning* in general proved to be a crucial component to Artificial Intelligence and also specifically in NLP. Machine Learning algorithms are designed with the goal that given an increasing number of examples a system improves its performance according to some measure of success. Reflecting the structure of ML itself and the popularity of ML within the field, NLP research follows a task-oriented methodology: researchers borrow or collect data sets, define measures of success and develop or apply learning algorithms. Chapters 5, 6 and 7 closely follow this blueprint.

From the rule based times of ”engineering grammars” researchers moved onto ”engineering features” to *represent* the textual data as input to general-purpose pattern recognition algorithms such as decision trees, support-vector machines or conditional random fields. A large set of these feature templates are still based on various formal-linguistic theories requiring various linguistic taggers and parsers to assign structure to raw texts. Intuitively, different applications such as machine translation or goal-oriented dialogue systems require different input representations. Furthermore, one would assume that various languages require different feature-extraction pipelines reflecting the typological differences across languages.

Linguistic representation learning challenges this intuition and is interested in discovering general principles that allows machines to *learn* linguistic representations from *raw data*, which are more or less generally applicable. This line of work, as well as the approaches presented in the thesis, fit in the general *representation learning* framework consisting of machine learning approaches that learn useful representations for various tasks from (close to) raw input.

The expression ”representation learning” is somewhat synonymous with ”deep learning” at the time of writing this thesis (Bengio et al., 2013). When mentioning representation learning in the deep learning context it is usually meant that the goal is to learn a function from raw input to target labels. In the context of this thesis, however, the emphasis is on learning representations of words, phrases and sentences that are potentially *generally useful*, meaning that they can be used as input to many tasks. This is sometimes referred to as *transfer learning* (Pratt, 1993) where we seek to identify unsupervised learning objectives, supervised tasks, self-supervision schemes or the combinations of these to learn representations that perform well on a large variety of problems.

1.2 Learning representations of words

Most attempts to build general representations for words are based on the *distributional hypothesis* of word meaning. It states that the degree to which words are similar is a function of the similarity of the linguistic context they appear in. In other words, similar words appear in similar contexts. Computational models of distributional semantics implement this intuition and generate real-valued word vectors based on co-occurrence statistics in large text corpora. To aid the reader with technical and historical context we introduce distributional semantics models using the *count-based*/*prediction-based* distinction borrowed from Baroni et al. (2014b). Section 2.1.1 introduces earlier *count-based* methods building word-context co-occurrence vectors-space, while Section 2.1.2 presents the *prediction-based* framework in a more detailed fashion as the techniques discussed here are closely related to the approaches presented in this thesis. Section 2.1.2.2 details efficient linear models for predictive word-learning for two main reasons: 1.) linear word-learning methods had a tremendous impact on shaping the current landscape of continuous linguistic representation learning and are still widely used at the time of writing this thesis, 2.) our main point of comparison for our word learning model in Chapter 3 is one of such models detailed in the section.

Word-representations within the prediction-based framework are an instance of *representation learning*: word representations – usually referred to as *word embeddings* – are learned through optimizing model parameters to predict context from words or words from context. Such learned word-representations have proven successful in many applications especially in recent years, however, they are not *realistic* in a certain sense. While they capture many aspects of syntax and semantics of natural language they are not connected to the real world outside of the large collections of texts. This leads us to the main topic of the thesis: *visual grounding* introduced in the following section.

1.3 Visually grounded word representations

Many theories of human cognition supported by empirical evidence state that human language and concept representation and acquisition is *grounded* in perceptual and sensori-motor experiences. Cross-situational word learning, an influential cognitive account of human word learning, supposes that humans learn the meanings of words exploiting repeated exposure to linguistic contexts paired with perceptual reality. Learning representations for linguistic units in a visually grounded manner brings computational language learning systems closer to human-like learning. Such theoretical considerations are detailed in Section 2.2.1.

Let us also consider the practical applicability of distributional language representations in the larger scope of Artificial Intelligence. One of the dreams of AI is to develop technology to power intelligent embodied agents taking the form of office assistants or emergency aid robots. These machines cannot implement natural language as an arbitrary symbol manipulation system akin to a calculator's understanding of magnitudes or slopes. Similarly to humans they need to link linguistic knowledge to the extra-linguistic world.

Furthermore, while certain aspects of meaning such as *encyclopedic* knowledge are abudant textual data, *perceptual* information can provide complementary valuable insights into physical properties rarely mentioned in texts such as *size*, *shape* and *color*. In practice harnessing the visual modality to learn language representations that link linguistic knowledge to the external world has been empirically shown to improve performance on several semantic tasks as detailed in Section 2.2.2.

In terms of computational modeling the jump from distributional to grounded models is conceptually simple: one needs to collect data where the *contexts* of linguistic units are *extra-linguistic* and represent these contexts such that they can be provided as input to representation learning algorithms. More concretely in terms of extra-linguistic context the present thesis focuses on the *visual modality*.

Linguistic-visual multi-modal representations on the word level have a well established albeit somewhat brief history (Section 2.2.2). Such methods were developed both within the *count-based* and *prediction-based* frameworks using computer vision techniques to represent the *visual modality* and NLP methods to represent texts. These separate spaces are then combined into a single multi-modal representation.

As the **first contribution** of the thesis in Chapter 3 we present an incremental cross-situational model of word learning introducing modern computer-vision techniques to computational cross-situational modeling of human language learning. Through our experiments we show that our presented model is competitive with state-of-the-art *prediction-based* distributional models and that our model can name relevant concepts given images.

1.4 Visually grounded sentence representations

When moving from *atomic* words to the *compositional* world of sentences we need flexible models that can represent word-order and hierarchical relationships. In Chapters 4, 5, 6 and 7 we use Recurrent Neural Networks, which form a powerful class of sequence models to represent sentences. Section 2.3 provides the reader with historical and technical background to the considerations behind this choice.

The study of general sentence representation learning has a much briefer history than word-representations and Section 2.4 situates the reader in the area. Most approaches to learn useful sentence representations to date are based also on the distributional hypothesis and formulate general purpose representation learning as a sort of linguistic context prediction, but on the sentence level.

Section 2.5 describes the general framework of learning visually-grounded sentence representations and their utility. The basic idea is still context prediction, however, we learn associations between sentences and their *visual* context i.e model parameters are optimized such that related image sentence pairs get pushed close together and unrelated pairs far from each other in a learned joint space.

As the **second contribution** of this thesis in Chapter 4 we train such an architecture and explore the learned representations. Our main interest and contribution here is the development of general techniques to *interpret* linguistic representations learned by Recurrent Neural Networks and use these techniques to contrast text-only language models with their grounded counter parts trained on the same sentences.

1.5 Visual modality bridging between languages

One of the intriguing aspects of using the visual modality as a naturally occurring meaning representation is that it is also naturally *universal* across languages. The visual modality anchors linguistic representations to perceptual reality, but also provides a natural bridge between various languages. Linguistic utterances that are similar to each other, intuitively, appear in the context of perceptually similar scenes across languages.

Utterances in multiple languages and corresponding perceptual stimuli can be conceptualized as *multiple views* of the same underlying abstract object. Learning to map these multiple views to the same feature space can lead to better representations as they have to be more *general* due to the model having to solve multiple tasks at the same time. This *multi-view learning* perspective is explained in more detail in Section 2.6.1 focusing on the specific case of multi-modal and multi-lingual representations we explore in Chapter 6 and 7.

The visual modality as pivot on the word-level can be used to find possible translations for words when no dictionary is available. Extending this idea from word to sentence level gives rise to techniques that use the visual modality as a pivot to translate full sentences. Approaches in this direction are discussed in Section 2.6.2.

The **third contribution** in the thesis combines visually grounded sentence representation learning with machine translation. More specifically in Chapter 5 we present a *multi-task* learning architecture that jointly learns to associate English sentences with images and to translate from English to German. We show that visually grounded learning improves translation quality in our domain and that it provides orthogonal improvements to having a large additional English-German parallel corpus.

The **fourth contribution** of this thesis is exploring visually grounded sentence representations learned for multiple languages jointly. In Chapter 6 we show that better grounded representations can be learned by training on multiple languages. We find a consistent pattern of improvement whereby multilingual visually grounded sentence representations outperform bilingual ones, which outperform monolingual representations. Furthermore, we provide empirical evidence that the quality of visually grounded sentence embeddings on lower-resource languages can be improved by jointly training together with data sets from higher-resource languages.

Lastly, our **fifth contribution** in Chapter 7 is exploring the benefit of multilinguality in visually grounded representation learning as in Chapter 6, but in the cross-domain setting. Here we consider a *disjoint* scenario where the image sentence data sets for different languages do not share images. We assess how the method applied in Chapter 6 performs under domain-shift. Furthermore, we introduce a technique we call *pseudopairs*, whereby we generate new image–caption data sets by creating pairs across data sets using the sentence similarities under the learned representations. We find that even though this technique does not require any additional external data source, models or other pipeline elements, it consistently improves image sentence ranking performance.

1.6 Published work

1.6.1 Chapters

Each of the following Chapters has been previously published. They are included with the only modification of re-aligning and re-sizing a few figures.

Chapter 3 Kádár, Á., Alishahi, A., & Chrupała, G. (2015a). Learning word meanings from images of natural scenes. *Traitement Automatique des Langues*, 55(3)

Chapter 4 Kádár, A., Chrupała, G., & Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4), 761–780

Chapter 5 Elliott, D. & Kádár, À. (2017). Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1 (pp. 130–141)

Chapter 6 Kádár, Á., Elliott, D., Côtė, M.-A., Chrupała, G., & Alishahi, A. (2018b). Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 402–412)

At the time of completing the thesis Chapter 7 has been submitted to the 2019 Conference on Empirical Methods in Natural Language Processing without modifications.

1.6.2 Publications completed during the PhD

These publications were completed during my PhD work, but have not been included in the thesis.

1.6.2.1 Publications on Vision and Language

- Chrupała, G., Kádár, À., & Alishahi, A. (2015). Learning language through pictures. In *Proceedings of the 53rd Annual*

Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), volume 2 (pp. 112–118)

- Kádár, A., Chrupała, G., & Alishahi, A. (2015b). Lingusitic analysis of multi-modal recurrent neural networks. In *Proceedings of the Fourth Workshop on Vision and Language* (pp. 8–9)

- Kahou, S. E., Atkinson, A., Michalski, V., Kádár, Á., Trischler, A., & Bengio, Y. (2018). Figureqa: An annotated figure dataset for visual reasoning

- van Miltenburg, E., Kádar, A., Koolen, R., & Krahmer, E. (2018). Didec: The dutch image description and eye-tracking corpus. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3658–3669)

1.6.2.2 Publications on other topics

- Chrupała, G., Gelderloos, L., Kádár, Á., & Alishahi, A. (2019). On the difficulty of a distributional semantics of spoken language. *Proceedings of the Society for Computation in Linguistics*, 2(1), 167–173

- Kádár, Á., Côté, M.-A., Chrupała, G., & Alishahi, A. (2018a). Revisiting the hierarchical multiscale lstm. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3215–3227)

- Ferreira, T. C., Moussallem, D., Kádár, Á., Wubben, S., & Krahmer, E. (2018). Neuralreg: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1959–1969)

- Côté, M.-A., Kádár, Á., Yuan, X., Kybartas, B., Barnes, T., Fine, E., Moore, J., Hausknecht, M., El Asri, L., Adada, M., et al. (2018). Textworld: A learning environment for text-based games. In *Proceedings of the Computer Games Workshop at ICML/IJCAI 2018* (pp. 1–29)

- Manjavacas, E., Kádár, Á., & Kestemont, M. (2019). Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

2

Background

2.1 Distributed word-representations

The distributional approach to word meaning hypothesizes that semantically related words tend to appear in similar contexts. This idea goes back in linguistics tradition to the earlier days of American structuralism (Nevin & Johnson, 2002). As early as in 1954 in his seminal paper ”Distributional structure” Harris claims that distribution should be taken as an explanation for word meaning and that similarity classes can be constructed based on co-occurrence statistics (Harris, 1954). Cruse & Cruse (1986) write that ”It is assumed that the semantic properties of a lexical item are fully reflected in the appropriate aspects of the relations it contracts with actual and potential contexts” and that ”the meaning of a word is constituted by its contextual relations”. Computational models of distributional semantics implement this intuition and generate real-valued word vectors based on co-occurrence statistics in large text corpora. Here we present such representations using the *count-based* and *prediction-based* distinction borrowed from Baroni et al. (2014b).

2.1.1 Count-based approaches

Early computational linguistics models of distributional semantics fall in the category of count-based approaches: they store the number of times target words appear in different contexts. In the resulting co-occurrence matrix each row corresponds to a word and each column to a context. Each cell in the matrix is the number of times a word appears in a particular context. The size of the co-occurrence matrix is then vocabulary size by the number of contexts.

Contexts are typically words appearing within a certain window size or text documents. To the counts in the co-occurrence matrix various re-weighting schemes are applied followed by some matrix factorization algorithm resulting in a lower dimensional dense representation.

The earliest approaches include Hyperspace Analogue to Language (Lund & Burgess, 1996), which constructs a term-term co-occurrence matrix and Latent Semantic Analysis (Dumais, 2004), which applies the tf-idf re-weighting scheme on a term-document matrix followed by singular value decomposition. More recent approaches apply different re-weighting schemes such as point-wise mutual information (Bullinaria & Levy, 2007) and local mutual information (Evert, 2005) or different matrix factorization algorithms such as non-negative matrix factorization (Baroni et al., 2014b).For a comprehensive set of empirical experiments on count-based approaches please consult Bullinaria & Levy (2007) and Bullinaria & Levy (2012).

2.1.2 Prediction-based approaches

In more recent years various deep learning methods have been applied to learn distributed word representations usually referred to as *word embeddings* in the literature. Contrary to count-based methods prediction-based approaches fit into the standard machine learning pipeline: they optimize a set of parameters to maximize the probability of words given contexts or contexts given words, where the word-embeddings themselves form a subset of the parameters of the full model.

2.1.2.1 Neural language models

Laying down the framework for recent developments the first modern approach to learn distributed word-representations from realistic data was the neural language model introduced by Bengio et al. (2003). They present a feed-forward multilayer perceptron with continuous word-embeddings, a single hidden layer and a softmax output layer. More precisely the model is parametrized by a 1.) word-embedding matrix, whose rows correspond to word-vectors and columns to a learned feature, 2.) hidden and output weight-matrices. The network takes as input the concatenation of n word-vectors preceding the target word as context representation and outputs the probability distribution over the current word. The model is trained to maximize the probability of the target word given the previous fixed number of words as context over a training corpus – n-gram language model –

trained with stochastic gradient descent (Cauchy, 1847) through the backpropagation algorithm (Rumelhart et al., 1985). Shortly after its publication the neural probabilistic feed-forward neural language model of Bengio et al. (2003) has been shown to improve performance in speech recognition (Schwenk & Gauvain, 2005).

Following a similar recipe, the convolutional architecture of Collobert & Weston (2008) based on the time-delay neural network model (Waibel et al., 1990) took several steps towards the by now standard practices in deep NLP. Contrary to the feed-forward network the convolution over-time structure can handle sequences of variable length. This is essential for NLP applications where typically sentences are composed of a varying number of words. Collobert & Weston (2008) introduce the idea of jointly learning many linguistic tasks at the same time such as part-of-speech tagging, chunking, named entity recognition and semantic role labeling through *multi-task learning* (Caruana, 1997). Their architecture was later refined in Collobert et al. (2011) and the pre-trained full model was made available alongside the standalone word-embeddings. Finally, Collobert & Weston (2008) were the first to show the utility of pre-trained word-embeddings in other tasks through transfer learning.

In this thesis we make extensive use of the multi-task learning strategy: In Chapters 4 we apply it to language modeling and image-sentence ranking, in Chapter 5 for machine translation and image-sentence ranking, while in Chapters 6 and 7 for image-sentence and sentence-sentence ranking in multiple languages.

The arguably most popular architecture in NLP, however, which we also use in Chapters 4, 5, 6 and 7 is the recurrent neural network (RNN). It was first introduced by the late Jeffrey L. Elman in his seminal paper "Finding Structure in Time" as a model of human language learning and processing (Elman, 1990). Recurrent neural networks take a variable length sequence as input and at every time-step they compute their next state based on the previous state and the current input. Recurrent networks essentially "read" the input left-to-right and keeps track of the context when encoutering a new input. Equation 2.1 provides the recursive definition of the computation in a RNN language model:

$$P(w_t|w_{<t}) = \text{softmax}(\mathbf{U}\mathbf{h_t} + \mathbf{b_o}) \tag{2.1}$$

$$\mathbf{h_t} = \tanh(\mathbf{W_h}\mathbf{h_{t-1}} + \mathbf{W_i}\mathbf{w_t} + \mathbf{b_h}) \tag{2.2}$$

In case of language modeling at each time-step t the network takes an input word vector $\mathbf{w_t}$ and its previous hidden state $\mathbf{h_{t-1}}$ maintained through previous time-steps. These are used to compute the current state $\mathbf{h_t}$ and to predict the probability distribution over the following word $P(w_t|w_{<t})$. It is parametrized by a word-embedding matrix $\mathbf{W} \in \mathbb{R}^{|V| \times d}$, an input to hidden weight matrix $\mathbf{W_i}$, a hidden to hidden weight matrix $\mathbf{W_h}$ and finally the hidden to output weight matrix $\mathbf{U}$ to predict the unnormalized probabilities over the vocabulary entries and additional hidden and output bias terms $\mathbf{b_o}$ and $\mathbf{b_h}$. This model is trained to maximize the probability of the training sequences, with the backpropagation through time algorithm (BPTT) (Robinson & Fallside, 1987; Werbos, 1988; Williams & Zipser, 1995).

Elman (1991) shows when trained on simple natural language-like input hidden states of the network $\mathbf{h_t}$ encode grammatical relations and hierarchical constituent structure. In Chapter 4 we also train a recurrent language model and compare it to its grounded counterpart on real-world data and explore similar questions about the learned opaque representations as Elman (1991) .

Despite the early successes, however, it turned out to be difficult to train Elman networks on practical applications with longer sequences due to the vanishing and exploding gradient phenomena (Bengio et al., 1994). The RNNLM implementation of Mikolov et al. (2010), however, established a new state-of-the-art on language modeling and RNNs regained their popularity in language processing. At the time of writing this thesis RNNs remain a widely used in language processing and are still trained with BPTT with various versions of stochastic gradient decent and smart initialization strategies.

While more complex training algorithms such as hessian-free optimizers (Martens & Sutskever, 2011) remain out of fashion, an overwhelming amount of empirical evidence shows that the more compliex long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997; Gers et al., 1999) and gated recurrent unit networks (GRU) (Cho et al., 2014b) recurrent network variants vastly outperform the simple Elman network in practice. We opted for GRUs in Chapters 4, 5, 6 and 7.

2.1.2.2 Efficient linear models

Despite the success of deep learning architectures in NLP it wasn't until the introduction of the much simpler and fast continuous bag-of-words and skip-gram with negative sampling (SGNS) algorithms Mikolov et al. (2013a) packaged into the easy-to-use `word2vec` toolkit that word-embeddings became ubiquitous in computational linguistics and NLP research. These algorithms rely on simple log-linear models as opposed to the more expensive neural networks leading to faster training on larger corpora. The more successful SGNS model has two learnable word embedding matrices one for the target words and a separate one for the contexts. The model is trained to maximize the dot product between true word-context pairs appearing in corpus and minimize the dot product between randomly sampled contrastive examples.

Similarly to the negative sampling algorithm of Mikolov et al. (2013a) the ranking objectives implemented in Chapters 5, 6 and 7 force the models to push images and corresponding sentences close and contrastive examples far from each other in the learned multimodal space.

The GloVe approach (Pennington et al., 2014) is another popular simple linear model with word and context embeddings, which represents a hybrid between count- and prediction-based techniques: it optimizes word-embeddings to predict the re-weighted log co-occurrence counts collected from large text corpora.

There has been work on finding relationships between count- and prediction-based methods (Levy & Goldberg, 2014) and using insights from both to develop novel improved variants (Levy et al., 2015).

2.2 Visually grounded representations of words

2.2.1 Language and perception

The discussion of learning distributed representations so far has focused on implementations of the distributional hypothesis and extracting information exclusively from text corpora. To human language learners, however, a plethora of perceptual information is available to aid the learning process and to enrich their mental representations. The link between human word and concept representation and acquisition and the perceptual-motor systems has been well established through behavioral neuroscientific experiments (Pulvermüller, 2005). The earliest words children learn tend to be names of concrete perceptual phenomena such as objects, colors and simple actions (Bornstein et al., 2004). Furthermore, children generalize to the names of novel objects based on perceptual cues such as shape or color (Landau et al., 1998). In general, the *embodiment*-based theories of concept representation and acquisition in the cognitive scientific literature put forward the view that a wide variety of cognitive processes are grounded in perception and action (Meteyard & Vigliocco, 2008). The precise role of sensori-motor information in language acquisition and representatio, however, is a highly debated topic (Meteyard et al., 2012).

Motivated by such cognitive theories and experimental data, various computational cognitive models of child language acquisition investigate the issue of learning word meanings from small scale or synthetic multi-modal data. The model presented by Yu (2005) uses visual information to learn the meanings of object names whereas the architecture of Roy (2002) learns to associate word sequences with simple shapes in a synthetically generated data setting.

Interestingly even the articles introducing Latent Semantic Analysis (Landauer & Dumais, 1997) and Hyperspace Analogue to Language (Lund & Burgess, 1996) mention that a possible limitation of the presented distributional semantic models is the lack of grounding in extra-linguistic reality. Landauer & Dumais (1997) puts it as ”But still, to be more than an abstract system like mathematics words must touch reality at least occasionally.” The lack of relationship between symbols and the external reality is usually referred to as the *grounding problem* in the literature (Harnad, 1990; Perfetti, 1998). On the defense of purely textual models Louwerse (2011) argues that the corpora used to train distributional semantic models are generated by humans and as such reflect the perceptual world. For a counter argument consider the few pieces of text that would state obvious perceptual facts such as "bananas are yellow" or how often objects with the property "yellow" would appear in similar textual contexts (Bruni et al., 2014).

In practice much work on multi-modal distributional semantics have found that text-only spaces tend to represent more encyclopedic knowledge, whereas multi-modal representations capture more concrete aspects (Andrews et al., 2009; Baroni & Lenci, 2008). In Chapter 3, where we develop a cross-situational cognitive model of word-learning, we also find that the word-representations learned by our model correlate better with human similarity judgements on more concrete than abstract words. In contrast, word-embeddings learned by the SGNS algorithm trained on the same sentences perform better on more abstract words.

One does not necessarily need to reach a conclusion on whether grounded or distributional models are superior; combining their merits in a pragmatic way is an attractive alternative (Riordan & Jones, 2011).

2.2.2 Combined distributional and visual spaces

When learning multimodal word representations we wish to construct a matrix, where each row corresponds to a word and each feature column represents a distributional feature, a perceptual feature or a mixture of the two.

The first approach to learn visual word representations from realistic data sets was introduced by Feng & Lapata (2010). They develop a multi-modal topic model trained on a BBC News data set containing text articles with image illustrations. Each document considered by their model is a pair of a text document and an image generated by a mixture of latent multi-modal topics. The documents are represented as bag-of-words vectors (BoW), while images are representedas bag-of-visual-words (BoVW) (Csurka et al., 2004) using a difference-of-Gaussians point detector for image segmentations SIFT features for local region descriptors (Lowe, 1999). Each document is then represented as the concatenation of BoW textual an BoVW visual features. A Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic model is trained on joint representations and after convergence each word is represented by a vector, where each component corresponds to the conditional probability of that word given a particular multi-modal topic. Feng & Lapata (2010) shows that the multi-modal LDA model outperforms the text-only LDA representations by a large margin on word association and word similarity experiments.

The perceptually grounded word representations of Bruni et al. (2012) also combines distributional semantic models and BoVW pipelines. Rather than having a collection of documents they consider a set of words for which both distributional and image features are available. For visual word representations they use images labeled with tags by annotators. Similarly to Feng & Lapata (2010) they apply a BoVW pipeline feature represent images. Contrary to Feng & Lapata (2010), however, a visual-only representation is created for each tag-word by summing over the features for all images corresponding to the tag. For text-only models Bruni et al. (2012)construct several types of distributional semantic spaces from text-only corpora unrelated to the images. As in Feng & Lapata (2010) they wish to create a multi-modal word representation and applied separate pipelines to extract textual and visual features. Finally they create the multi-modal space through concatenation.

On word-similarity benchmarks Bruni et al. (2012) show that the text-only model performs better than visual-only and that the combination of the two surpasses both. They also find that distributional semantics models perform poorly on finding the typical colors of concrete nouns, whereas the visual and multi-modal models perform perfectly. In these experiments distributional semantics models fail to capture the obvious fact that "the grass is green" providing evidence against the theoretical argument that perceptual information is available in large collections of texts and so grounded representations are superfluous (Louwerse, 2011).

Combining BoVW and count-based distributional word-representations remained the standard methodology in many other works on multi-modal word representations at the time (Bruni et al., 2011; Leong & Mihalcea, 2011a,b). Bruni et al. (2014) frames multi-modal distributional semantics under a general framework: create separate textual and visual features for words followed by re-weighting and matrix factorization. Researchers can apply vector-space models to create the distributional space computer vision techniques to create visual feature representations for the same words. The separate feature spaces are mixed together first by concatenation, which is optionally followed by singular value decomposition to combine the two modalities.[1]

For example Kiela & Bottou (2014) run the SGNS algorithm on large text corpora and take the word-embedding parameters as the distributional space and apply pre-trained convolutional neural networks (CNN) as black-box image feature extractors. We apply CNNs as image feature extractors in all chapters.

Convolutional neural networks learn a hierarchy of blocks of image filters followed by pre-defined pooling operations optimized for a particular task. It has been observed in the computer vision community that the lower layers of deep CNNs trained across various data sets and tasks tend to learn filter maps that resemble Gabor filters and color blobs. Intuitively these low-level features appear to be general and as such afford *transfer*. There is an extensive body of work on exploring the transferability of CNN features to various computer vision tasks through fine-tuning (Donahue et al., 2014b; Oquab et al., 2014) or simply taking the last layer representation of CNNs as high-level features as inputs to linear classifiers (Girshick et al., 2014; Sharif Razavian et al., 2014). Given their success in transfer learning in computer vision it is natural to apply CNNs in the visually grounded language learning community as black-box image feature extractors. Similarly to Bruni et al. (2014) in Kiela & Bottou (2014) the visual features of words are computed as the summary the feature vectors extracted from all images they co-occur with. Applying

[1]We remark that discovering latent factors through the application of singular value decomposition on the concatenated textual and visual spaces is similar in spirit to applying topic models such as LDA.

the simple concatenation operation to the two spaces they show that CNN features outperform BoVW features on word similarity experiments with the multi-modal word representations.

All approaches described so far require both textual and visual information for the same concepts and representations for these modalities are learned separately and are fused later. The multi-modal skip-gram (Lazaridou et al., 2015) model was developed to alleviate such limitation: it is a multi-task extension of the skip-gram algorithm predicting both the context of words, but also the visual representations of concrete nouns. This optimizes word representations to be predictive of both their visual and textual contexts jointly. Ground truth visual representations are constructed by averaging the CNN representations (Krizhevsky et al., 2012) of 100 pictures sampled from ImageNet (Deng et al., 2009). This architecture was later proposed as a model of child language learning and was applied to the CHILDES corpus (MacWhinney, 2014) with modifications to model referential uncertainty and social cues Lazaridou et al. (2016) present in language learning. Later it was compared to human performance in terms of learning the meaning of novel words from minimal exposure (Lazaridou et al., 2017).

Chapter 3 presents a computational cognitive model of word learning using only visual information developed at the same time as the multi-modal skip-gram approach. Similarly to Feng & Lapata (2010) we assume pairs of text and images. However, in our data set images represent everyday scenes and are paired with descriptive sentences to mimic the language environment of children on a high level. Our model is inspired by the cross-situational account to language learning and assumes that the representations of words are learned exclusively through the co-occurrences between visual features and words. Rather than building a matrix of image-features per word and then summarizing as in Kiela & Bottou (2014), we apply an online expectation-maximization-based algorithm (Dempster et al., 1977) to align words with image features mimicking child language learning. In essence, our approach combines the cross-situational incremental word-learning model of Fazly et al. (2010b) with the larger realistic data sets, modern convolutional image representations and extends it to operate on real-valued scene representations. The result of the learning process is a word embedding matrix, where each row corresponds to a word, each column to a CNN feature and each entry to the strength of the relationship between the word and an image-feature.

We show through word-similarity experiments that, while our approach performs on par with the SGNS trained on the same text data, there is a qualitative difference between the learned embeddings: the correlation between our visual word-embeddings and human similarity judgements is significantly higher for concrete than abstract nouns. As each word embedding is represented in the image-feature space as in Kiela & Bottou (2014), we show that our model can label images with related nouns through simple nearest neighbor search.

2.3 From words to sentences

Applying the distributional intuition to model the meaning of sentences is not as straightforward as it is for words. In a sentence-embedding matrix each row would correspond to a possible sentence and each column to a feature. Intuitively however, the number of words in a corpus is much lower than the number of sentences: one can assume a large, but finite set of existing words and an infinite set of potential sentences to *compose* from them. Words can be thought of as *atomic units* and in downstream applications one can use a lookup operation on a word embedding matrix to represent units in the input. However, it is infeasible to look up full sentences. In fact, from a method that represents sentences in a continuous space one would expect to also represent and generalize to unseen sentences at test time. Furthermore, given two sentences *John loves Mary* and *Mary loves John* we wish our sentence encoding function ϕ to represent the meaningful difference stemming from the underlying syntactic structure $\phi(\textit{John loves Mary}) \neq \phi(\textit{Mary loves John})$. As such we seek to learn a sentence encoder that is sensitive to syntactic structure and semantic compositionality i.e.: the notion that the meaning of an expression is a function of its parts and the rules combining them (Montague, 1970).[2]

The compositional distributional semantics framework produces continuous representations for phrases up to sentences using additive and multiplicative interactions of count-based distributed word-representations (Mitchell & Lapata, 2008) or combine symbolic and continuous representations with tensor-products (Clark & Pulman, 2007). The latter line of work culminated in a number of unified theories of distributional semantics and formal type logical and categorical grammars (Coecke et al., 2010; Clarke, 2012; Baroni et al., 2014a). These approaches assume that words are represented by distributional word embeddings and define compositional operators on

[2]Similarly one could argue that *morphemes* are atomic units from which words are composed. In fact, sub-word representations form an important field of research (Bojanowski et al., 2017).

top of them motivated by particular formal semantic considerations. From the point of view of theoretical linguistics arguably one of the most intriguing aspects of such theories of meaning is that they provide an elegant data-driven solution to deal with the representation of the lexical entries of content words – nouns, verbs and adjectives. Within applied NLP, however, this line of work has not resulted in practical machine learning approaches to solve natural language tasks on real world data sets. This is likely due to the different scope and the computationally expensive high-order tensor operations involved (Bowman, 2016).

Lastly, before going forward with the more recent neural models in the next section, it is only fair to mention that bag-of-words based representations bypassing the issue of compositionality form a set of very strong baselines for a number of sentence level tasks (Hill et al., 2016). These simple baselines include using multinomial naive Bayes uni- and bigram log-count features within support vector machines (Wang & Manning, 2012) or feeding the average of the word-embeddings in a sentence into a softmax classifier (Joulin et al., 2016).

2.4 Neural sentence representations

In Section 2.1 we have discussed the feed-forward (Bengio et al., 2003) and recurrent network (Mikolov et al., 2010) language models from the perspective of learning word-representations. However, both architectures learn embeddings of not only single words, but also learn to represent sequences of multiple words such as sentences. An intriguing property of such approaches is that they represent various linguistic objects in the same space as activations of the neural models.[3] When learning transferable sentence representations there are two main considerations we will discuss: 1.) which architecture to choose and 2.) what objective to optimize. Various neural network architectures have been proposed that handle variable-sized data structures useful for language processing: recurrent networks take the input sequentially one word or character at a time, convolutional neural networks (Kalchbrenner et al., 2014; Zhang et al., 2015; Conneau et al., 2016; Chen et al., 2013) process sequences in fixed-sized n-gram patterns up to a large window, recursive neural networks (Goller & Kuchler, 1996; Socher et al., 2011; Tai et al., 2015) take a tree data structure as input such as a sentence according to the traversal of a constituency and graph neural networks operate on graphs (Marcheggiani & Titov, 2017) such as syntactic/semantic dependency or abstract meaning representations. All the aforementioned architectures take word-embeddings as input and compute fixed vectors for sentences. These representations are tuned to a specific task such as sequence tagging, sentence classification, machine translation, parsing or language modeling.

In Chapters 4, 5, 6 and 7 we decided to apply recurrent neural networks as sentence encoders. Recursive neural networks provide a principled approach to compute representations along the nodes of constituency or dependency parse trees (Socher et al., 2013, 2014; Le & Zuidema, 2015; Tai et al., 2015). In practice, however, these architectures require parse trees as input, which makes them impractical for our mission of learning visually grounded representations for mul-

[3]This property is not unique to neural models e.g. the pregroup algebra based compositional distributional semantics framework of Coecke et al. (2010) sentences of any grammatical structure live in the same inner-product space.

tiple languages. For each language considered, the training procedure requires finding a good pipeline from text pre-processing to parsing to generate the input representations for the networks. Furthermore, tree structures by nature do not afford straightforward batched computation directly and tend to run dramatically slower than recurrent or convolutional models. In terms of performance the jury is still out, however, so far only modest improvements have been observed over recurrent models on specific tasks in specific settings (Li et al., 2015; Tai et al., 2015). For graph structured neural networks the same argument holds. Two equally practical alternatives to RNNs that operate on raw sequences are convolutional neural networks (Bai et al., 2018) and transformers (Vaswani et al., 2017) and both could replace the RNNs in Chapters 5, 6 and 7.

As mentioned earlier contrary to word-embeddings the representations and the composition function these models learn are task-specific and not *universal*. For learning transferable and general distributed sentence representations the first notable approach is the skip-thought vectors model (Kiros et al., 2015). It extends distributional semantics intuition to the sentence level: they train a sentence encoder on contiguous sentence sequences – such as books – which learns sentence representations predictive of the sentences around it. More specifically they train a recurrent encoder to encode a target sentence and use two separate recurrent decoders to generate its pre-context and post-context. This method was confirmed to be a successful self-supervised method for transfer learning in a number of sentence classification tasks beating simpler approaches such as bag-of-words approaches based on skip-gram or CBOW embeddings, tf-idf vectors and auto-encoders (Hill et al., 2016). A convolutional variant of the encoder was introduced in Gan et al. (2016) and several other works train simple sum/average pre-trained word-embedding encoders using the same sentential context prediction objective (Kenter et al., 2016; Hill et al., 2016). The larger context of paragraphs is explored in Jernite et al. (2017) where the task is to predict discourse coherence labels in a self-supervised fashion.

Some later approaches have moved away from distributional cues and identifyied supervised tasks that lead to representations that transfer well to a wide variety of other tasks. The task of Natural Language Inference (Bowman et al., 2015; Williams et al., 2017) as an objective was identified to learn good sentence embeddings (Conneau et al., 2017; Kiros & Chan, 2018) and Subramanian et al. (2018) combine a number of other supervised tasks with self-supervised training through multi-task learning.

The state-of-the-art in learning universal sentence representations at the time of writing is represented by neural language models with a large number of parameters trained over huge corpora (Peters et al., 2018; Devlin et al., 2018). These approaches go back to exploiting only distributional cues and train a stack of convolutional and/or recurrent and/or transfomer layers on large-scale language modeling. When transferring the networks to novel tasks they are either fine-tuned, a separate smaller network is trained on top of their representations or they are used as fixed feature extractors (Howard & Ruder, 2018; Peters et al., 2019).

2.5 Visually grounded sentence representations

Universal sentence representations are in general learned from text-only corpora. The most successful current trend is large-scale language modeling based on the distributional semantics intuition of the general usefulness of linguistic context prediction. However, this leaves the resulting sentence representations blind to the language-external reality leading to the grounding problem as discussed in Section 2.2. Given that visual information has been shown to contain useful information for word-representations it is a natural question to ask whether this observation generalizes to sentence embeddings. The idea of context prediction coming from the distributional hypothesis can be adopted to visual grounding in a conceptually straightforward manner: train sentence embeddings to be predictive of their visual context.

The larger family of techniques that our visually grounded sentence learning approach technically belongs to is *learning to rank* (Li, 2011). Some of the earlier attempts at multi-modal ranking did not consider full sentences rather, images paired with (mostly) noun tags such as Weston et al. (2010).

Framing the learning of multi-modal vision and language spaces as a ranking problem from images to descriptions and conversely from descriptions to images was put forward by Hodosh et al. (2013). They argue that evaluating grounded learning through image–description generation is plagued by the lack of straightforward performance measures. This issue was later discussed in Elliott & Keller (2014) who demonstrate low to moderate correlation between automatic measures such as BLUE, METEOR and ROUGE with human judgements.

From the image–sentence ranking perspective a joint space between language and vision reflects accurately the underlying semantics if given one modality as *query* the other modality can be accurately *retrieved*. This leads to a straightforward evaluation protocol adopting metrics from information retrieval literature such as Recall@N, Precision@N, Mean Reciprocal Rank, Median Rank or Mean Rank. From the practical point of view it also unifies image–annotation and image–search based on language queries.

The standard benchmark data sets we used for this purpose annotate images found in online resources with descriptions through crowd-sourcing. These descriptions are largely *conceptual*, *concrete* and *generic*. This means that descriptions do not focus too much on *perceptual* information such as colors, contrast or picture quality; they do not mention many *abstract* notions about images such as mood and finally the descriptions are not *specific* meaning that they do not mention the names of cities, people or brands of objects. What they do end up mentioning are entities depicted in the images (frisbee, dog, boy) their attributes (yellow, fluffy, young) and the relations between them. The images depict common real-life scenes such as a *bus turning left* or *people playing soccer in the park*. As such, annotation collected independently from different crowd-source workers end up focusing on different aspects of these scenes. For a comprehensive overview on image-description data sets please consult Bernardi et al. (2016).

Following the ranking formulation of grounded learning by Hodosh et al. (2013) the earliest works applying a combination of sentence and image encoder neural networks to image–sentence ranking were put forward by Kiros et al. (2014) and Socher et al. (2014): they both apply pre-trained convolutional neural networks as image encoders and while Socher et al. (2014) use recursive neural network variants to encode sentences Kiros et al. (2014) apply recurrent networks. Rather than matching whole images with full sentences the alternative approach of learning latent alignments between image–regions and sentence fragments have also been explored concurrently (Karpathy et al., 2014; Karpathy & Fei-Fei, 2015).

Following Kiros et al. (2014) to predict images from the sentences – and conversely sentences from the images – the architecture we chose in Chapters 4, 5, 6 and 7 combines a recurrent neural network to represent sentences and a pre-trained convolutional neural network followed by an affine transformation we train for the task to extract features from images . The image-context prediction from sentences in Chapter 4 is formulated as minimizing the cosine distance between the learned sentence and image representations in a training set of image–caption pairs.

Later we follow the formulations of Vendrov et al. (2016) and Faghri et al. (2017) and apply the sum-of-hinges loss in Chapter 5 and max-of-hinges ranking objective in Chapters 6 and 7 . These loss functions push relevant image–sentence pairs close, while contrastive pairs far from each other in a joint embedding space. Given a mini-batch of e.g. 100 samples, for each sample the contrastive pairs are generated by taking the wrong pairings from that batch leading to 99 contrastive pair per sample. The ranking losses are minimized in both image $\rightarrow$ sentence and sentence $\rightarrow$ image directions.

We empirically found both ranking losses to perform better than minimizing the cosine distance alone and max-of-hinges to perform consistently better in our experiments than sum-of-hinges. The use of these various objectives across chapters reflects the evolving common practices in the field at the time.

The representations learned by such image–sentence ranking models have been shown to improve performance when combined with skip-thought embeddings on a large number of semantic sentence classifications tasks compared to skip-thought only (Kiela et al., 2017). These findings were confirmed and improved upon using a self-attention mechanism on top of the RNN encoder (Yoo et al., 2017).

To investigate the difference between the representations learned by (text-only) language models and image–sentence ranking models in Chapter 4 we develop novel visualization and analysis methods. Expading on the findings of (Kiela et al., 2017) and (Yoo et al., 2017) in Chapter 5 we show that visually grounded learning through an image–sentence ranking objective leads to better translations in the visually descriptive domain. Furthermore, we show that learning multi-modal representations provides gains on top of learning from larger bilingual corpora. Finally in Chapters 6 and 7 we apply the image–sentence ranking framework (Vendrov et al., 2016; Faghri et al., 2017) in the multi-lingual setting and demonstrate that better visually grounded representations can be learned when training on multiple languages jointly.

2.6 Visually grounded multilingual representations

On top of the visual modality anchoring linguistic representations to perceptual reality it also provides a universal meaning representation bridging between languages. The intuition being that words or sentences with similar meanings appear within similar perceptual contexts independently of the language. First in Section 2.6.1 we discuss the multi-view representation learning perspective of considering images annotated with multiple descriptions in different languages as multiple views of the same underlying semantic concepts. Our aim in Chapters 6 and 7 is to learn better visually grounded sentence representations by learning from these multiple views simultaneously. Furthermore, in Chapter 5 we show that we can improve translation performance by learning better sentence representations through adding the visual modality as an additional view. Next we discuss how images can be used as pivots in practice for translation on word-level and on sentence-level in Section 2.6.2.

2.6.1 Multi-view representation learning perspective

Images and their descriptions in multiple languages can be taken as different views of the same underlying semantic concepts. From this multi-view perspective learning common representations of multiple languages and perceptual stimuli can potentially exploit the complementary information between views to learn better representations. Being able to extract a shared representation from only a single view also leads to practical applications such as cross-modal and cross-lingual retrieval or similarity calculation.

Multi-view learning traditionally considers paired samples where each pair of rows in two feature matrices are two measurements of the same underlying phenomena. In other words we assume two sets of variables representing the same data point. The two main multi-view learning paradigms put forward in recent literature are based on autoencoders and canonical correlation analysis (CCA) (Wang et al., 2015).

Ngiam et al. (2011) introduced the idea of multi-modal autoencoders to learn multi-modal joint representations of audio and video. Let they have a shared encoder neural network extracting features from both modalities and two modality specific decoders. Their approach learns shared representations such that one view can be reconstructed from another and the activations of the shared encoder learn a multi-modal joint space. Autoencoder approaches remain one of the standard family of models to study the learning of visual-linguistic multi-modal spaces (Silberer & Lapata, 2014; Silberer et al., 2017; Wang et al., 2018).

For the discussion of CCA based approaches let us consider the deep canonical correlation analysis (DCCA) put forward by Andrew et al. (2013). In this approach view specific networks are applied to extract non-linear features and the canonical correlation between these representations is maximized. This optimization process amounts to maximizing the correlation between the projections of the two data views subject to the constraint that the projected dimensions are uncorrelated.

The *workhorse* of the image–sentence ranking experiments in the foundational work on the topic from Hodosh et al. (2013) was in fact the Kernel CCA method (Akaho, 2006). Other CCA based approaches were also applied to the image–sentence data sets we consider in our work (Gong et al., 2014; Klein et al., 2015; Eisenschtat & Wolf, 2017).

A third direction that is also explored in the literature is combining the reconstruction objective of autoencoders with an additional correlation loss, but without the whitening constraints of CCA (AP et al., 2014; Chandar et al., 2016).

In the multi-lingual multi-modal setting Funaki & Nakayama (2015) apply Generalized CCA (Horst, 1961) – a variant of CCA generalized to multiple views as opposed to only two – to learn correlated representations of images and multiple languages. The deep partial canonical correlation analysis approach – a deep learning extension of the partial canonical correlation (Rao, 1969) – learns multilingual English-German sentence embeddings conditioned on the representation of the images they are aligned to (Rotman et al., 2018). They show that their model using the visual modality as an extra view finds better multilingual sentence and word representations as demonstrated by cross-lingual paraphrasing and word-similarity results.

The bridge correlational neural network approach (Rajendran et al., 2015) combines autoencoders and correlation objectives to learn common representations in a setting when the different views only need to be aligned with one pivot view. They preform image-sentence retrieval experiments in French or German where the image-caption data set is only available for English, however, there is parallel corpora between German or French and English. In other words English acts as a pivot. A similar combination of autencoder and correlational training was applied to *bridge* image–captioning (Saha et al., 2016).

Our formulation of the problem does not fall in the set of approaches generally considered as multi-view learning. However, it is similar to the CCA based techniques in that we train two subnetworks – one for the linguistic and another for the image modality – and do not rely on decoder networks to compute a reconstruction loss as in the autoencoder approaches. Another connections is that in Chapter 4 we minimize the cosine distance between the learned representations which is related to the CCA objective: when the feature matrices are not centered the CCA objective corresponds to maximizing the cosine similarity instead of the correlation.

One of the main benefits of the learning to rank approach we opted for is its flexibility: 1.) in Chapter 4 we train an image–sentence ranking model in a single language, 2.) we apply the same building blocks to train on multiple languages where the same images are shared between languages in Chapter 6, 3.) in Chapter 7 we explore the setup without such an alignment and finally 4.) in Chapter 5 we improve the automatic translation performance by adding the image–sentence ranking objective, incorporating an additional view to help us learn better sentence representations.

In our setup additional views are incorporated in the sentence representations by full parameter sharing through multi-task learning: given multiple image–caption data sets at each iteration we sample a batch from one of them and perform an update to the encoders using the ranking loss function. Gella et al. (2017) applies the image–sentence ranking framework in the multilingual setup considering images as a pivots bridging English and German and train a multilingual image–sentence ranking model. Their results suggest that the multilingual models outperform the monolingual ones on image-sentence ranking, however, they do not show consistent gains across languages and model settings. In Chapter 6 we implement a similar setup to Gella et al. (2017) and show that both English and German image-sentence ranking performance is reliably improved by bilingual joint training using our setup. We expand on the results further and provide evidence that more gains can be achieved by adding more views: on top of English and German, we add French and Czech captions and show the monolingual model is consistently outperformed by the bilingual and the latter by the multilingual. We apply the same approach to improve the performance on the lower resource French and Czech languages by adding the larger English and German image-caption sets; showing successful multilingual transfer in the vision-language domain.

2.6.2 Images as pivots for translation

On word level images have been used to link languages and induce bilingual lexicons without parallel corpora. The lack of bi-text in this setting has been traditionally solved by methods relying on textual features such as ortographic similarity (Haghighi et al., 2008) or similar diachronic distributional statistical trends between words (Schafer & Yarowsky, 2002). However, images tagged with various labels in a multitude of languages are available on the internet and a model of image-similarity can be used to exploit images as pivots to find translations between such tags. Alternatively representing words in multiple languages in a joint space with images – as discussed in Section 2.2 for the monolingual case – allows word translation to be simply performed through nearest neighbour search.

The first approach to construct a dictionary based on image similarity (Bergsma & Van Durme, 2011) used Google image search to find relevant images for the names of physical objects in multiple languages. These images are represented by BoVW vectors based on SIFT and color features. Given a source word and a list of possible translations Google is queried to find n images per word. For each image the feature vectors are extracted, which is followed by computing the similarity between the feature vectors of the images corresponding to the source word and all target words. Finally the word in the target vocabulary with the highest similarity is chosen.

This method is vastly improved upon by a later approach applying pre-trained convolutional neural networks to represent words in the visual space (Kiela et al., 2015). Their approach closely follows the monolingual visual-word representations of Kiela et al. (2014): each word is represented as the summary of CNN representations of images they co-occur with. Given a word in the source language the candidate translation is found simply by performing nearest neighbor search on the target vocabulary.

Exploring the limitations of image-pivoting for bilingual lexicon induction Hartmann & Søgaard (2017) present a negative result showing that such techniques scale poorly to non-noun words such as adjectives or verbs. However, combining image-pivot based bilingual word-representations with more traditional multi-lingual word-embedding techniques leads to superior performance compared to their uni-modal counterparts (Vulić et al., 2016).

The first large-scale vision based bilingual dictionary induction data set was put forward by Hewitt et al. (2018). They create a data set of 200K words and 100 images per each using Google Image

Search and perform experiments with 32 languages. They confirm the finding of Hartmann & Søgaard (2017) that image-pivoting is most effective for nouns, however, find that using their larger data set adjectives can also be translated reliably.

Images have also been used as pivots for translating full sentences. In automatic machine translation a pivot based approach is applied when there are parallel corpora available between language pairs $A \rightarrow C$ and $C \rightarrow B$, but there is no data for $A \rightarrow B$. The problem is solved by first translating A to C and then C to B. Image–pivoting refers to a setup where we assume the existence of a data set of images paired with words or sentences in different languages $A \leftrightarrow I_A$ and $B \leftrightarrow I_B$ and translation is done through the image space going through $A \rightarrow I$ to $I \rightarrow C$.

Nakayama & Nishida (2017) apply image-pivoting in such a zero-resource machine translation setting. The task is to translate from English to German without aligned parallel corpora, however, separate image–description data sets are available in both languages. They solve the problem by training two components: 1.) visual-sentence encoder that matches images and their descriptions, 2.) image-description generator maximizing the likelihood of gold standard captions given the images. At test time the visual-sentence encoder representation of the source sentence is fed to the image–description generator to produce the translation. Their results were improved later by modeling the image-pivot based zero-resource translation setup as a multi-agent communication game between encoder and decoder (Chen et al., 2018; Lee et al., 2017b).

2.7 Interpreting continuous representations

The linguistic representations learned through neural architectures are notoriously opaque. Contrary to count-based methods the features extracted by deep networks from text input appear as arbitrary dense vectors to the human eye. In experiments with grounded learning throughout this thesis we find that visual grounding improves translation peformance and that multilingual representations outperform monolingual ones in image–sentence ranking. But where do these improvement come from? What are the linguistic regularities represented in the recurrent states that lead to the final performance metrics? Did the model learn to exploit trivial artifacts in the training data or did it learn to meaningfully generalize? What characterizes the individual features recurrent networks extract from the input sentences?

The main topic of this thesis is learning visually grounded representations for linguistic units. For a complete picture, however, it is crucial to assess the difference in the representations between textual only and multimodal representations not only from a quantitative point of view, but also from a qualitative linguistics angle. This is what Chapter 4 is dedicated to.

Developing techniques for interpreting machine learning models have multiple goals. From the practical point of view as learning algorithms make their way into critical applications such as medicine humans and machines need to be able to co-operate to avoid catastrophic outcomes (Caruana et al., 2015). As such there is a growing interest in deriving methods to *explain* the decision of such architectures.

One of the approaches is to assign a real-valued ”relevance” score to each unit in the input signal, signifying how much impact it had on the final prediction of the model. One of the first paradigms in generating such relevance scores is gradient based methods: they take the gradient of the output of the network with respect to the input (Simonyan et al., 2014). Deep neural models of language tasks learn distributed representations of input symbols and as such further operations have be applied to reduce the resulting gradient vectors to single scalars e.g.: using ℓ_2 norm (Bansal et al., 2016).

Another prominent and well studied approach still based on gradient information is layerwise relevance propagation (LRP) (Bach et al., 2015). The output of the final layer is written as the sum of the relevance-scores from the input and similarly to the back-propagation algorithm the relevance of each neuron recursively depends on the lower-layer all the way down to the input signal. Different versions of LRP run the backward pass with different rules taking as input gradient information and activation values. It was later theoretically analyzed and generalized into the deep Taylor decomposition method (Binder et al., 2016).

Perturbation based methods are gradient free and are algorithmically very simple: they involve generating pseudo-samples according to some procedure and measuring how the models' response changes between each pseudo-sample and the original. LIME (Ribeiro et al., 2016) and its NLP specific LIMSSE extension (Poerner et al., 2018b) perturbs the input creating a local neighborhood around it and fits interpretable linear models to explain the predictions of any complex black box classifier. Even simpler perturbation based techniques measure the difference between the original input and the various perturbed candidates such as the erasure (Li et al., 2016b) and our omission (Kádár et al., 2017) method in Chapter 4.

Apart from practical considerations of model interpretation training complex and opaque models from close to raw input can help us discover interesting patterns in the input data that are crucial in solving the task. Deep neural networks learn to solve tasks from close to raw input, similar to what humans receive. As such the regularities they learn can also shed light on the patterns humans might extract from data to cope with certain tasks. Recent methodology in probing the learned representations of LSTM language models, in fact, resemble psycholinguistic studies. A number of experiments using the agreement prediction paradigm (Bock & Miller, 1991) suggest that LSTM language models successfully learn syntactic regularities as opposed to memorizing surface patterns (Linzen et al., 2016; Enguehard et al., 2017; Bernardy & Lappin, 2017; Gulordava et al., 2018).

For our purpuses in Chapter 4 we develop our explanation method to shed light on the linguistic characteristics of the input grounded learning models learn in contrast to text-only language models.

3

Learning word meanings from images of natural scenes

Abstract Children early on face the challenge of learning the meaning of words from noisy and ambiguous contexts. Utterances that guide their learning are emitted in complex scenes rendering the mapping between visual and linguistic cues difficult. A key challenge in computational modeling of the acquisition of word meanings is to provide representations of scenes that contain sources of information and statistical properties similar in complexity to natural data. We propose a novel computational model of cross-situational word learning that takes images of natural scenes paired with their descriptions as input and incrementally learns probabilistic associations between words and image features. Through a set of experiments we show that the model learns meaning representations that correlate with

human similarity judgments, and that given an image of a scene it produces words conceptually related to the image.

This chapter is based on Kádár, Á., Alishahi, A., & Chrupala, G. (2015). Learning word meanings from images of natural scenes. *Traitement Automatique des Langues.*

3.1 Introduction

Children learn most of their vocabulary from hearing words in noisy and ambiguous contexts, where there are often many possible mappings between words and concepts. They attend to the visual environment to establish such mappings, but given that the visual context is often very rich and dynamic, elaborate cognitive processes are required for successful word learning from observation. Consider a language learner hearing the utterance *"the gull took my sandwich"* while watching a bird stealing someone's food. For the word *gull*, such information suggests potential mappings to the bird, the person, the action, or any other part of the observed scene. Further exposure to usages of this word and relying on structural cues from the sentence structure is necessary to narrow down the range of its possible meanings.

3.1.1 Cross-situational learning

A well-established account of word learning from perceptual context is called cross-situational learning, a bottom-up strategy in which the learner draws on the patterns of co-occurrence between a word and its referent across situations in order to reduce the number of possible mappings (Quine, 1960; Carey, 1978; Pinker, 1989). Various experimental studies have shown that both children and adults use cross-situational evidence for learning new words (Yu & Smith, 2007; Smith & Yu, 2008; Vouloumanos, 2008; Vouloumanos & Werker, 2009).

Cognitive word learning models have been extensively used to study how children learn robust word-meaning associations despite the high rate of noise and ambiguity in the input they receive. Most of the existing models are either simple associative networks that gradually learn to predict a word form based on a set of semantic features (Li et al., 2004; Regier, 2005), or are rule-based or probabilistic implementations which use statistical regularities observed in the input to detect associations between linguistic labels and visual features or concepts (Siskind, 1996; Frank et al., 2007; Yu, 2008; Fazly et al., 2010b). These models all implement different (implicit or explicit) variations of the cross-situational learning mechanism, and demonstrate its efficiency in learning robust mappings between words and meaning representations in presence of noise and perceptual ambiguity.

However, a main obstacle to developing realistic models of child word learning is lack of resources for reconstructing perceptual context. The input to a usage-based cognitive model must contain the same information components and statistical properties as naturally-occurring data children are exposed to. A large collection of transcriptions and video recordings of child-adult interactions has been accumulated over the years (MacWhinney, 2014), but few of these resources provide adequate semantic annotations that can be automatically used by a computational model. As a result, the existing models of word learning have relied on artificially generated input (Siskind, 1996). The meaning of each word is represented as a symbol or a set of semantic features that are selected arbitrarily or from lexical resources such as WordNet (Fellbaum, 1998), and the visual context is built by sampling these symbols. Some models add additional noise to data by randomly adding or removing meaning symbols to/from the perceptual input (Fazly et al., 2010b).

Carefully constructed artificial input is useful in testing the plausibility of a learning mechanism, but comparisons with manually annotated visual scenes show that these artificially generated data sets often do not show the same level of complexity and ambiguity as naturally occurring perceptual context (Matusevych et al., 2013; Beekhuizen et al., 2013).

3.1.2 Learning meanings from images

To investigate the plausibility of cross-situational learning in a more naturalistic setting, we propose to use visual features from collections of images and their captions as input to a word learning model. In the domain of human-computer interaction (HCI) and robotics, a number of models have investigated the acquisition of terminology for visual concepts such as color and shape from visual data. Such concepts are learned based on communication with human users (Fleischman & Roy, 2005; Skocaj et al., 2011). Because of the HCI setting, they need to make simplifying assumptions about the level of ambiguity and uncertainty about the visual context.

The input data we exploit in this research has been used for much recent work in NLP and machine learning whose goal is to develop multimodal systems for practical tasks such as automatic image captioning. This is a fast-growing field and a detailed discussion of it is beyond the scope of this paper. Recent systems include Karpathy et al. (2014), Mao et al. (2014b), Kiros et al. (2014), Donahue et al. (2014a), Vinyals et al. (2015b), Venugopalan et al. (2014), Chen & Zitnick (2014), Fang et al. (2014). The majority of these approaches rely on convolutional neural networks for deriving representations of visual input, and then generate the captions using various versions of recurrent neural network language models conditioned on image representations. For example Vinyals et al. (2015b) use the deep convolutional neural network of Szegedy et al. (2014) trained on ImageNet to encode the image into a vector. This representation is then decoded into a sentence using a Long Short-Term Memory recurrent neural network (Hochreiter & Schmidhuber, 1997). Words are represented by embedding them into a multidimensional space where similar words are close to each other. The parameters of this embedding are trainable together with the rest of the model, and are analogous to the vector representations learned by the model proposed in this paper. The authors show some example embeddings but do not analyze or evaluate them quantitatively, as their main focus is on the captioning performance.

Perhaps the approach most similar to ours is the model of Bruni et al. (2014). In their work, they train multimodal distributional semantics models on both textual information and bag-of-visual-words features extracted from captioned images. They use the induced semantic vectors for simulating word similarity judgments by humans, and show that a combination of text and image-based vectors can replicate human judgments better than using uni-modal vectors. This is a batch model and is not meant to simulate human word learning from noisy context, but their evaluation scheme is suitable for our purposes.

Lazaridou et al. (2015) propose a multimodal model which learns word representations from both word co-occurrences and from visual features of images associated with words. Their input data consists of a large corpus of text (without visual information) and additionally of the ImageNet dataset (Deng et al., 2009) where images are labeled with WordNet synsets.[1] Thus, strictly speaking their model does not implement cross-situational learning because a subset of words is unambiguously associated with certain images.

3.1.3 Our study

In this paper we investigate the plausibility of cross-situational learning of word meanings in a more naturalistic setting. Our goal is to simulate this mechanism under the same constraints that humans face when learning a language, most importantly by learning in a piecemeal and incremental fashion, and facing noise and ambiguity in their perceptual environment. (We do not investigate the role of sentence structure on word learning in this study, but we discuss this issue in Section 3.5).

For simulation of the visual context we use two collections of images of natural scenes, Flickr8K (F8k) (Rashtchian et al., 2010) and Flickr30K (F30k) (Young et al., 2014), where each image is associated with several captions describing the scene. We extract visual features from the images and learn to associate words with probability distributions over these features. This has the advantage that we do not need to simulate ambiguity or referential uncertainty – the data has these characteristics naturally.

The challenge is that, unlike in much previous work on cross-situational learning of word meanings, we do not know the ground-truth word meanings, and thus cannot directly measure the progress and effectiveness of learning. Instead, we use indirect measures such as (i) the correlation of the similarity of learned word meanings to

[1] The synsets of WordNet are groups of synonyms that represent an abstract concept.

word similarities as judged by humans, and (ii) the accuracy of producing words in response to an image. Our results show that from pairings of scenes and descriptions it is feasible to learn meaning representations that approximate human similarity judgments. Furthermore, we show that our model is able to name image descriptors considerably better than the frequency baseline and names a large variety of these target concepts. In addition we present a pilot experiment for word production using the ImageNet data set and qualitatively show that our model names words that are conceptually related to the images.

3.2 Word learning model

Latest existing cross-situational models formulate word learning as a translation problem, where the learner must decide which words in an utterance correspond to which symbols (or potential referents) in the perceptual context (Yu & Ballard, 2007; Fazly et al., 2010b). For each new utterance paired with a symbolic representation of the visual scene, first the model decides which word is *aligned* with which symbol based on previous associations between the two. Next, it uses the estimated alignments to update the meaning representation associated with each word.

We introduce a novel computational model for cross-situational word learning from captioned images. We reformulate the problem of learning the meaning of words as a translation problem between words and a *continuous* representation of the scene; that is, the visual features extracted from the image. In this setting, the model learns word representations by taking images and their descriptions one pair at a time. To learn correspondences between English words and image features, we borrow and adapt the translation-table estimation component of the IBM Model 1 (Brown et al., 1993). The learning results in a translation table between words and image-features, i.e. a list of probabilities of image-features given a word.

3.2.1 Visual input

The features of the images are extracted by training a 16-layer convolutional neural network (CNN) (Simonyan & Zisserman, 2015) on an object recognition task.[2] The network is trained to discriminate among 1,000 different object labels on the ImageNet dataset (Deng et al., 2009). The last layer of the CNN before the classification layer contains high level visual features of the images, invariant to particulars such as position, orientation or size. We use the activation vector from this layer as a representation of the visual scene described in the corresponding caption. Each caption is paired with such a 4,096-dimensional vector and used as input to a cross-situational word learner. Figure 3.1 shows three sample images from the F8k dataset most closely aligned with a particular dimension, as measured by the cosine similarity between the image and a unit vector parallel to the dimension axis. For example, dimension 1,000 seems to be related to water, 2,000 to dogs or perhaps grass, and 3,000 to children.

[2]We used the F8k and F30k features available at `http://cs.stanford.edu/people/karpathy/deepimagesent/` and the data handling utilities from `https://github.com/karpathy/neuraltalk` for our experiments. The pre-trained CNN can be used through the Caffe framework (Jia et al., 2014) and is available at the ModelZoo `https://github.com/BVLC/caffe/wiki/Model-Zoo`.

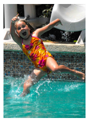
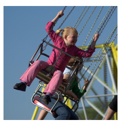
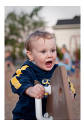
Dimension	Top 3 images		
1,000			
2,000			
3,000			

Figure 3.1: *Dimensions with three most closely aligned images from F8k.*

3.2.2 Learning algorithm

We adapt the IBM model 1 estimation algorithm in the following ways[3]: (i) like Fazly et al. (2010b) we run it in an online fashion, and (ii) instead of two sequences of words, our input consists of one sequence of words on one side, and a vector of real values representing the image on the other side. The dimensions are indexes into the visual feature "vocabulary", while the values are interpreted as weights of these "vocabulary items". In order to get an intuitive understanding of how the model treats the values in the feature vector, we could informally liken these weights to word counts. As an example consider the following input with a sentence and a vector of 5 dimensions (i.e. 5 features):

- The blue sky
- $(2, 0, 2, 1, 0)$

Our model treats this equivalently to the following input, with the values of the dimensions converted to "feature occurrences" of each feature f_n.

- The blue sky
- f_1 f_1 f_3 f_3 f_4

The actual values in the image vectors are always non-negative, since they come from a rectified linear (ReLu) activation. However, they can be fractional, and thus strictly speaking cannot be

[3]The source code for our model is available at https://github.com/kadarakos/IBMVisual.

literal counts. We simply treat them as generalized, fractional feature "counts". The end result is that given the lists of words in the image descriptions and the corresponding image vectors the model learns a probability distribution $t(f|w)$ over feature-vector indexes f for every word w in the descriptions.

Algorithm 1 Sentence-vector alignment model (Visual)

Input: visual feature vectors paired with sentences $((V_1, S_1), \ldots, (V_N, S_N))$
Output: translation table $t(f|w)$
$D \leftarrow$ dimensionality of feature vectors
$\epsilon \leftarrow 1$ ▷ Smoothing coefficient
$a[f, w] \leftarrow 0, \ \forall f, w$ ▷ Initialize count tables
$a[\cdot, w] \leftarrow 0, \ \forall w$
$t(f|w) \leftarrow \frac{1}{D}$ ▷ Translation probability $t(f|w)$
for each input pair (vector V, sentence S) **do**
    for each feature index $f \in \{1, \ldots, D\}$ **do**
        $Z_f \leftarrow \sum_{w \in S} t(f|w)$ ▷ Normalization constant Z_f
        for each word w in sentence S **do**
            $c \leftarrow \frac{1}{Z_f} \times V[f] \times t(f|w)$ ▷ Expected count c
            $a[f, w] \leftarrow a[f, w] + c$
            $a[\cdot, w] \leftarrow a[\cdot, w] + c$ ▷ Updates to count tables
            $t(f|w) \leftarrow \frac{a[f,w]+\epsilon}{a[\cdot,w]+\epsilon D}$ ▷ Recompute translation probabilities
        end for
    end for
end for

This is our sentence-vector alignment model, Visual. In the interest of cognitive plausibility, we train it using a single-pass, online algorithm. Algorithm 1 shows the pseudo-code. Our input is a sequence of pairs of D-dimensional feature vectors and sentences, and the output is a translation table $t(f|w)$. We maintain two count tables of expected counts $a[f, w]$ and $a[\cdot, w]$ which are used to incrementally recompute the translation probabilities $t(f|w)$. The initial translation probabilities are uniform (line 7). In lines 12-14 the count tables are updated, based on translation probabilities weighted by the feature value $V[f]$, and normalized over all the words in the sentence. In line 15 the translation table is in turn updated.

3.2.3 Baseline models

To asses the quality of the meaning representations learned by our sentence-vector alignment model VISUAL, we compare its performance in a set of tasks to the following baselines:

- MONOLING: instead of aligning each sentence with its corresponding visual vector, this variation aligns two copies of each sentence with each other, and thus learns word representations based on word-word co-occurrences[4].

- WORD2VEC: for comparison we also report results with the skip-gram embedding model, also known as WORD2VEC which builds word representations based on word-word co-occurrences as well (Mikolov et al., 2013a,b). WORD2VEC learns a vector representation (embedding) of a word which maximizes performance on predicting words in a small window around it.

[4]This model does not estimate probabilities of translation of a word to itself, that is probabilities of the form $t(w|w)$.

3.3 Experiments

3.3.1 Image datasets

We use image-caption datasets for our experiments. F8k (Rashtchian et al., 2010) consists of 8000 images and five captions for each image. F30k (Young et al., 2014) extends the F8k and contains 31,783 images with five captions each summing up to 158,915 sentences. For both data sets we use the splits from Karpathy et al. (2014), leaving out 1000 images for validation and 1000 for testing from each set. Table 3.1 summarizes the statistics of the Flickr image-caption datasets.

	F8k	F30k
Train images	6,000	29,780
Validation images	1,000	1,000
Test images	1,000	1,000
Image in total	8,000	31,780
Captions per image	5	5
Captions in total	40,000	158,900

Table 3.1: *Flickr image caption datasets.*

For the Single-concept image descriptions experiments reported in section 3.3.4, we also use the ILSVRC2012 subset of ImageNet (Russakovsky et al., 2015), a widely-used data set in the computer vision community. It is an image database that annotates the WordNet noun synset hierarchy with images. It contains 500 images per synset on average.

3.3.2 Word similarity experiments

A common evaluation task for assessing the quality of learned semantic vectors for words is measuring word similarity. A number of experiments have elicited human ratings on the similarity and/or relatedness of a list of word pairs. For instance one of the data sets we used was the SimLex999 data set, which contains similarity judgments for 666 noun pairs (organ-liver 6.15), 222 verb pairs (occur-happen 1.38) and 111 adjective pairs (nice-cruel 0.67) elicited by 500 participants recruited from Mechanical Turk. These types of data sets are commonly used as benchmarks for models of distributional semantics, where the learned representations are expected to show a significant positive correlation with human similarity judgments on a large number of word pairs.

We selected a subset of the existing benchmarks according to the size of their word pairs that overlap with our restricted vocabulary. We ran a statistical power analysis test to estimate the minimum number of required word pairs needed in our experiments. The projected sample size was $N = 210$ with $p = .05$, effect-size $r = .2$ and $power = 0.9$. Thus some of the well-known benchmarks were excluded due to their small sample size after we excluded words not present in our datasets.[5]

The four standard benchmarks that contain the minimum number of word pairs are: the full WS-353 (Finkelstein et al., 2001), MTurk-771 (Radinsky et al., 2011), MEN (Bruni et al., 2014) and SimLex999 (Hill et al., 2015). Note that the MTurk dataset only contains similarity judgments for nouns. Also, a portion of the full WordSim-353

[5] These include RG-65 (Rubenstein & Goodenough, 1965), MC-30 (Miller & Charles, 1991) and YP-130 (Yang & Powers, 2006).

dataset reports relatedness ratings instead of word similarity.

3.3.3 Effect of concreteness on similarity judgments

The word similarity judgments provide a macro evaluation about the overall quality of the learned word representations. For more fine-grained analysis we turn to the dichotomy of concrete (e.g. *chair, car*) versus abstract (e.g. *love, sorrow*) nouns. Evidence presented by Recchia & Jones (2012) shows that in naming and lexical decision tasks the early activation of abstract concepts is facilitated by rich linguistic contexts, while physical contexts promote the activation of concrete concepts. Based on these recent findings, Bruni et al. (2014) suggest that in case of computational models *concrete* words (such as names for physical objects and visual properties) are easier to learn from perceptual/visual input and *abstract* words are mainly learned based on their co-occurrence with other words in text. Following Bruni et al. (2014), but using novel methodology, we also test this idea and examine whether more concrete words benefit more from visual features compared with less concrete ones.

In their work Bruni et al. (2014) use the automatic method from Turney et al. (2011) to assign concreteness values to words and split the MEN corpus in concrete and abstract chunks. From their experiments they draw the conclusion that visual information boosts their models' performance on concrete nouns. However, whereas the multi-modal embeddings of Bruni et al. (2014) are trained using an unbalanced corpus of large quantities of textual information and far poorer visual stimuli, our visual embeddings are learned on a parallel corpus of sentences paired with images. To our purposes, this balance in the sources of information is critical as we aim at modeling word learning in humans. As a consequence of this setting we rather hypothesized that solely relying on visual features would result in better performance on more concrete words than on abstract ones and conversely, learning language solely from textual features would lead to higher correlations on the more abstract portion of the vocabulary.

To test this hypothesis, MEN, MTurk and Simlex999 datasets were split in two halves based on concreteness score of the word pairs. The "abstract" and "concrete" subclasses for each data set are obtained by ordering the pairs according to their concreteness and then partition the ordered tuples in halves. We defined the concreteness of a word pair as the product of the concreteness scores of the two words. The scores are taken from the University of South Florida Free Association Norms dataset (Nelson et al., 1998). Table 3.2 provides an overview of the benchmarks we use in this study. Column "Concreteness" shows the average concreteness scores of all words pairs per data set, while columns "Concrete" and "Abstract" contain the average concreteness of the concrete and abstract halves of the word-pairs respectively.

3.3.4 Word production

Learning multi-modal word representations gives us the advantage of replicating real-life tasks such as naming visual entities. In this study, we simulate a word production task as follows: given an image from the test set, we rank all words in our vocabulary according to their cosine similarity to the visual vector representing the image. We evaluate these ranked lists in two different ways.

	#Pairs			Concreteness		
	Total	F8k	F30k	Full set	Concrete	Abstract
WS353	353	104	232	25.09	35.44	16.22
SimLex999	999	412	733	23.86	35.72	11.99
MEN	3000	2069	2839	29.77	36.28	23.26
MTurk771	771	295	594	25.89	34.02	16.16

Table 3.2: *Summary of the word-similarity benchmarks, showing the number of word pairs in the benchmarks and the size of their overlap with the F8k and F30k data sets. The table also reports the average concreteness of the whole, concrete and abstract portions of the benchmarks.*

3.3.4.1 Multi-word image descriptions.

We use images from the test portion of the F8k and F30k datasets as benchmarks. These images are each labeled with up to five captions, or multi-word descriptions of the content of the image. To evaluate the performance of our model in producing words for each image, we construct the target description of an image as the union of the words in all its captions (with stop-words[6] removed). We compare this set with the top N words in our predicted ranked word list. As a baseline for this experiment we implemented a simple frequency baseline FREQ, which for every image retrieves the top N most frequent words. The second model COSINE uses our VISUAL word-embeddings and ranks the words based on their cosine similarity to the given image. The final model PRIOR implements a probabilistic interpretation of the task

$$P(w_i|i_j) \propto P(i_j|w_i) \times P(w_i), \tag{3.1}$$

[6]Function words such as *the, is, at, what, there*; we used the stop-word list from the Python library NLTK.

where w_i is a word from the vocabulary of the captions and i_j is an image from the collections of images I. The probability of an image given a word is defined as

$$P(i_j|w_i) = \frac{\text{cosine}(i_j, w_i)}{\sum_{k=1}^{|I|} \text{cosine}(i_k, w_i),} \tag{3.2}$$

where $\text{cosine}(i_j, w_i)$ is the cosine between the vectorial representation of i_j and the VISUAL word-embedding w_i. Since in any natural language corpus the distribution of word frequencies is expected to be very heavy tailed, in the model PRIOR, rather than using maximum likelihood estimation, we reduce the importance of the differences in word-frequencies and smooth the prior probability $P(w_i)$ as described by equation 3.3, where N is the number of words in the vocabulary.

$$P(w_i) = \frac{\log(count(w_i))}{\sum_{j=1}^{N} \log(count(w_j))} \tag{3.3}$$

As a measure of performance, we report Precision at 5 (P@5) between the ranked word list and the target descriptions; i.e., proportion of correct target words among the top 5 predicted ranked words. Figure 3.2 shows an example of an image and its multi-word captions in the validation portion of the F30k dataset.

3.3.4.2 Single-concept image descriptions

Even though we use separate portions of F8k and F30k for training and testing, these subsets are still very similar. To test how general the VISUAL word representations are, we use images from the ILSVRC2012 subset of ImageNet (Russakovsky et al., 2015) as benchmark. The major difference between these images and the ones from

A boy in a blue shirt and white helmet is riding a white bike
A boy in blue is riding his bike in a skate park
A boy on a BMX bike
A cyclist riding on their front wheel on the asphalt
The man is on a black and white bike

Descriptors: blue boy skate shirt asphalt helmet
park cyclist black bike wheel front
riding white bmx man

Predicted: bike bicycle riding man biker
Overlap: riding bike man
P@5: 0.6

Figure 3.2: *Multiword image description example. Below the image are the 5 captions describing the image, the union of words that we take as targets, the top 5 predicted and the list of correct words and the P@5 score for the given test case.*

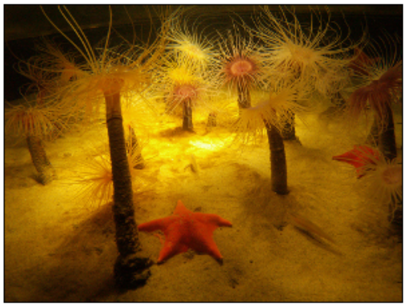


Figure 3.3: *Example of the Single-concept image description task from the validation portion of the ILSVRC2012 subset of ImageNet. The terms "sea anemone" and "anemone" are unknown to* VISUAL *and "animal" is the first word among it's hypernyms that appear in the vocabulary of F30k.*

F8k and F30k datasets is that labels of the images in ImageNet are synsets from WordNet, which identify a single concept present in the image instead of providing a natural descriptions of its full content. Providing a quantitative evaluation in this case is not straightforward for two main reasons. First, the vocabulary of our model is restricted and the synsets in the ImageNet dataset are quite varied. Second, the synset labels can be very precise, much more so than the descriptions provided in the captions that we use as our training data.

To attempt to solve the vocabulary mismatch problem, we use synset hypernyms from WordNet as substitute target descriptors. If none of the lemmas in the target synset are in the vocabulary of the model, the lemmas in the hypernym synset are taken as new targets, until we reach the root of the taxonomy. However, we find that in a large number of cases these hypernyms are unrealistically general given the image. Figure 3.3 illustrates these issues.

3.4 Results

We evaluate our model on two main tasks: simulating human judgments of word similarity[7] and producing labels for images. For all performance measures in this sections (Spearman's ρ, P@5), we estimated the confidence intervals using the Bias-corrected Accelerated bootstrapping method[8] (Efron, 1982).

3.4.1 Word similarity

We simulate the word similarity judgment task using the induced word vectors by three models: VISUAL, MONOLING, and WORD2VEC. All models were trained on the tokenized training portion of the F30k data set. While VISUAL is presented with pairs of captions and the 4,096 dimensional image-vectors, MONOLING and WORD2VEC[9] are trained solely on the sentences in the captions. The smoothing coefficient $\epsilon = 1.0$ was used for VISUAL and MONOLING. The WORD2VEC model was run for one iteration with default parameters, except for the minimum word count (as our models also consider each word in each sentence): feature-vector-size=100, alpha=0.025, window-size=5, min-count=5, downsampling=False, alpha=0.0001, model=skip-gram, hierarchical-sampling=True, negative-sampling=False.

Figure 3.4 illustrates the correlation of the similarity judgments by the three models with those of humans on four datasets. Table 3.3

[7]We made available the source code used for running word similarity/relatedness experiments on https://bitbucket.org/kadar_akos/wordsims.

[8] Provided by the scikits-bootstrap Python package https://github.com/cgevans/scikits-bootstrap.

[9]We used the Word2Vec implementation from the gensim Python package available at https://radimrehurek.com/gensim/models/word2vec.html.

shows the results in full detail: it reports the Spearman rank-order correlation coefficient between the human similarity judgments and the pairwise cosine similarities of the word vectors per data set along with the confidence intervals estimated by using bootstrap (the correlation values marked by a * were significant at level $p < 0.05$).

Although VISUAL achieves a higher correlation than the other two models on all datasets, the overlapping confidence intervals suggest that, in this particular setting, both VISUAL and WORD2VEC perform very similarly in approximating human similarity judgments. This result is particularly interesting as these models exploit different sources of information: The input to WORD2VEC is text only (i.e., the set of captions) and it learns from word-word co-occurrences, while VISUAL takes pairs of image vectors and sentences as input, and thus learns from word-scene co-occurrences.

The significant medium-sized correlation ($p < .001$, $\rho = 0.47$ 95% CI [0.44, 0.50]) with reasonably narrow confidence intervals on the large number of samples, $N = 2,839$, of the MEN data set supports the hypothesis that the similarities between the meaning representations learned by VISUAL mirror the distance between word pairs as estimated by humans. This result suggests that it is feasible to learn word meanings from co-occurrences of sentences with noisy visual scenes. However, on all other data sets, the effect sizes for all models are small and their performances vary considerably given different subsamples of the benchmarks.

3.4.1.1 Concreteness

Based on the previous findings of Bruni et al. (2014), we expected that models relying on perceptual cues perform better on the concrete

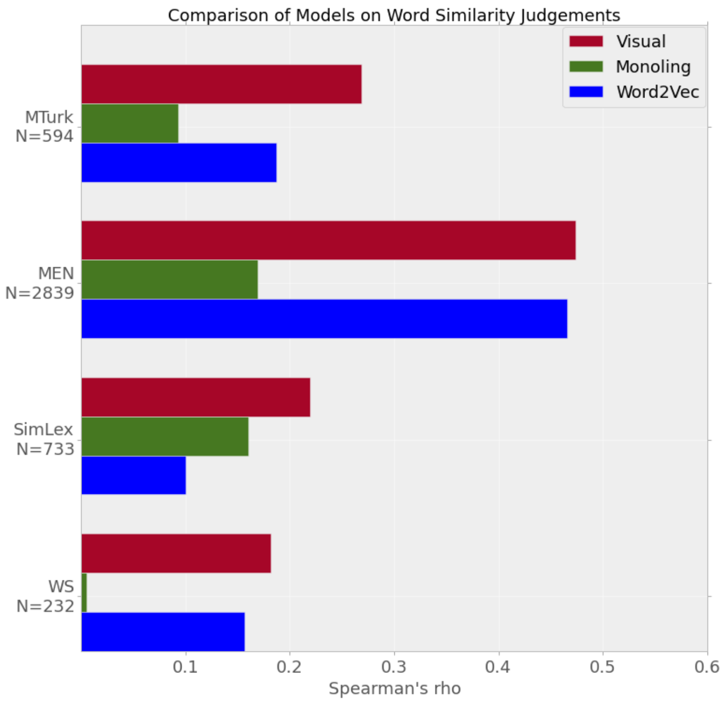


Figure 3.4: *Comparison of models on approximating word similarity judgments. The length of the bars indicate the size of the correlation measured by Spearman's ρ, longer bars indicate better similarity between the models' predictions and the human data. The labels on the y-axis contain the names of the data sets and indicate the number of overlapping word pairs with the vocabulary of the F30k data set. All models were trained on the training portion of the F30k data set.*

	WS	SimLex	MEN	MTurk
VISUAL	**0.18***	**0.22***	**0.47***	**0.27***
	CI[0.05, 0.32]	CI[0.15, 0.29]	CI[0.44, 0.50]	CI[0.19, 0.34]
MONOLING	0.08	0.18*	0.23*	0.17*
	CI[-0.06, 0.21]	CI[0.11, 0.25]	CI[0.19, 0.26]	CI[0.04, 0.19]
WORD2VEC	0.16*	0.10*	0.47*	0.19*
	CI[0.02, 0.28]	CI[0.02, 0.17]	CI[0.43 0.49]	CI[0.11, 0.26]

Table 3.3: *Word similarity correlations with human judgments measured by Spearman's ρ. Models were trained on the training portion of the F30k data set. The * next to the values marks the significance of the correlation at level $p < 0.05$. The confidence intervals for the correlation are estimated using bootstrap.*

portion of the word-pairs in the word-similarity benchmarks. Furthermore, we expected approximating human word similarity judgments on concrete word-pairs to be generally easier. As discussed in section 3.3.3, we split the data sets into *abstract* and *concrete* halves and ran the word similarity experiments on the resulting portions of the word-pairs for comparison. Table 3.4 only reports the results on MEN and Simlex999 as these were the only benchmarks that had at least 200 word-pairs after partitioning. Table 3.2 summarizes the average concreteness of the different portions of the data sets.

On all data sets, VISUAL seems to perform considerably better on the concrete word-pairs then on abstract ones. On the abstract half of the MEN data set, the performance of VISUAL is $\rho = 0.35$, 95% $CI[0.29, 0.41]$, while it is $\rho = 0.56$, 95% $CI[0.49, 0.59]$ on the concrete portion. The non-overlapping confidence intervals support the hypothesis that VISUAL does significantly better on the concrete word pairs. This pattern, however, is not observed for WORD2VEC as there is no significant difference in its performance given the different concreteness levels of the word pairs. Splitting the word pairs in two

	MEN		SimLex	
	Abstract	Concrete	Abstract	Concrete
Visual	0.35* CI[0.29, 0.41]	0.55* CI[0.49, 0.59]	0.16* CI[0.04, 0.25]	0.39* CI[0.28, 0.47]
Word2Vec	0.48 CI[0.43, 0.53]	0.45 CI[0.39, 0.50]	0.14 CI[0.02, 0.25]	0.18 CI[0.07, 0.29]

Table 3.4: *The table reports the Spearman rank-order correlation coefficient on the abstract and concrete portions of the data sets separately as well as the confidence intervals around the effect-sizes estimated by using bootstrap. The * next to the values indicates significance at level* $p < 0.05$.

groups based on their concreteness scores reveals that performance of Visual is affected by concreteness and that it only performs better than Word2Vec on the more concrete word pairs. Another pattern that the analysis reveals is that the average concreteness of the data sets is reflected in the performance of the models: for both Visual and Word2Vec the rank of their performance follows the rank of concreteness of the benchmarks.

3.4.2 Word production

In this set of experiments, we evaluate the word meaning vectors learned by Visual by simulating the task of word production for an image, as described in Section 3.3.4. These experiments can be viewed as computational simulations of a language task where human subjects associate words to given images. Words were ranked according to their cosine similarity to a given image vector. The Visual model was trained on the training portion of the F8k and F30k data sets. We report results on two variations of the word production task: multi-word image descriptors, and single-concept image descriptors.

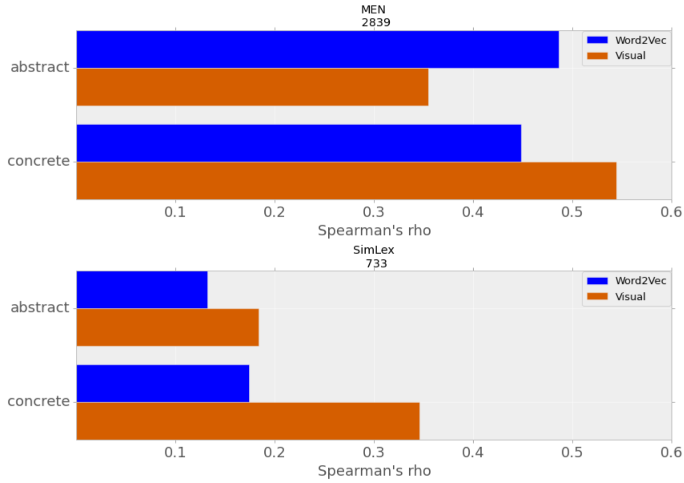


Figure 3.5: *Models' performance on word similarity judgments as a function of the concreteness of the word pairs.*

3.4.2.1 Multi-word image descriptors

The objective of the model in this experiment is to rank only words in the top N that occur in the set containing all words from the concatenation of the 5 captions of a given image with stop-words removed. The ranking models used for these experiments (FREQ, COSINE, and PRIOR) are described in section 3.3.4. Table 3.5 reports the results of the experiments on the respective test portions of the F8k and F30k datasets as estimated by P@5. We estimated the variability of the models' performance by calculating these measures per sample and estimating the confidence intervals around the means using bootstrap.

On these particular data sets the naive frequency baseline can perform particularly well: by only retrieving the sequence *wearing, woman, people, shirt, blue* the ranking model FREQ scores a P@5=.27

on F30k. Incorporating both the meaning representations learned by VISUAL and the prior probabilities of the words, the non-overlapping confidence intervals suggest that PRIOR significantly outperforms FREQ — P@5=0.42, 95% $CI[0.41, 0.44]$.

In addition to P@5, we also report the number of word types that were retrieved correctly given the images (column Words@5 on table 3.5). This measure was inspired by the observation that by focusing only on the precision scores it seems like incorporating visual information rather than just using raw word-frequency statistics provides a significant, but small advantage. However, taking into consideration that PRIOR retrieves 178 word types correctly suggests that it can retrieve less generic words that are especially descriptive of fewer scenes.

To have a more intuitive grasp on the performance of PRIOR, it is worth taking also into consideration the distribution of P@5 scores over the test cases. When trained and tested on F30k in most cases (34%), PRIOR retrieves two words correctly in the top 5 and in 23% and 25% of the cases it retrieves one and three respectively. In only 6% of the time $P@5 = 0$, which means that it is very unlikely that PRIOR named unrelated concepts given an image. These results suggest that VISUAL learns word meanings that allow for labeling unseen images with reasonable accuracy using a large variety of words.

3.4.2.2 Single-concept image descriptors

The motivation for this experiment was to assess the generalizability of the word-representations learned by VISUAL. Similarly to the previous task, the goal here is to associate words to a given image, but in this case the images are drawn from the validation set of

	F8k		F30k	
	P@5	Words@5	P@5	Words@5
FREQ	0.20	5	0.27	5
	CI[0.19, 0.21]		CI[0.26, 0.29]	
COSINE	0.16	310	0.14	371
	CI[0.15, 0.17]		CI[0.13, 0.15]	
PRIOR	**0.44**	135	**0.42**	178
	CI[0.42, 0.45]		CI[0.41, 0.44]	

Table 3.5: *Results for the multi-word image descriptors experiments reported on the test sets of F8k and F30k. Words@5 the number of correctly retrieved word types in the top 5. The confidence intervals below P@5 scores were estimated using bootstrap.*

ILSVRC2012 portion of ImageNet (Russakovsky et al., 2015). Providing quantitative results is not as straightforward as in the case of multi-word image descriptors, since these images are not labeled with target descriptions, but with a synset from WordNet. As demonstrated in Figure 3.6, some of the lemmas in the target synsets are far too specific or unnatural for our purposes, for example *schooner* for an image depicting a sailboat or *alp* for an image of a mountain. In other cases, a particular object is named which might not be the most salient one, for example *freight car* for a picture of a graffiti with three pine trees on the side of railway carriage.

We made an attempt to search through the lemmas in the hypernym paths of the synsets until a known target lemma is reached. However, as demonstrated by examples in Figurer̃effig:pretty, these hypernyms are often very general (e.g. *device*) and predicting such high-level concepts as descriptors of the image is unrealistic. In other cases, the lemmas from the hypernym synsets are simply misleading; for example, *wood* for describing a wooden wind instrument. As can

Figure 3.6: *The caption above the images show the target labels, the hypernyms that were considered as a new target if the original was not in the vocabulary and the top N predicted words. In a large number of cases the guesses of the model are conceptually similar to the images, although, do not actually overlap with the labels or the hypernyms.*

be seen in the examples in Figure 3.6, the top ranked words predicted by our model are in fact conceptually more similar to the images covering a variety of objects and concepts than the labels specified in the dataset.

We conclude that in the future, to quantitatively investigate the cognitive plausibility of cross-situational models of word learning, the collection of feature production norms for ImageNet (Russakovsky et al., 2015) would be largely beneficial.

3.5 Discussion and conclusion

We have presented a computational cross-situational word learning model that learns word meanings from pairs images and their natural language descriptions. Unlike previous word learning studies which often rely on artificially generated perceptual input, the visual features we extract from images of natural scenes offers a more realistic simulation of the cognitive tasks humans face, since our data includes a significant level of ambiguity and referential uncertainty.

Our results suggest that the proposed model can learn meaningful representations for individual words from varied scenes and their multiword descriptions. Learning takes place incrementally and without assuming access to single-word unambiguous utterances or corrective feedback. When using the learned visual vector representations for simulating human ratings of word-pair similarity, our model shows significant correlation with human similarity judgments on a number of benchmarks. Moreover, it moderately outperforms other models that only rely on word-word co-occurrence statistics to learn word meaning.

The comparable performance of visual versus word-based models seems to be in line with Louwerse (2011), who argues that linguistic and perceptual information show a strong correlation, and therefore meaning representations solely based on linguistic data are not distinguishable from representations learned from perceptual information. However, an analysis of the impact of word concreteness on the performance of our model shows that visual features are especially useful when estimating the similarity of more concrete word pairs. In contrast, models that rely on word-based cues do not show such improvement when judging the similarity of concrete word pairs. These results suggest that these two sources of information might best be viewed as complementary, as also argued by Bruni et al. (2014).

We also used the word meaning representations that our model learns from visual input to predict the best label for a given image. This task is similar to word production in language learners. Our quantitative and qualitative analyses show that the learned representations are informative and the model can produce intuitive labels for the images in our dataset. However, as discussed in the previous section, the available image collections and their labels are not developed to suit our purpose, as most of the ImageNet labels are too detailed and at a taxonomic level which is not compatible with how language learners name a visual concept.

Finally, a natural next step for this model is to also take into account cues from sentence structure. For example, Alishahi & Chrupała (2012) try to include basic syntactic structure by introducing a separate category learning module into their model. Alternatively, learning sequential structure and visual features could be modeled in an integrated rather than modular fashion, as done by the multimodal captioning systems based on recurrent neural nets (see section 3.1.2). We are currently developing this style of integrated model to investigate the impact of structure on word learning from a cognitive point of view.

4

Representation of linguistic form and function in recurrent neural networks

abstract We present novel methods for analyzing the activation patterns of RNNs from a linguistic point of view and explore the types of linguistic structure they learn. As a case study, we use a standard standalone language model, and a multi-task gated recurrent network architecture consisting of two parallel pathways with shared word embeddings; The VISUAL pathway is trained on predicting the representations of the visual scene corresponding to an input sentence, whereas the TEXTUAL pathway is trained to predict the next word in the same sentence. We propose a method for estimating the amount of contribution of individual tokens in the input to

the final prediction of the networks. Using this method, we show that the VISUAL pathway pays selective attention to lexical categories and grammatical functions that carry semantic information, and learns to treat word types differently depending on their grammatical function and their position in the sequential structure of the sentence. In contrast, the language models are comparatively more sensitive to words with a syntactic function. Further analysis of the most informative n-gram contexts for each model shows that in comparison to the VISUAL pathway, the language models react more strongly to abstract contexts that represent syntactic constructions.

This chapter is based on Kádár, Á., Chrupała, G., & Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4), 761-780.

4.1 Introduction

Recurrent neural networks (RNNs) were introduced by Elman (1990) as a connectionist architecture with the ability to model the temporal dimension. They have proved popular for modeling language data as they learn representations of words and larger linguistic units directly from the input data, without feature engineering. Variations of the RNN architecture have been applied in several NLP domains such as parsing (Vinyals et al., 2015a) and machine translation (Bahdanau et al., 2015), as well as in computer vision applications such as image generation (Gregor et al., 2015) and object segmentation (Visin et al., 2016). RNNs are also important components of systems integrating vision and language, e.g. image (Karpathy & Fei-Fei, 2015) and video captioning (Yu et al., 2015).

These networks can represent variable-length linguistic expressions by encoding them into a fixed-size low-dimensional vector. The nature and the role of the components of these representations are not directly interpretable as they are a complex, non-linear function of the input. There have recently been numerous efforts to visualize deep models such as convolutional neural networks in the domain of computer vision, but much less so for variants of RNNs and for language processing.

The present paper develops novel methods for uncovering abstract linguistic knowledge encoded by the distributed representations of RNNs, with a specific focus on analyzing the hidden activation patterns rather than word embeddings and on the syntactic generalizations that models learn to capture. In the current work we apply our methods to a specific architecture trained on specific tasks, but also provide pointers about how to generalize the proposed analysis to other settings.

As our case study we picked the Imaginet model introduced by Chrupała et al. (2015). It is a multi-task, multi-modal architecture consisting of two Gated-Recurrent Unit (GRU) (Cho et al., 2014a; Chung et al., 2014) pathways and a shared word embedding matrix. One of the GRUs (Visual) is trained to predict image vectors given image descriptions, while the other pathway (Textual) is a language model, trained to sequentially predict each word in the descriptions. This particular architecture allows a comparative analysis of the hidden activation patterns between networks trained on two different tasks, while keeping the training data and the word embeddings fixed. Recurrent neural language models akin to Textual which are trained to predict the next symbol in a sequence are relatively well understood, and there have been some attempts to analyze their internal states (Elman, 1991; Karpathy et al., 2016, among others). In constrast, Visual maps a complete sequence of words to a representation of a corresponding visual scene and is a less commonly encountered, but a more interesting model from the point of view of representing meaning conveyed via linguistic structure. For comparison, we also consider a standard standalone language model.

We report a thorough quantitative analysis to provide a linguistic interpretation of the networks' activation patterns. We present a series of experiments using a novel method we call *omission score* to measure the importance of input tokens to the final prediction of models that compute distributed representations of sentences. Furthermore, we introduce a more global measure for estimating the informativeness of various types of n-gram contexts for each model. These techniques can be applied to various RNN architectures, Recursive Neural Networks and Convolutional Neural Networks.

Our experiments show that the VISUAL pathway in general pays special attention to syntactic categories which carry semantic content, and particularly to nouns. More surprisingly, this pathway also learns to treat word types differently depending on their grammatical function and their position in the sequential structure of the sentence. In contrast, the TEXTUAL pathway and the standalone language model are especially sensitive to the local syntactic characteristics of the input sentences. Further analysis of the most informative n-gram contexts for each model shows that while the VISUAL pathway is mostly sensitive to lexical (i.e., token n-gram) contexts, the language models react more strongly to abstract contexts (i.e., dependency relation n-grams) that represent syntactic constructions.

4.2 Related work

The direct predecessors of modern architectures were first proposed in the seminal paper of Elman (1990). He modifies the recurrent neural network architecture of Jordan (1986) by changing the output-to-memory feedback connections to hidden-to-memory recurrence, enabling Elman networks to represent arbitrary dynamic systems. Elman (1991) trains an RNN on a small synthetic sentence dataset and analyzes the activation patterns of the hidden layer. His analysis shows that these distributed representations encode lexical categories, grammatical relations and hierarchical constituent structures. Giles et al. (1991) train RNNs similar to Elman networks on strings generated by small deterministic regular grammars with the objective to recognize grammatical and reject ungrammatical strings, and develop the *dynamic state partitioning* technique to extract the learned grammar from the networks in the form of deterministic finite state automatons.

More closely related is the recent work of Li et al. (2016a), who develop techniques for a deeper understanding of the activation patterns of RNNs, but focus on models with modern architectures trained on large scale data sets. More specifically, they train Long Short-Term Memory networks (LSTM) (Hochreiter & Schmidhuber, 1997) for phrase-level sentiment analysis and present novel methods to explore the inner workings of RNNs. They measure the salience of tokens in sentences by taking the first-order derivatives of the loss with respect to the word embeddings and provide evidence that LSTMs can learn to attend to important tokens in sentences. Furthermore, they plot the activation values of hidden units through time using heat maps and visualize local semantic compositionality in RNNs. In comparison, the present work goes beyond the importance of single words and focuses more on exploring structure learning in RNNs, as well as on developing methods for a comparative analysis between RNNs that are focused on different modalities (language versus vision).

Adding an explicit attention mechanism that allows the RNNs to focus on different parts of the input was recently introduced by Bahdanau et al. (2015) in the context of extending the sequence-to-sequence RNN architecture for neural machine translation. At the decoding side this neural module assigns weights to the hidden states of the decoder, which allows the decoder to selectively pay varying degrees of attention to different phrases in the source sentence at different decoding time-steps. They also provide qualitative analysis by visualizing the attention weigths and exploring the importance of the source encodings at various decoding steps. Similarly Rocktäschel et al. (2016) use an attentive neural network architecture to perform natural language inference and visualize which parts of the hypotheses and premises the model pays attention to when deciding on the entailment relationship. Conversely, the present work focuses on RNNs without an explicit attention mechanism.

Karpathy et al. (2016) also take up the challenge of rendering RNN activation patterns understandable, but use character level language models and rather than taking a linguistic point of view, focus on error analysis and training dynamics of LSTMs and GRUs. They show that certain dimensions in the RNN hidden activation vectors have specific and interpretable functions. Similarly, Li et al. (2016d) use a Convolutional Neural Networks (CNN) based on the architecture of Krizhevsky et al. (2012), and train it on the ImageNet dataset using different random initializations. For each layer in all networks they store the activation values produced on the validation set of ILSVRC and align similar neurons of different networks. They conclude that while some features are learned across networks, some seem to depend on the initialization. Other works on visualizing the role of individual hidden units in deep models for vision synthesize images by optimizing random images through backpropagation to maximize the activity of units (Erhan et al., 2009; Simonyan et al., 2014; Yosinski et al., 2015; Nguyen et al., 2016) or to approximate the activation vectors of particular layers (Mahendran & Vedaldi, 2016; Dosovitskiy & Brox, 2015).

While this paper was under review, a number of articles appeared which also investigate linguistic representations in LSTM architectures. In an approach similar to ours, Li et al. (2016b) study the contribution of individual input tokens as well as hidden units and word embedding dimensions by erasing them from the representation and analyzing how this affects the model. They focus on text-only tasks and do not take other modalities such as visual input into account. Adi et al. (2017) take an alternative approach by introducing prediction tasks to analyze information encoded in sentence embeddings about sentence length, sentence content and word order. Finally, Linzen et al. (2016) examine the acquisition of long-distance dependencies through the study of number agreement in different variations of an LSTM model with different objectives (number prediction, grammaticality judgment, and language modeling). Their results show that such dependencies can be captured with very high accuracy when the model receives a strong supervision signal (that is, whether the subject is plural or singular), but simple language models still capture the majority of test cases. While they focus on an in-depth analysis of a single phenomenon, in our work we are interested in methods which make it possible to uncover a broad variety of patterns of bebavior in RNNs.

In general, there has been a growing interest within computer vision in understanding deep models, with a number of papers dedicated to visualizing learned CNN filters and pixel saliencies (Simonyan et al., 2014; Yosinski et al., 2015; Mahendran & Vedaldi, 2015). These techniques have also led to improvements in model performance (Eigen et al., 2014) and transferability of features (Zhou et al., 2015). To date there has been much less work on such issues within computational linguistics. We aim to fill this gap by adapting existing methods as well as developing novel techniques to explore the linguistic structure learned by recurrent networks.

4.3 Models

In our analyses of the acquired linguist knowledge, we apply our methods to the following models:

- IMAGINET: A multi-modal Gated Recurrent Unit (GRU) network consisting of two pathways, VISUAL and TEXTUAL, coupled via word embeddings.
- LM: A (unimodal) language model consisting of a GRU network.
- SUM: A network with the same objective as the VISUAL pathway of IMAGINET, but which uses sum of word embeddings instead of a GRU.

The rest of this section gives a detailed description of these models.

4.3.1 Gated Recurrent Neural Networks

One of the main difficulties for training traditional Elman networks arises from the fact that they overwrite their hidden states at every time step with a new value computed from the current input $\mathbf{x_t}$ and the previous hidden state $\mathbf{h_{t-1}}$. Similarly to LSTMs, Gated Recurrent Unit networks introduce a mechanism which facilitates the retention of information over multiple time steps. Specifically, the GRU computes the hidden state at current time step $\mathbf{h}_t$, as the linear combination of previous activation $\mathbf{h_{t-1}}$, and a new *candidate* activation $\tilde{\mathbf{h}}_t$:

$$\mathrm{GRU}(\mathbf{h}_{t-1}, \mathbf{x}_t) = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \tag{4.1}$$

where $\odot$ is elementwise multiplication, and the update gate activation $\mathbf{z_t}$ determines the amount of new information mixed in the current state:

$$\mathbf{z}_t = \sigma_s(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1}) \tag{4.2}$$

The candidate activation is computed as:

$$\tilde{\mathbf{h}}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1})) \tag{4.3}$$

The reset gate $\mathbf{r_t}$ determines how much of the current input $\mathbf{x_t}$ is mixed in the previous state $\mathbf{h}_{t-1}$ to form the candidate activation:

$$\mathbf{r}_t = \sigma_s(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1}) \tag{4.4}$$

where $\mathbf{W}$, $\mathbf{U}$, $\mathbf{W}_z$, $\mathbf{U}_z$, $\mathbf{W}_r$ and $\mathbf{U}_r$ are learnable parameters.

4.3.2 Imaginet

Imaginet introduced in Chrupała et al. (2015) is a multi-modal GRU network architecture that learns visually grounded meaning representations from textual and visual input. It acquires linguistic knowledge through language comprehension, by receiving a description of a scene and trying to visualise it through predicting a visual representation for the textual description, while concurrently predicting the next word in the sequence.

Figure 4.1 shows the structure of Imaginet. As can be seen from the figure, the model consists of two GRU pathways, Textual and Visual, with a shared word embedding matrix. The inputs to the model are pairs of image descriptions and their corresponding images. The Textual pathway predicts the next word at each position in

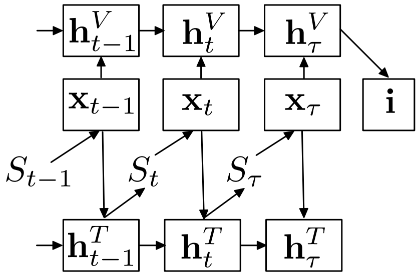


Figure 4.1: Structure of IMAGINET, adapted from Chrupała et al. (2015).

the sequence of words in each caption, whereas the VISUAL pathway predicts a visual representation of the image that depicts the scene described by the caption after the final word is received.

Formally, each sentence is mapped to two sequences of hidden states, one by VISUAL and the other by TEXTUAL:

$$\mathbf{h}^V_t = \mathrm{GRU}^V(\mathbf{h}^V_{t-1}, \mathbf{x}_t) \tag{4.5}$$
$$\mathbf{h}^T_t = \mathrm{GRU}^T(\mathbf{h}^T_{t-1}, \mathbf{x}_t) \tag{4.6}$$

At each time step TEXTUAL predicts the next word in the sentence S from its current hidden state $\mathbf{h}^T_t$, while VISUAL predicts the image-vector[1] $\hat{\mathbf{i}}$ from its last hidden representation $\mathbf{h}^V_t$.

$$\hat{\mathbf{i}} = \mathbf{V}\mathbf{h}^V_\tau \tag{4.7}$$
$$p(S_{t+1}|S_{1:t}) = \mathrm{softmax}(\mathbf{L}\mathbf{h}^T_t) \tag{4.8}$$

The loss function is a multi-task objective which penalizes error on the visual and the textual targets simultaneously. The objective combines

[1]Representing the full image, extracted from the pre-trained Convolutional Neural Network of Simonyan & Zisserman (2015).

cross-entropy loss L^T for the word predictions and cosine distance L^V for the image predictions[2], weighting them with the parameter α (set to 0.1).

$$L^T(\theta) = -\frac{1}{\tau}\sum_{t=1}^{\tau} \log p(S_t|S_{1:t}) \tag{4.9}$$

$$L^V(\theta) = 1 - \frac{\hat{\mathbf{i}} \cdot \mathbf{i}}{||\hat{\mathbf{i}}|| \, ||\mathbf{i}||} \tag{4.10}$$

$$L = \alpha L^T + (1-\alpha) L^V \tag{4.11}$$

For more details about the Imaginet model and its performance see Chrupała et al. (2015). Note that we introduce a small change in the image representation: we observe that using standardized image vectors, where each dimension is transformed by subtracting the mean and dividing by standard deviation, improves performance.

4.3.3 Unimodal language model

The model LM is a language model analogous to the Textual pathway of Imaginet with the difference that its word embeddings are not shared, and its loss function is the cross-entropy on word prediction. Using this model we remove the visual objective as a factor, as the model does not use the images corresponding to captions in any way.

[2]Note that the original formulation in Chrupała et al. (2015) uses mean squared error instead; as the performance of Visual is measured on image-retrieval which is based on cosine distances, we use cosine distance as the visual loss here.

4.3.4 Sum of word embeddings

The model SUM is a stripped-down version of the VISUAL pathway, which does not share word embeddings, only uses the cosine loss function, and replaces the GRU network with a summation over word embeddings. This removes the effect of word order from consideration. We use this model as a baseline in the sections which focus on language structure.

4.4 Experiments

In this section, we report a series of experiments in which we explore the kinds of linguistic regularities the networks learn from word-level input. In Section 4.4.1 we introduce *omission score*, a metric to measure the contribution of each token to the prediction of the networks, and in Section 4.4.2 we analyze how omission scores are distributed over dependency relations and part-of-speech categories. In Section 4.4.3 we investigate the extent to which the importance of words for the different networks depend on the words themselves, their sequential position, and their grammatical function in the sentences. Finally, in Section 4.4.4 we systematically compare the types of n-gram contexts that trigger individual dimensions in the hidden layers of the networks, and discuss their level of abstractness.

In all these experiments we report our findings based on the IMAGINET model, and whenever appropriate compare it to our two other models LM and SUM. For all the experiments, we trained the models on the training portion of the MSCOCO image-caption dataset (Lin et al., 2014), and analyzed the representations of the sentences in the validation set corresponding to 5000 randomly chosen images. The

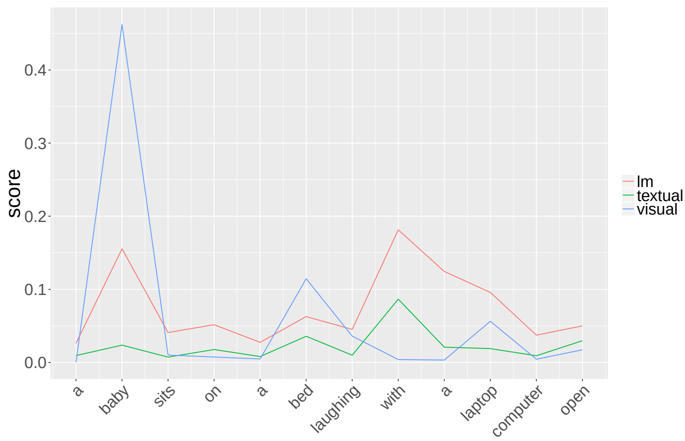


Figure 4.2: Omission scores for the example sentence *a baby sits on a bed laughing with a laptop computer open* for LM and the two pathways, TEXTUAL and VISUAL, of IMAGINET.

target image representations were extracted from the pre-softmax layer of the 16-layer CNN of Simonyan & Zisserman (2015).

4.4.1 Computing Omission Scores

We propose a novel technique for interpreting the activation patterns of neural networks trained on language tasks from a linguistic point of view, and focus on the high-level understanding of what parts of the input sentence the networks pay most attention to. Furthermore, we investigate if the networks learn to assign different amounts of importance to tokens depending on their position and grammatical function in the sentences.

In all the models the full sentences are represented by the acti-

Figure 4.3: Images retrieved for the example sentence *a baby sits on a bed laughing with a laptop computer open* (left) and the same sentence with the second word omitted (right).

vation vector at the end-of-sentence symbol ($\mathbf{h}_{\text{end}}$). We measure the salience of each word S_i in an input sentence $S_{1:n}$ based on how much the representation of the partial sentence $S_{\setminus i} = S_{1:i-1}S_{i+1:n}$, with the omitted word S_i, deviates from that of the original sentence representation. For example, the distance between $\mathbf{h}_{\text{end}}$("*the black dog is running*") and $\mathbf{h}_{\text{end}}$("*the dog is running*") determines the importance of *black* in the first sentence. We introduce the measure omission(i, S) for estimating the salience of a word S_i:

$$\text{omission}(i, S) = 1 - \text{cosine}(\mathbf{h}_{\text{end}}(S), \mathbf{h}_{\text{end}}(S_{\setminus i})) \tag{4.12}$$

Figure 4.2 demonstrates the omission scores for the LM, Visual and Textual models for an example caption. Figure 4.3 shows the images retrieved by Visual for the full caption and for the one with the word *baby* omitted. The images are retrieved from the validation set of MS-COCO by: 1) computing the image representation of the given sentence with Visual; 2) extracting the CNN features for the images from the set; and 3) finding the image that minimizes the cosine distance to the query. The omission scores for VISUAL show that the model paid attention mostly to *baby* and *bed* and slightly to *laptop*, and retrieved an image depicting a baby sitting on a bed with a laptop. Removing the word *baby* leads to an image that depicts an adult male laying on a bed. Figure 4.2 also shows that in contrast to VISUAL, TEXTUAL distributes its attention more evenly across time steps instead of focusing on the types of words related to the corresponding visual scene. The peaks for LM are the same as for TEXTUAL, but the variance of the omission scores is higher, suggesting that the unimodal language model is more sensitive overall to input perturbations than TEXTUAL.

4.4.2 Omission score distributions

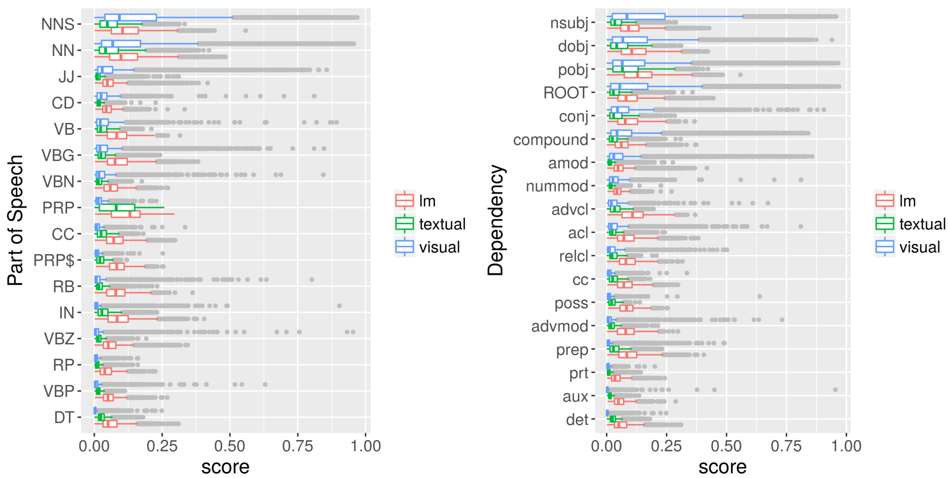


Figure 4.4: Distribution of omission scores for POS (left) and dependency labels (right), for the TEXTUAL and VISUAL pathways and for LM. Only labels which occur at least 1250 times are included.

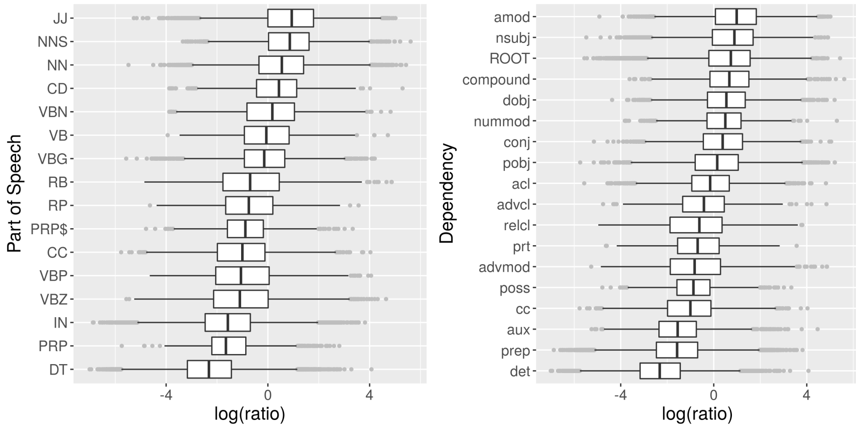


Figure 4.5: Distributions of log ratios of omission scores of TEXTUAL to VISUAL per POS (left) and dependency labels (right). Only labels which occur at least 1250 times are included.

The omission scores can be used not only to estimate the importance of individual words, but also of syntactic categories. We estimate the salience of each syntactic category by accumulating the omission scores for all words in that category. We tag every word in a sentence with the part-of-speech (POS) category and the dependency relation (deprel) label of its incoming arc. For example, for the sentence *the black dog*, we get (*the*, DT, det), (*black*, JJ, amod), (*dog*, NN, root). Both POS tagging and dependency parsing are performed using the `en_core_web_md` dependency parser from the Spacy package.[3]

Figure 4.4 shows the distribution of omission scores per POS and dependency label for the two pathways of IMAGINET and for LM.[4]

[3] Available at `https://spacy.io/`.

[4] The boxplots in this and subsequent figures are Tukey boxplots and should be interpreted as follows: the box extends from the 25th to the 75th percentile of the data; the line across the box is the 50th percentile, while the whiskers

The general trend is that for the VISUAL pathway, the omission scores are high for a small subset of labels - corresponding mostly to nouns, less so for adjectives and even less for verbs - and low for the rest (mostly function words and various types of verbs). For TEXTUAL the differences are smaller, and the pathway seems to be sensitive to the omission of most types of words. For LM the distribution over categories is also relatively uniform, but the omission scores are higher overall than for TEXTUAL.

Figure 4.5 compares the two pathways of IMAGINET directly using the log of the ratio of the VISUAL to TEXTUAL omission scores, and plots the distribution of this ratio for different POS and dependency labels. Log ratios above zero indicate stronger association with the VISUAL pathway and below zero with the TEXTUAL pathway. We see that in relative terms, VISUAL is more sensitive to adjectives (JJ), nouns (NNS, NN), numerals (CD) and participles (VBN), and TEXTUAL to determiners (DT), pronouns (PRP), prepositions (IN) and finite verbs (VBZ, VBP).

This picture is complemented by the analysis of the relative importance of dependency relations: VISUAL pays most attention to the relations AMOD, NSUBJ, ROOT, COMPOUND, DOBJ, NUMMOD whereas TEXTUAL is more sensitive to DET, PREP, AUX, CC, POSS, ADVMOD, PRT, RELCL. As expected, VISUAL is more focused on grammatical functions typically filled by semantically contentful words, while TEXTUAL distributes its attention more uniformly and attends relatively more to purely grammatical functions.

It is worth noting, however, the relatively low omission scores for verbs in the case of VISUAL. One might expect that the task of image

extend past the lower and upper quartile to 1.5× the interquartile range (i.e. 75th percentile - 25th percentile); the points are outliers.

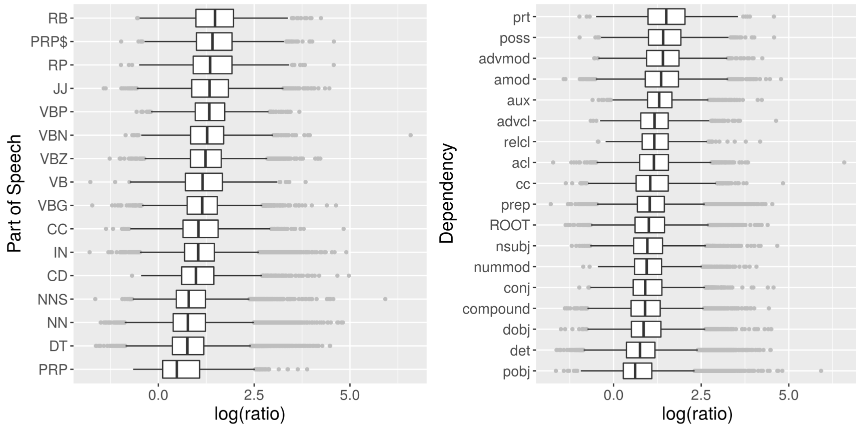


Figure 4.6: Distributions of log ratios of omission scores of LM to TEXTUAL per POS (left) and dependency labels (right). Only labels which occur at least 1250 times are included.

prediction from descriptions requires general language understanding and so high omission scores for all content words in general; however, the results suggest that this setting is not optimal for learning useful representations of verbs, which possibly leads to representations that are too task-specific and not transferable across tasks.

Figure 4.6 shows a similar analysis contrasting LM with the TEXTUAL pathway of IMAGINET. The first observation is that the range of values of the log ratios is narrow, indicating that the differences between these two networks regarding which grammatical categories they are sensitive to is less pronounced than when comparing VISUAL to TEXTUAL. While the size of the effect is weak, there also seems to be a tendency for the TEXTUAL model to pay relatively more attention to content and less to function words, compared to LM: it may be that the VISUAL pathway pulls TEXTUAL in this direction by sharing word embeddings with it.

Most of our findings up to this point conform reasonably well to prior expectations about effects that particular learning objectives should have. This fact serves to validate our methods. In the next section we go on to investigate less straightforward patterns.

4.4.3 Beyond Lexical Cues

Models that utilize the sequential structure of language have the capacity to interpret the same word type differently depending on the context. The omission score distributions in Section 4.4.2 show that in the case of IMAGINET the pathways are differentially sensitive to content vs. function words. In principle, this may be either just due to purely lexical features or the model may actually learn to pay more attention to the same word type in appropriate contexts. This section investigates to what extent our models discriminate between occurrences of a given word in different positions and grammatical functions.

We fit four L2-penalized linear regression models which predict the omission scores per token with the following predictor variables:

1. LR WORD: word type
2. LR +DEP: word type, dependency label and their interaction
3. LR +POS: word type, position (binned as FIRST, SECOND, THIRD, MIDDLE, ANTEPENULT, PENULT, LAST) and their interaction
4. LR FULL: word type, dependency label, position, word:dependency interaction, word:position interaction

Table 4.1: Proportion of variance in omission scores explained by linear regression.

	word	+pos	+dep	full
SUM	0.654	0.661	0.670	0.670
LM	0.358	0.586	0.415	0.601
TEXTUAL	0.364	0.703	0.451	0.715
VISUAL	0.490	0.506	0.515	0.523

We use the 5000-image portion of MSCOCO validation data for training and test. The captions contain about 260,000 words in total, of which we use 100,000 to fit the regression models. We then use the rest of the words to compute the proportion of variance explained by the models. For comparison we also use the SUM model which composes word embeddings via summation, and uses the same loss function as VISUAL. This model is unable to encode information about word order, and thus is a good baseline here as we investigate the sensitivity of the networks to positional and structural cues.

Table 4.1 shows the proportion of variance R^2 in omission scores explained by the linear regression with the different predictors. The raw R^2 scores show that for the language models LM and TEXTUAL, the word type predicts the omission-score to much smaller degree compared to VISUAL. Moreover, adding information about either the position or the dependency labels increases the explained variance for all models. However, for the TEXTUAL and LM models the position of the word adds considerable amount of information. This is not surprising considering that the omission scores are measured with respect to the final activation state, and given the fact that in a language model the recent history is most important for accurate prediction.

Figure 4.7 offers a different view of the data, showing the increase or decrease in R^2 for the models relative to LR +POS to emphasise the importance of syntactic structure beyond the position in the sentence. Interestingly, for the VISUAL model, dependency labels are more informative than linear position, hinting at the importance of syntactic structure beyond linear order. There is a sizeable increase in R^2 between LR +POS and LR FULL in the case of VISUAL, suggesting that the omission scores for VISUAL depend on the words' grammatical function in sentences, *even after controlling for word identity and linear position.* In contrast, adding additional information on top of lexical features in the case of SUM increases the explained variance only slightly, which is most likely due to the unseen words in the held out set.

Overall, when regressing on word identities, word position and dependency labels, the VISUAL model's omission scores are the hardest to predict of the four models. This suggests that VISUAL may be encoding additional structural features not captured by these predictors. We will look more deeply into such potential features in the following sections.

4.4.3.1 Sensitivity to grammatical function

In order to find out some of the specific syntactic configurations leading to an increase in R^2 between the LR WORD and LR +DEP predictors in the case of VISUAL, we next considered all word types with occurrence counts of at least 100 and ranked them according to how much better, on average, LR +DEP predicted their omission scores compared to LR WORD.

Figure 4.8 shows the per-dependency omission score distributions for seven top-ranked words. There are clear and large differences in

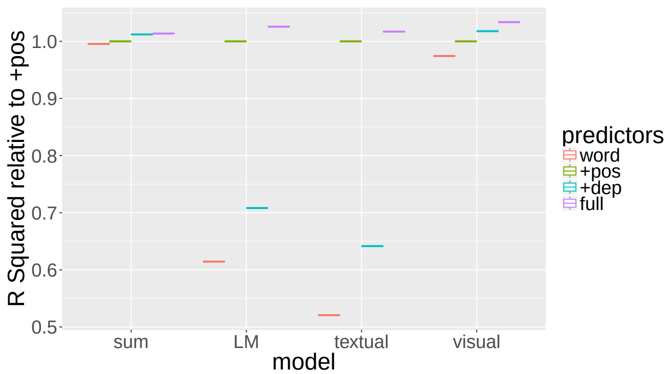


Figure 4.7: Proportion of variance in omission scores explained by the linear regression models for SUM, LM, VISUAL and TEXTUAL, relative to regressing on word identity and position only.

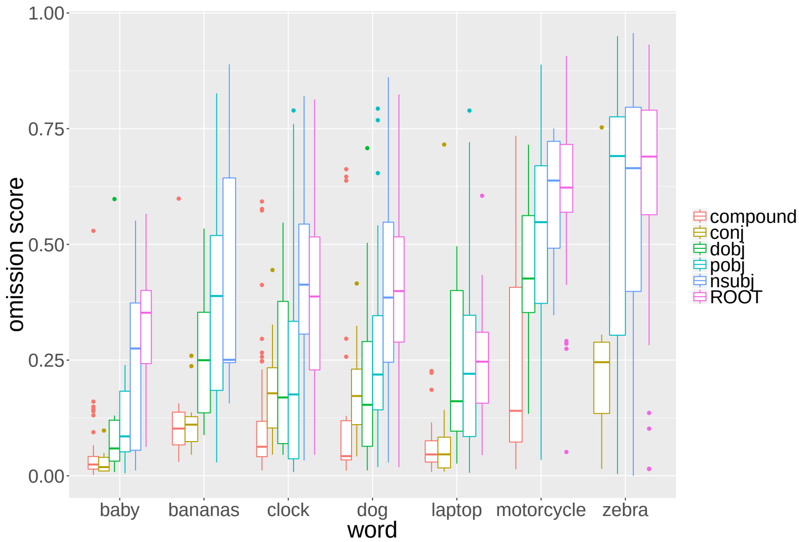


Figure 4.8: Distribution of omission scores per dependency label for the selected word types.

how these words impact the network's representation depending on what grammatical function they fulfil. They all have large omission scores when they occur as NSUBJ (nominal subject) or ROOT, likely due to the fact that these grammatical functions typically have a large contribution to the complete meaning of a sentence. Conversely, all have small omission scores when appearing as CONJ (conjunct): this is probably because in this position they share their contribution with the first, often more important, member of the conjunction, for example in *A cow and its baby eating grass*.

4.4.3.2 Sensitivity to linear structure

As observed in Section 4.4.3, adding extra information about the position of words explains more of the variance in the case of VISUAL and

Figure 4.9: Coefficients on the y-axis of LR FULL corresponding to the position variables on the x-axis.

especially TEXTUAL and LM. Figure 4.9 shows the coefficients corresponding to the position variables in LR FULL. Since the omission scores are measured at the end-of-sentence token, the expectation is that for TEXTUAL and LM, as language models, the words appearing closer to the end of the sentence would have a stronger effect on the omission scores. This seems to be confirmed by the plot as the coefficients for these two networks up until the *antepenult* are all negative.

For the VISUAL model it is less clear what to expect: on the one hand due to their chain structure, RNNs are better at keeping track of short-distance rather than long-distance dependencies and thus we can expect tokens in positions closer to the end of the sentence to be more important. On the other hand, in English the information structure of a single sentence is expressed via linear ordering: the TOPIC of a sentence appears sentence-initially, and the COMMENT follows. In the context of other text types such as dialog or multi-sentence narrative structure, we would expect COMMENT to often be more important than TOPIC as COMMENT will often contain new information in these cases. In our setting of image captions however, sentences are not part of a larger discourse; it is sentence initial material that typically contains the most important objects depicted in the image, e.g. *two zebras are grazing in tall grass on a savannah.* Thus, for the task of predicting features of the visual scene, it would be advantageous to detect the topic of the sentence and up-weight its importance in the final meaning representation. Figure 4.9 appears to support this hypothesis and the network does learn to pay more attention to words appearing sentence-initially. This effect seems to be to some extent mixed with the recency bias of RNNs as perhaps indicated by the relatively high coefficient of the *last* position for VISUAL.

4.4.4 Lexical versus abstract contexts

We would like to further analyze the kinds of linguistic features that the hidden dimensions of RNNs encode. Previous work (Karpathy et al., 2016; Li et al., 2016d) has shown that in response to the task the networks are trained for, individual dimensions in the hidden layers of RNNs can become *specialised* in responding to certain types of triggers, including the tokens or token types at each time step, as well as the preceding context of each token in the input sentence.

Here we perform a further comparison between the models based on the hypothesis that due to their different objectives, the activations of the dimensions of the last hidden layer of VISUAL are more characterized by semantic relations within contexts, whereas the hidden dimensions in TEXTUAL and LM are more focused on extracting syntactic patterns. In order to quantitatively test this hypothesis, we measure the strength of association between activations of hidden dimensions and either lexical (token n-grams) or structural (dependency label n-grams) types of context.

For each pathway, we define A_i as a discrete random variable corresponding to a binned activation over time steps at hidden dimension i, and C as a discrete random variable indicating the context (where C can be of type 'word trigram' or 'dependency label bigram', for example). The strength of association between A_i and C can be measured by their mutual information:

$$\mathrm{I}(A_i; C) = \sum_{a \in A_i} \sum_{c \in C} p(a, c) \log \left(\frac{p(a, c)}{p(a)p(c)} \right) \tag{4.13}$$

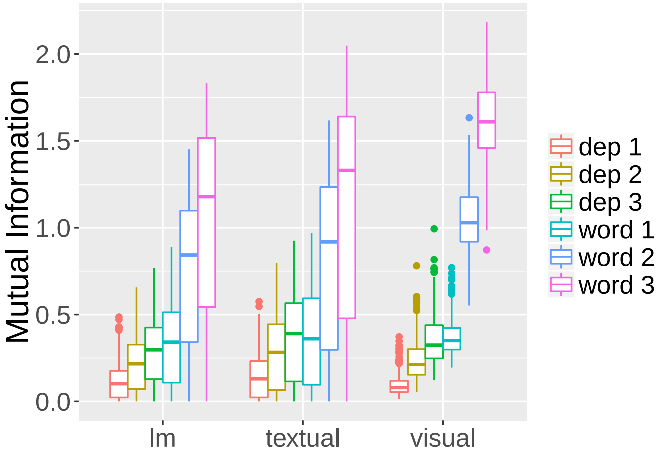


Figure 4.10: Distributions of the mutual information scores for the three networks and the six context types.

Similarly to Li et al. (2016d), the activation value distributions are discretized into percentile bins per dimension, such that each bin contains 5% of the marginal density. For context types, we used unigrams, bigrams and trigrams of both dependency labels and words. Figure 4.10 shows the distributions of the mutual information scores for the three networks and the six context types. Note that the scores are not easily comparable between context types, due the different support of the distributions; they are, however, comparable across the networks. The figure shows LM and Textual as being very similar, while Visual exhibits a different distribution. We next compare the models' scores pairwise to pinpoint the nature of the differences.

We use the notation MI_C^{LM}, MI_C^{T} and MI_C^{V} to denote the median mutual information score over all dimensions of LM, Textual and Visual respectively, when considering context C. We then compute log ratios $\log(\mathrm{MI}_C^{T}/\mathrm{MI}_C^{V})$ and $\log(\mathrm{MI}_C^{LM}/\mathrm{MI}_C^{T})$ for all six context types

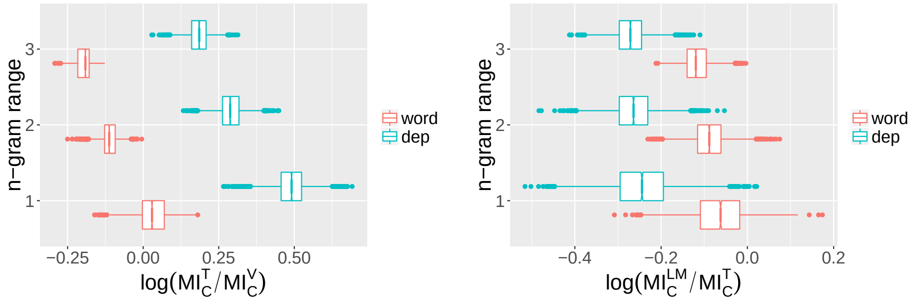


Figure 4.11: Bootstrap distributions of log ratios of median mutual information scores for word and dependency contexts. Left: TEXTUAL vs VISUAL; right: LM vs TEXTUAL

C. In order to quantify variability we bootstrap this statistic with 5000 replicates. Figure 4.11 shows the resulting bootstrap distributions for uni-, bi-, and trigram contexts, in the word and dependency conditions.

The clear pattern is that for TEXTUAL versus VISUAL, the log ratios are much higher in the case of the dependency contexts, with no overlap between the bootstrap distributions. Thus, in general, the size of the relative difference between TEXTUAL and VISUAL median mutual information score is much more pronounced for dependency context types. This suggests that features that are encoded by the hidden dimensions of the models are indeed different, and that the features encoded by TEXTUAL are more associated with syntactic constructions than in the case of VISUAL. In contrast, when comparing LM with TEXTUAL, the difference between context types is much less pronounced, with distributions overlapping. Though the difference is small, it goes in the direction of the dimensions of the TEXTUAL model showing higher sensitivity towards dependency contexts.

The mutual information scores can be used to pinpoint specific dimensions of the hidden activation vectors which are strongly associated with a particular type of context. Table 4.2 lists for each network the dimension with the highest mutual information score with respect to the *dependency trigram* context type, together with the top five contexts where these dimensions carry the highest value. In spite of the quantitative difference between the networks discussed above, the dimensions which come up top seem to be capturing something quite similar for the three networks: (a part of) a construction with an animate root or subject modified by a participle or a prepositional phrase, though this is somewhat less clean-cut for the VISUAL pathway where only two out of five top context clearly conform to this pattern. Other interesting templates can be found by visual inspection of the contexts where high-scoring dimensions are active; for example, dimension 324 of LM is high for *word bigram* contexts including *people preparing, gets ready, man preparing, woman preparing, teenager preparing.*

4.5 Discussion

The goal of our paper is to propose novel methods for the analysis of the encoding of linguistic knowledge in RNNs trained on language tasks. We focused on developing quantitative methods to measure the importance of different kinds of words for the performance of such models. Furthermore, we proposed techniques to explore what kinds of linguistic features the models learn to exploit beyond lexical cues.

Using the IMAGINET model as our case study, our analyses of

Table 4.2: Dimensions most strongly associated with the dependency trigram context type, and the top five contexts in which these dimensions have high values.

Network	Dimension	Examples
LM	511	cookie/pobj attached/acl to/prep
		people/pobj sitting/acl in/prep
		purses/pobj sitting/pcomp on/prep
		and/cc talks/conj on/prep
		desserts/pobj sitting/acl next/advmod
TEXTUAL	735	male/root on/prep a/det
		person/nsubj rides/root a/det
		man/root carrying/acl a/det
		man/root on/prep a/det
		person/root on/prep a/det
VISUAL	875	man/root riding/acl a/det
		man/root wearing/acl a/det
		is/aux wearing/conj a/det
		a/det post/pobj next/advmod
		one/nummod person/nsubj is/aux

the hidden activation patterns show that the VISUAL model learns an abstract representation of the information structure of a single sentence in the language, and pays selective attention to lexical categories and grammatical functions that carry semantic information. In contrast, the language model TEXTUAL is sensitive to features of a more syntactic nature. We have also shown that each network contains specialized units which are tuned to both lexical and structural patterns that are useful for the task at hand.

4.5.1 Generalizing to other architectures

For other RNN architectures such as LSTMs and their bi-directional variants, measuring the contribution of tokens to their predictions (or the omission scores) can be straight-forwardly computed using their hidden state at the last time step used for prediction. Furthermore, the technique can be applied in general to other architectures which map variable-length linguistic expressions to the same fixed dimensional space and perform predictions based on these embeddings. This includes tree-structured Recursive Neural Network models such as the Tree-LSTM introduced in Tai et al. (2015), or the CNN architecture of Kim (2014) for sentence classification. However, the presented analysis and results regarding word positions can only be meaningful for Recurrent Neural Networks as they compute their representations sequentially and are not limited by fixed window sizes.

A limitation of the generalizability of our analysis is that in the case of bi-directional architectures, the interpretation of the features extracted by the RNNs that process the input tokens in the reversed order might be hard from a linguistic point of view.

4.5.2 Future directions

In future we would like to apply the techniques introduced in this paper to analyze the encoding of linguistic form and function of recurrent neural models trained on different objectives, such as neural machine translation systems (Sutskever et al., 2014) or the purely distributional sentence embedding system of Kiros et al. (2015). A number of recurrent neural models rely on a so-called attention mechanism, first introduced by Bahdanau et al. (2015) under the name of soft alignment. In these networks attention is explicitly represented, and it would be interesting to see how our method of discovering implicit attention, the omission score, compares. For future work we also propose to collect data where humans assess the importance of each word in a sentence and explore the relationship between omission scores for various models and human annotations. Finally, one of the benefits of understanding how linguistic form and function is represented in RNNs is that it can provide insight into how to improve systems. We plan to draw on lessons learned from our analyses in order to develop models with better general-purpose sentence representations.

5

Imagination Improves Multimodal Translation

abstract We decompose multimodal translation into two sub-tasks: learning to translate and learning visually grounded representations. In a multitask learning framework, translations are learned in an attention-based encoder-decoder, and grounded representations are learned through image representation prediction. Our approach improves translation performance compared to the state of the art on the Multi30K dataset. Furthermore, it is equally effective if we train the image prediction task on the external MS COCO dataset, and we find improvements if we train the translation model on the external News Commentary parallel text.

This chapter is based on Elliott, D., & Kádár, Á. (2017, November). Imagination Improves Multimodal Translation. *In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 130-141).

5.1 Introduction

Multimodal machine translation is the task of translating sentences in context, such as images paired with a parallel text (Specia et al., 2016). This is an emerging task in the area of multilingual multimodal natural language processing. Progress on this task may prove useful for translating the captions of the images illustrating online news articles, and for multilingual closed captioning in international television and cinema.

Initial efforts have not convincingly demonstrated that visual context can improve translation quality. In the results of the First Multimodal Translation Shared Task, only three systems outperformed an off-the-shelf text-only phrase-based machine translation model, and the best performing system was equally effective with or without the visual features (Specia et al., 2016). There remains an open question about how translation models should take advantage of visual context.

We present a multitask learning model that decomposes multimodal translation into learning a translation model and learning visually grounded representations. This decomposition means that our model can be trained over external datasets of parallel text or described images, making it possible to take advantage of existing resources. Figure 5.1 presents an overview of our model, Imagination, in which source language representations are shared between tasks through the Shared Encoder. The translation decoder is an attention-based neural machine translation model (Bahdanau et al., 2015), and the image prediction decoder is trained to predict a global feature vector of an image that is associated with a sentence (Chrupała et al.,

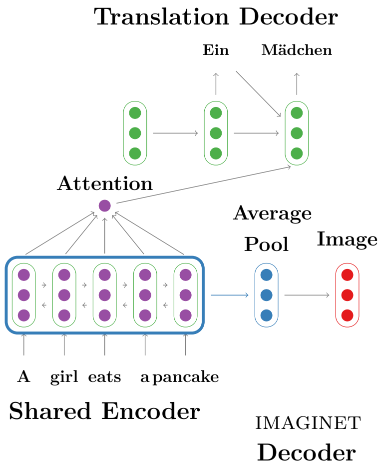


Figure 5.1: The Imagination model learns visually-grounded representations by sharing the encoder network between the Translation Decoder with image prediction in the IMAGINET Decoder.

2015, IMAGINET). This decomposition encourages grounded learning in the shared encoder because the IMAGINET decoder is trained to imagine the image associated with a sentence. It has been shown that grounded representations are qualitatively different from their text-only counterparts (Kádár et al., 2017) and correlate better with human similarity judgements (Chrupała et al., 2015). We assess the success of the grounded learning by evaluating the image prediction model on an image–sentence ranking task to determine if the shared representations are useful for image retrieval (Hodosh et al., 2013). In contrast with most previous work, our model does not take images as input at translation time, rather it learns grounded representations in the shared encoder.

We evaluate Imagination on the Multi30K dataset (Elliott et al., 2016) using a combination of in-domain and out-of-domain data. In the in-domain experiments, we find that multitasking translation with image prediction is competitive with the state of the art. Our model achieves 55.8 Meteor as a single model trained on multimodal in-domain data, and 57.6 Meteor as an ensemble.

In the experiments with out-of-domain resources, we find that the improvement in translation quality holds when training the IMAGINET decoder on the MS COCO dataset of described images (Chen et al., 2015). Furthermore, if we significantly improve our text-only baseline using out-of-domain parallel text from the News Commentary corpus (Tiedemann, 2012), we still find improvements in translation quality from the auxiliary image prediction task. Finally, we report a state-of-the-art result of 59.3 Meteor on the Multi30K corpus when ensembling models trained on in- and out-of-domain resources.

The main contributions of this paper are:

- We show how to apply multitask learning to multimodal translation. This makes it possible to train models for this task using external resources alongside the expensive triple-aligned source-target-image data.

- We decompose multimodal translation into two tasks: learning to translate and learning grounded representations. We show that each task can be trained on large-scale external resources, e.g. parallel news text or images described in a single language.

- We present a model that achieves state of the art results without using images as an input. Instead, our model learns visually grounded source language representations using an auxiliary image prediction objective. Our model does not need any additional parameters to translate unseen sentences.

5.2 Problem Formulation

Multimodal translation is the task of producing target language translation y, given the source language sentence x and additional context, such as an image v (Specia et al., 2016). Let x be a source language sentence consisting of N tokens: x_1, x_2, ..., x_n and let y be a target language sentence consisting of M tokens: y_1, y_2, ..., y_m. The training data consists of tuples $\mathcal{D} \in (x, y, v)$, where x is a description of image v, and y is a translation of x.

Multimodal translation has previously been framed as minimising the negative log-likelihood of a translation model that is additionally conditioned on the image, i.e. $J(\theta) = -\sum_j \log \mathrm{p}(y_j | y_{<j}, x, v)$. Here, we decompose the problem into learning to translate and learning visually grounded representations. The decomposition is based on sharing parameters θ between these two tasks, and learning task-specific parameters ϕ. We learn the parameters in a multitask model with shared parameters in the source language encoder. The translation model has task-specific parameters ϕ^t in the attention-based decoder, which are optimized through the translation loss $J_T(\theta, \phi^t)$. Grounded representations are learned through an image prediction model with task-specific parameters ϕ^g in the image-prediction decoder by minimizing $J_G(\theta, \phi^g)$. The joint objective is given by mixing the translation and image prediction tasks with the parameter w:

$$J(\theta, \phi) = wJ_T(\theta, \phi^t) + (1 - w)J_G(\theta, \phi^g) \tag{5.1}$$

Our decomposition of the problem makes it straightforward to optimise this objective without paired tuples, e.g. where we have an external dataset of described images $\mathcal{D}_{image} \in (x, v)$ or an external parallel corpus $\mathcal{D}_{text} \in (x, y)$.

We train our multitask model following the approach of Luong et al. (2016). We define a primary task and an auxiliary task, and a set of parameters θ to be shared between the tasks. A minibatch of updates is performed for the primary task with probability w, and for the auxiliary task with $1 - w$. The primary task is trained until convergence and weight w determines the frequency of parameter updates for the auxiliary task.

5.3 Imagination Model

5.3.1 Shared Encoder

The encoder network of our model learns a representation of a sequence of N tokens $x_{1 \ldots n}$ in the source language with a bidirectional recurrent neural network (Schuster & Paliwal, 1997). This representation is shared between the different tasks. Each token is represented by a one-hot vector $\mathbf{x_i}$, which is mapped into an embedding $\mathbf{e_i}$ through a learned matrix $\mathbf{E}$:

$$\mathbf{e_i} = \mathbf{x_i} \cdot \mathbf{E} \tag{5.2}$$

A sentence is processed by a pair of recurrent neural networks, where one captures the sequence left-to-right (forward), and the other captures the sequence right-to-left (backward). The initial state of the encoder $\mathbf{h_{-1}}$ is a learned parameter:

$$\overrightarrow{\mathbf{h_i}} = \overrightarrow{\text{RNN}}(\overrightarrow{\mathbf{h_{i-1}}}, \mathbf{e_i}) \tag{5.3}$$
$$\overleftarrow{\mathbf{h_i}} = \overleftarrow{\text{RNN}}(\overleftarrow{\mathbf{h_{i-1}}}, \mathbf{e_i}) \tag{5.4}$$

Each token in the source language input sequence is represented by a concatenation of the forward and backward hidden state vectors:

$$\mathbf{h_i} = [\overrightarrow{\mathbf{h_i}}; \overleftarrow{\mathbf{h_i}}] \tag{5.5}$$

5.3.2 Neural Machine Translation Decoder

The translation model decoder is an attention-based recurrent neural network (Bahdanau et al., 2015). Tokens in the decoder are represented by a one-hot vector $\mathbf{y_j}$, which is mapped into an embedding $\mathbf{e_j}$ through a learned matrix $\mathbf{E_y}$:

$$\mathbf{e_j} = \mathbf{y_j} \cdot \mathbf{E_y} \tag{5.6}$$

The inputs to the decoder are the previously predicted token $\mathbf{y_{j-1}}$, the previous decoder state $\mathbf{d_{j-1}}$, and a timestep-dependent context vector $\mathbf{c_j}$ calculated over the encoder hidden states:

$$\mathbf{d_j} = \mathrm{RNN}(\mathbf{d_{j-1}}, \mathbf{y_{j-1}}, \mathbf{e_j}) \tag{5.7}$$

The initial state of the decoder $\mathbf{d}_{\text{-}1}$ is a nonlinear transform of the mean of the encoder states, where $\mathbf{W}_{init}$ is a learned parameter:

$$\mathbf{d}_{\text{-}1} = \tanh(\mathbf{W}_{init} \cdot \frac{1}{N} \sum_{i}^{N} \mathbf{h_i}) \tag{5.8}$$

The context vector c_j is a weighted sum over the encoder hidden states, where N denotes the length of the source sentence:

$$\mathbf{c_j} = \sum_{i=1}^{N} \alpha_{ji} \mathbf{h_i} \tag{5.9}$$

The α_{ji} values are the proportion of which the encoder hidden state vectors $\mathbf{h_{1...n}}$ contribute to the decoder hidden state when producing the jth token in the translation. They are computed by a feed-forward neural network, where $\mathbf{v_a}$, $\mathbf{W_a}$ and $\mathbf{U_a}$ are learned parameters:

$$\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_{l=1}^{N} \exp(e_{li})} \tag{5.10}$$

$$e_{ji} = \mathbf{v_a} \cdot \tanh(\mathbf{W_a} \cdot \mathbf{d_{j-1}} + \mathbf{U_a} \cdot \mathbf{h_i}) \tag{5.11}$$

From the hidden state $\mathbf{d_j}$ the network predicts the conditional distribution of the next token y_j, given a target language embedding $\mathbf{e_{j-1}}$ of the previous token, the current hidden state $\mathbf{d_j}$, and the calculated context vector $\mathbf{c_j}$. Note that at training time, y_{j-1} is the true observed token; whereas for unseen data we use the inferred token $\hat{y}_{j-1}$ sampled from the output of the softmax:

$$p(y_j|y_{<j}, c) = \text{softmax}(\tanh(\mathbf{e_{j-1}} + \mathbf{d_j} + \mathbf{c_j})) \tag{5.12}$$

The translation model is trained to minimise the negative log likelihood of predicting the target language output:

$$J_{NLL}(\theta, \phi^t) = -\sum_j \log \text{p}(y_j|y_{<j}, x) \tag{5.13}$$

5.3.3 Imaginet Decoder

The image prediction decoder is trained to predict the visual feature vector of the image associated with a sentence (Chrupała et al., 2015). It encourages the shared encoder to learn grounded representations for the source language.

A source language sentence is encoded using the Shared Encoder, as described in Section 5.3.1. Then we transform the shared encoder representation into a single vector by taking the mean pool over the hidden state annotations, the same way we initialise the hidden state of the translation decoder (Eqn. 5.8). This sentence representation is the input to a feed-forward neural network that predicts the visual feature vector $\hat{\mathbf{v}}$ associated with a sentence with parameters $\mathbf{W_{vis}}$:

$$\hat{\mathbf{v}} = \tanh(\mathbf{W_{vis}} \cdot \frac{1}{N} \sum_{i}^{N} \mathbf{h_i}) \tag{5.14}$$

This decoder is trained to predict the true image vector $\mathbf{v}$ with a margin-based objective, parameterised by the minimum margin α, and the cosine distance $d(\cdot, \cdot)$. A margin-based objective has previously been used in grounded representation learning (Vendrov et al., 2016; Chrupała et al., 2017). The contrastive examples $\mathbf{v}'$ are drawn from the other instances in a minibatch:

$$\begin{aligned} J_{MAR}(\theta, \phi^t) = \sum_{\mathbf{v}' \neq \mathbf{v}} \max\{0, \alpha - d(\hat{\mathbf{v}}, \mathbf{v}) \\ + d(\hat{\mathbf{v}}, \mathbf{v}')\} \end{aligned} \tag{5.15}$$

	Size	Tokens	Types	Images
Multi30K: parallel text with images				
En	31K	377K	10K	31K
De		368K	16K	
MS COCO: external described images				
En	414K	4.3M	24K	83K
News Commentary: external parallel text				
En	240K	8.31M	17K	–
De		8.95M		–

Table 5.1: The datasets used in our experiments.

5.4 Data

We evaluate our model using the benchmark Multi30K dataset (Elliott et al., 2016), which is the largest collection of images paired with sentences in multiple languages. This dataset contains 31,014 images paired with an English language sentence and a German language translation: 29,000 instances are reserved for training, 1,014 for development, and 1,000 for evaluation.[1]

The English and German sentences are preprocessed by normalising the punctuation, lowercasing and tokenizing the text using the Moses toolkit. We additionally decompound the German text using Zmorge Sennrich & Kunz (2014). This results in vocabulary sizes of 10,214 types for English and 16,022 for German.

[1]The Multi30K dataset also contains 155K independently collected descriptions in German and English. In order to make our experiments more comparable with previous work, we do not make use of this data.

We also use two external datasets to evaluate our model: the MS COCO dataset of English described images Chen et al. (2015), and the English-German News Commentary parallel corpus (Tiedemann, 2012). When we perform experiments with the News Commentary corpus, we first calculate a 17,597 sub-word vocabulary using SentencePiece (Schuster & Nakajima, 2012) over the concatentation of the Multi30K and News Commentary datasets. This gives us a shared vocabulary for the external data that reduces the number of out-of-vocabulary tokens.

Images are represented by 2048D vectors extracted from the `pool5/7x7_s1' layer of the GoogLeNet v3 CNN (Szegedy et al., 2015).

5.5 Experiments

We evaluate our multitasking approach with in- and out-of-domain resources. We start by reporting results of models trained using only the Multi30K dataset. We also report the results of training the IMAGINET decoder with the COCO dataset. Finally, we report results on incorporating the external News Commentary parallel text into our model. Throughout, we report performance of the En→De translation using Meteor (Denkowski & Lavie, 2014) and BLEU (Papineni et al., 2002) against lowercased tokenized references.

5.5.1 Hyperparameters

The encoder is a 1000D Gated Recurrent Unit bidirectional recurrent neural network (Cho et al., 2014b, GRU) with 620D embeddings. We share all of the encoder parameters between the primary and auxiliary task. The translation decoder is a 1000D GRU recurrent neural network, with a 2000D context vector over the encoder states, and 620D word embeddings (Sennrich et al., 2017). The Imaginet decoder is a single-layer feed-forward network, where we learn the parameters $\mathbf{W_{vis}} \in \mathbb{R}^{2048\text{x}2000}$ to predict the true image vector with $\alpha = 0.1$ for the Imaginet objective (Equation 5.15). The models are trained using the Adam optimiser with the default hyperparameters (Kingma & Ba, 2014) in minibatches of 80 instances. The translation task is defined as the primary task and convergence is reached when BLEU has not increased for five epochs on the validation data. Gradients are clipped when their norm exceeds 1.0. Dropout is set to 0.2 for the embeddings and the recurrent connections in both tasks (Gal & Ghahramani, 2016). Translations are decoded using beam search with 12 hypotheses.

5.5.2 In-domain experiments

We start by presenting the results of our multitask model trained using only the Multi30K dataset. We compare against state-of-the-art approaches and text-only baselines. Moses is the phrase-based machine translation model (Koehn et al., 2007) reported in Specia et al. (2016). NMT is a text-only neural machine translation model. Calixto et al. (2017a) is a double-attention model over the source language and the image. Calixto & Liu (2017a) is a multimodal translation model that conditions the decoder on semantic image vector extracted from the VGG-19 CNN. Hitschler et al. (2016) uses visual features in a target-side retrieval model for translation. Toyama et al. (2016) is most comparable to our approach: it is a multimodal variational NMT model that infers latent variables to represent the source

	Meteor	BLEU
NMT	54.0 ± 0.6	35.5 ± 0.8
Calixto et al. (2017a)	55.0	36.5
Calixto & Liu (2017a)	55.1	37.3
Imagination	55.8 ± 0.4	36.8 ± 0.8
Toyama et al. (2016)	56.0	36.5
Hitschler et al. (2016)	56.1	34.3
Moses	56.9	36.9

Table 5.2: En→De translation results on the Multi30K dataset. Our Imagination model is competitive with the state of the art when it is trained on in-domain data. We report the mean and standard deviation of three random initialisations.

language semantics from the image and linguistic data.

Table 5.2 shows the results of this experiment. We can see that the combination of the attention-based translation model and the image prediction model is a 1.8 Meteor point improvement over the NMT baseline, but it is 1.1 Meteor points worse than the strong Moses baseline. Our approach is competitive with previous approaches that use visual features as inputs to the decoder and the target-side reranking model. It also competitive with Toyama et al. (2016), which also only uses images for training. These results confirm that our multitasking approach uses the image prediction task to improve the encoder of the translation model.

	Meteor	BLEU
Imagination	55.8 ± 0.4	36.8 ± 0.8
Imagination (COCO)	55.6 ± 0.5	36.4 ± 1.2

Table 5.3: Translation results when using out-of-domain described images. Our approach is still effective when the image prediction model is trained over the COCO dataset.

	Meteor	BLEU
NMT	52.8 ± 0.6	33.4 ± 0.6
+ NC	56.7 ± 0.3	37.2 ± 0.7
+ Imagination	56.7 ± 0.1	37.4 ± 0.3
+ Imagination (COCO)	57.1 ± 0.2	37.8 ± 0.7
Calixto et al. (2017a)	56.8	39.0

Table 5.4: Translation results with out-of-domain parallel text and described images. We find further improvements when we multitask with the News Commentary (NC) and COCO datasets.

5.5.3 External described image data

Recall from Section 5.2 that we are interested in scenarios where x, y, and v are drawn from different sources. We now experiment with separating the translation data from the described image data using $\mathcal{D}_{image}$: MS COCO dataset of 83K described images[2] and $\mathcal{D}_{text}$: Multi30K parallel text.

[2]Due to differences in the vocabularies of the respective datasets, we do not train on examples where more than 10% of the tokens are out-of-vocabulary in the Multi30K dataset.

	Parallel text		Described images			
	Multi30K	News Commentary	Multi30K	COCO	Meteor	BLEU
Zmorge	✓				56.2	37.8
	✓		✓		57.6	39.0
Sub-word	✓				54.4	35.0
	✓	✓			58.6	39.4
	✓	✓	✓		59.0	39.5
	✓	✓		✓	**59.3**	**40.2**

Table 5.5: Ensemble decoding results. Zmorge denotes models trained with decompounded German words; Sub-word denotes joint SentencePiece word splitting (see Section 6.4 for more details).

Table 5.3 shows the results of this experiment. We find that there is no significant difference between training the IMAGINET decoder on in-domain (Multi30K) or out-of-domain data (COCO). This result confirms that we can separate the parallel text from the described images.

5.5.4 External parallel text data

We now experiment with training our model on a combination of the Multi30K and the News Commentary English-German data. In these experiments, we concatenate the Multi30K and News Commentary datasets into a single $\mathcal{D}_{text}$ training dataset, similar to Freitag & Al-Onaizan (2016). We compare our model against Calixto et al. (2017a), who pre-train their model on the WMT'15 English-German parallel text and back-translate (Sennrich et al., 2016) additional sentences from the bilingual independent descriptions in the Multi30K dataset (Footnote 2).

Table 5.4 presents the results. The text-only NMT model using sub-words is 1.2 Meteor points lower than decompounding the German text. Nevertheless, the model trained over a concatenation of the parallel texts is a 2.7 Meteor point improvement over this baseline (+ NC) and matches the performance of our Multitasking model that uses only in-domain data (Section 5.5.2). We do not see an additive improvement for the multitasking model with the concatenated parallel text and the in-domain data (+ Imagination) using a training objective interpolation of $w = 0.89$ (the ratio of the training dataset sizes). This may be because we are essentially learning a translation model and the updates from the IMAGINET decoder are forgotten. Therefore, we experiment with multitasking the concatenated parallel text and the COCO dataset ($w = 0.5$). We find that balancing the datasets improves over the concatenated text model by 0.4 Meteor (+ Imagination (COCO)). Our multitasking approach improves upon Calixto et al. by 0.3 Meteor points. Our model can be trained in 48 hours using 240K parallel sentences and 414K described images from out-of-domain datasets. Furthermore, recall that our model does not use images as an input for translating unseen data, which results in 6.2% fewer parameters compared to using the 2048D Inception-V3 visual features to initialise the hidden state of the decoder.

5.5.5 Ensemble results

Table 5.5 presents the results of ensembling different randomly initialised models. We achieve a start-of-the-art result of 57.6 Meteor for a model trained on only in-domain data. The improvements are more pronounced for the models trained using sub-words and out-of-domain data. An ensemble of baselines trained on sub-words is

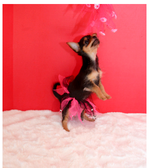
	Source:	two children on their stomachs lay on the ground under a pipe
	NMT:	zwei kinder auf ihren gesichtern liegen unter dem boden auf dem boden
	Ours:	zwei kinder liegen bäuchlings auf dem boden unter einer schaukel
	Source:	small dog in costume stands on hind legs to reach dangling flowers
	NMT:	ein kleiner hund steht auf dem hinterbeinen und läuft , nach links von blumen zu sehen
	Ours:	ein kleiner hund in einem kostüm steht auf den hinterbeinen , um die blumen zu erreichen
	Source:	a bird flies across the water
	NMT:	ein vogel fliegt über das wasser
	Ours:	ein vogel fliegt durch das wasser

Table 5.6: Examples where our model improves or worsens the translation compared to the NMT baseline. Top: NMT translates the wrong body part; both models skip “pipe”. Middle: NMT incorrectly translates the verb and misses several nouns. Bottom: Our model incorrectly translates the preposition.

initially worse than an ensemble trained on Zmorge decompounded words. However, we always see an improvement from ensembling models trained on in- and out-of-domain data. Our best ensemble is trained on Multi30K parallel text, the News Commentary parallel text, and the COCO descriptions to set a new state-of-the-art result of 59.3 Meteor.

5.5.6 Multi30K 2017 results

We also evaluate our approach against 16 submissions to the WMT Shared Task on Multimodal Translation and Multilingual Image Description (Elliott & Kádár, 2017). This shared task features a new evaluation dataset: Multi30K Test 2017 (Elliott & Kádár, 2017), which contains 1,000 new evaluation images. The shared task submissions are evaluated with Meteor and human direct assessment (Graham et al., 2017). We submitted two systems, based on whether they used only the Multi30K dataset (constrained) or used additional external resources (unconstrained). Our constrained submission is an ensemble of three Imagination models trained over only the Multi30K training data. This achieves a Meteor score of 51.2, and a joint 3rd place ranking according to human assessment. Our unconstrained submission is an ensemble of three Imagination models trained with the Multi30K, News Commentary, and MS COCO datasets. It achieves a Meteor score of 53.5, and 2nd place in the human assessment.

5.5.7 Qualitative examples

Table 5.6 shows examples of where the multitasking model improves or worsens translation performance compared to the baseline model[3]. The first example shows that the baseline model makes a significant error in translating the pose of the children, translating "on their stomachs" as "on their faces"). The middle example demonstrates that the baseline model translates the dog as walking ("läuft") and then makes grammatical and sense errors after the clause marker. Both models neglect to translate the word "dangling", which is a low-frequency word in the training data. There are instances where the baseline produces better translations than the multitask model: In the bottom example, our model translates a bird flying through the water ("durch") instead of "over" the water.

[3]We used MT-ComparEval (Klejch et al., 2015)

5.6 Discussion

5.6.1 Does the model learn grounded representations?

A natural question to ask if whether the multitask model is actually learning representations that are relevant for the images. We answer this question by evaluating the Imaginet decoder in an image–sentence ranking task. Here the input is a source language sentence, from which we predict its image vector $\hat{\mathbf{v}}$. The predicted vector $\hat{\mathbf{v}}$ can be compared against the true image vectors $\mathbf{v}$ in the evaluation data using the cosine distance to produce a ranked order of the images. Our model returns a median rank of 11.0 for the true image compared to the predicted image vector. Figure 5.2 shows examples of the nearest neighbours of the images predicted by our multitask model. We can see that the combination of the multitask source language representations and IMAGINET decoder leads to the prediction of relevant images. This confirms that the shared encoder is indeed learning visually grounded representations.

5.6.2 The effect of visual feature vectors

We now study the effect of varying the Convolutional Neural Network used to extract the visual features used in the Imaginet decoder. It has previously been shown that the choice of visual features can affect the performance of vision and language models (Jabri et al., 2016; Kiela et al., 2016). We compare the effect of training the IMAGINET decoder to predict different types of image features, namely: 4096D features extracted from the 'fc7'' layer of the VGG-19 model

(a) Nearest neighbours for "a native woman is working on a craft project ."

(b) Nearest neighbours for "there is a cafe on the street corner with an oval painting on the side of the building ."

Figure 5.2: We can interpret the IMAGINET Decoder by visualising the predictions made by our model.

(Simonyan & Zisserman, 2015), 2048D features extracted from the 'pool5/7x7_s1' layer of InceptionNet V3 (Szegedy et al., 2015), and 2048D features extracted from 'avg_pool' layer of ResNet-50 (He et al., 2016). Table 5.7 shows the results of this experiment. There is a clear difference between predicting the 2048D vectors (Inception-V3 and ResNet-50) compared to the 4096D vector from VGG-19). This difference is reflected in both the translation Meteor score and the Median rank of the images in the validation dataset. This is likely because it is easier to learn the parameters of the image prediction model that has fewer parameters (8.192 million for VGG-19 vs. 4.096 million for Inception-V3 and ResNet-50). However, it is not clear why there is such a pronounced difference between the Inception-V3 and ResNet-50 models[4].

[4] We used pre-trained CNNs (`https://github.com/fchollet/deep-learning-models`), which claim equal ILSVRC object recognition performance for both models: 7.8% top-5 error with a single-model and single-crop.

	Meteor	Median Rank
Inception-V3	56.0 ± 0.1	11.0 ± 0.0
Resnet-50	54.7 ± 0.4	11.7 ± 0.5
VGG-19	53.6 ± 1.8	13.0 ± 0.0

Table 5.7: The type of visual features predicted by the IMAGINET Decoder has a strong impact on the Multitask model performance.

5.7 Related work

Initial work on multimodal translation used semantic or spatially-preserving image features as inputs to a translation model. Semantic image features are typically extracted from the final layer of a pre-trained object recognition CNN, e.g. 'pool5/7x7_s1' in GoogLeNet (Szegedy et al., 2015). This type of vector has been used as input to the encoder Elliott et al. (2015); Huang et al. (2016), the decoder (Libovický et al., 2016), or as features in a phrase-based translation model (Shah et al., 2016; Hitschler et al., 2016). Spatially-preserving image features are extracted from deeper inside a CNN, where the position of a feature is related to its position in the image. These features have been used in "double-attention models", which calculate independent context vectors for the source language and a convolutional image features (Calixto et al., 2016; Caglayan et al., 2016; Calixto et al., 2017a). We use an attention-based translation model but our multitask model does not use images for translation.

More related to our work is an extension of Variational Neural Machine Translation to infer latent variables to *explicitly* model the semantics of source sentences from visual and linguistic information (Toyama et al., 2016). They report improvements on the Multi30K data set but their model needs additional parameters in the "neural inferrer" modules. In our model, the grounded semantics are represented *implicitly* in the shared encoder. They assume Source-Target-Image training data, whereas our approach achieves equally good results if we train on separate Source-Image and Source-Target datasets. Saha et al. (2016) study cross-lingual image description where the task is to generate a sentence in language L_1 given the image, using only Image-L_2 and L_1-L_2 training corpora. They propose a Correlational Encoder-Decoder to model the Image-L_2 and L_1-L_2 data, which learns correlated representations for paired Image-L_2 data and decodes L_1 from the joint representation. Similar to our work, the encoder is trained by minimizing two loss functions: the Image-L_2 correlation loss, and the L_1 decoding cross-entropy loss. Nakayama & Nishida (2017) consider a zero-resource problem, where the task is to translate from L_1 to L_2 with only Image-L_1 and Image-L_2 corpora. Their model embeds the image, L_1, and L_2 in a joint multimodal space learned by minimizing a multi-task ranking loss between both pairs of examples. In this paper, we focus on *enriching* source language representations with visual information instead of zero-resource learning.

Multitask Learning improves the generalisability of a model by requiring it to be useful for more than one task (Caruana, 1997). This approach has recently been used to improve the performance of sentence compression using eye gaze as an auxiliary task (Klerke et al., 2016), and to improve shallow parsing accuracy through the auxiliary task of predicting keystrokes in an out-of-domain corpus (Plank, 2016). More recently, Bingel & Søgaard (2017) analysed the beneficial relationships between primary and auxiliary sequential prediction tasks. In the translation literature, multitask learning has been used to learn a one-to-many languages translation model (Dong et al., 2015), a multi-lingual translation model with a single attention mechanism shared across multiple languages (Firat et al., 2016), and in multitask sequence-to-sequence learning without an attention-based decoder (Luong et al., 2016). We explore the benefits of grounded learning in the specific case of multimodal translation. We combine sequence prediction with continuous (image) vector prediction, compared to previous work which multitasks different sequence prediction tasks.

Visual representation prediction has been studied using static images or videos. Lin & Parikh (2015) use a conditional random field to imagine the composition of a clip-art scene for visual paraphrasing and fill-in-the-blank tasks. Chrupała et al. (2015) predict the image vector associated with a sentence using an L2 loss; they found this improves multi-modal word similarity compared to text-only baselines. Gelderloos & Chrupała (2016) predict the image vector associated with a sequence of phonemes using a max-margin loss, similar to our image prediction objective. Collell et al. (2017) learn to predict the visual feature vector associated with a word for word similarity and relatedness tasks. As a video reconstruction problem, Srivastava et al. (2015) propose an LSTM Autoencoder to predict video frames as a reconstruction task or as a future prediction task. Pasunuru & Bansal (2017) propose a multitask model for video description that combines unsupervised video reconstruction, lexical entailment, and video description. They find improvements from using out-of-domain resources for entailment and video prediction, similar to the improvements we find from using out-of-domain parallel text and described images.

5.8 Conclusion

We decompose multimodal translation into two sub-problems: learning to translate and learning visually grounded representations. In a multitask learning framework, we show how these sub-problems can be addressed by sharing an encoder between a translation model and an image prediction model[5]. Our approach achieves state-of-the-art results on the Multi30K dataset without using images for translation. We show that training on separate parallel text and described image datasets does not hurt performance, encouraging future research on multitasking with diverse sources of data. Furthermore, we still find improvements from image prediction when we improve our text-only baseline with the out-of-domain parallel text. Future work includes adapting our decomposition to other NLP tasks that may benefit from out-of-domain resources, such as semantic role labelling, dependency parsing, and question-answering; exploring methods for inputting the (predicted) image into the translation model; experimenting with different image prediction architectures; multitasking different translation languages into a single shared encoder; and multitasking in both the encoder and decoder(s).

Acknowledgments

We are grateful to the anonymous reviewers for their feedback. We thank Joost Bastings for sharing his multitasking Nematus model,

[5]Code: `http://github.com/elliottd/imagination`

Wilker Aziz for discussions about formulating the problem, Stella Frank for finding and explaining the qualitative examples to us, and Afra Alishahi, Grzegorz Chrupała, and Philip Schulz for feedback on earlier drafts of the paper. DE acknowledges the support of an Amazon Research Award, NWO Vici grant nr. 277-89-002 awarded to K. Sima'an, and a hardware grant from the NVIDIA Corporation.

6

Lessons learned in multilingual grounded language learning

abstract Recent work has shown how to learn better visual-semantic embeddings by leveraging image descriptions in more than one language. Here, we investigate in detail which conditions affect the performance of this type of grounded language learning model. We show that multilingual training improves over bilingual training, and that low-resource languages benefit from training with higher-resource languages. We demonstrate that a multilingual model can be trained equally well on either translations or comparable sentence pairs, and that annotating the same set of images in multiple language enables further improvements via an additional caption-caption ranking objective.

This chapter is based on Kádár, Á., Elliott, D., Côté, M. A., Chrupała, G., & Alishahi, A. (2018). Lessons Learned in Multilingual Grounded Language Learning. *In Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 402-412).

6.1 Introduction

Multimodal representation learning is largely motivated by evidence of perceptual grounding in human concept acquisition and representation (Barsalou et al., 2003). It has been shown that visually grounded word and sentence-representations (Kiela et al., 2014; Baroni, 2016; Elliott & Kádár, 2017; Kiela et al., 2017; Yoo et al., 2017) improve performance on the downstream tasks of paraphrase identification, semantic entailment, and multimodal machine translation (Dolan et al., 2004; Marelli et al., 2014b; Specia et al., 2016). Multilingual sentence representations have also been successfully applied to many-languages-to-one character-level machine translation (Chung et al., 2016) and multilingual dependency parsing (Ammar et al., 2016).

Recently, Gella et al. (2017) proposed to learn both bilingual and multimodal sentence representations using images paired with captions independently collected in English and German. Their results show that bilingual training improves image-sentence ranking performance over a monolingual baseline, and it improves performance on semantic textual similarity benchmarks Agirre et al. (2014, 2015). These findings suggest that it may be beneficial to consider another language as another *modality* in a monolingual grounded language learning model. In the grounded learning scenario, descriptions of an image in multiple languages can be considered as multiple views of the same or closely related data. These additional views can help overcome the problems of data sparsity, and have practical implications for efficiently collecting image-text datasets in different languages. In real-life applications, many tasks and domains can involve code switching (Barman et al., 2014), which is easier to deal with using a

[*]Work carried out at the University of Edinburgh.

En: A group of people are eating noodles.

De: Eine Gruppe von Leuten isst Nudeln.

Fr: Un groupe de gens mangent des nouilles.

Cs: Skupina lidí jedí nudle.

(a) A translation tuple

En: Several asian people eating around a table.

De: Drei Männer und zwei Frauen südostasiatischen Aussehens sitzen, aus Schälchen essend, an einem schwarzen, Tisch, auf dem sich u.a. auch Pappbecher und eine Tasche befinden, im Hintergrund sind weitere Personen und Tische.[1]

(b) A comparable pair

Figure 6.1: An example taken from the *Translation* and *Comparable* portions of the Multi30K dataset. The translation portion (a) contains professional translations of the English captions into German, French, and Czech. The comparable portion (b) consists of five independently crowdsourced English and German descriptions, given only the image. Note that the sentences in (b) convey different information from the English–German translation pair in (a).

multilingual model. Furthermore, it is more convenient to maintain a single multilingual system than one system for each considered language. However, there is a need for a systematic exploration of the conditions under which it is useful to add additional views of the data. We investigate the impact of the following conditions on the performance of a multilingual grounded language learning model in sentence and image retrieval tasks:

Additional languages. Multilingual models have not been explored yet in a multimodal setting. We investigate the contribution of adding more than one language by performing bilingual experiments on English and German (Section 6.5) as well as adding French and Czech captioned images (Section 6.6).

Data alignment: We assess the performance of a multilingual models trained using either captions that are translations of each other, or captions that are independently collected in different languages for the same set of images. The two scenarios are illustrated in Figure 6.1. Additionally we consider the setup when non-overlapping sets of images and their captions are collected in different languages. Such disjoint settings have been explored in pivot-based multimodal representation learning (Funaki & Nakayama, 2015; Rajendran et al., 2015) or zero-shot multi-modal machine translation (Nakayama & Nishida, 2017). We compare translated vs. independently collected captions in Sections 6.5.2 and 6.6.1, and overlapping vs. disjoint images in Section 6.5.3.

High-to-low resource transfer: In Section 6.6.2 we investigate whether low-resource languages benefit from jointly training on larger data sets from higher-resource languages. This type of transfer has previously been shown to be effective in machine translation (e.g., Zoph et al., 2016).

Training objective: In addition to learning to map images to sentences, we study the effect of also learning relationships between captions of the same image in different languages (Gella et al., 2017). We assess the contribution of such a caption–caption ranking objective throughout our experiments.

Our results show that multilingual joint training improves upon bilingual joint training, and that grounded sentence representations for a low-resource language can be substantially improved with data from different high-resource languages. Our results suggest that independently-collected captions are more useful than translated captions, for the task of learning multilingual multimodal sentence embeddings. Finally, we recommend to collect captions for the same set of images in multiple languages, due to the benefits of the additional caption–caption ranking objective function.

6.2 Related work

Learning visually grounded word-representations has been an active area of research in the fields of multi-modal semantics and cross-situational word-learning. Such perceptually-grounded word representations have been shown to lead to higher correlation with human judgements on word-similarity benchmarks such as WordSim353

[1]Gloss: Three men and two women with a South-East Asian appearance eat out of bowls at a black table, on which there are, among other things, paper cups and a bag; in the background there are other people and tables.

(Finkelstein et al., 2001) or SimLex999 (Hill et al., 2015) compared to uni-modal representations (Kádár et al., 2015a; Bruni et al., 2014; Kiela & Bottou, 2014).

Grounded representations of sentences that are learned from image–caption data sets also improve performance on a number of sentence-level tasks (Kiela et al., 2017; Yoo et al., 2017) when used as additional features to skip-thought vectors (Kiros et al., 2015). The model architectures used for these studies have the same overall structure as our model and coincide with image–sentence retrieval systems (Kiros et al., 2014; Karpathy & Fei-Fei, 2015): a pre-trained CNN is fixed or fine-tuned as image feature extractor, followed by a learned transformation, while sentence representations are learned by a randomly initialized recurrent neural network. These models are trained to push the true image–caption pairs closer together, and the false image–caption pairs further from each other, in a joint embedding space.

In addition to learning grounded representations for image-sentence ranking, joint vision and language systems have been proposed to solve a wide range of tasks across modalities such as image captioning (Mao et al., 2014a; Vinyals et al., 2015b; Xu et al., 2015), visual question answering (Antol et al., 2015; Fukui et al., 2016; Jabri et al., 2016), text-to-image synthesis (Reed et al., 2016) and multimodal machine translation (Libovicky & Helcl, 2017; Elliott & Kádár, 2017).

Our work is also closely related to multilingual joint representation learning. In this scenario, a single model is trained to solve a task across multiple languages. Ammar et al. (2016) train a multilingual dependency parser on the Universal Dependencies treebank (Nivre et al., 2015) and show that on average the single multilingual model outperforms the monolingual baselines. Johnson et al. (2016) present a zero-shot neural machine translation model that is jointly trained on language pairs $A \leftrightarrow B$ and $B \leftrightarrow C$ and show that the model is capable of performing well on the unseen language pair $A \leftrightarrow C$. Lee et al. (2017a) find that jointly training a many-languages-to-one translation model on unsegmented character sequences improves BLEU scores compared to monolingual training. They also show evidence that the model can handle intra-sentence code-switching. Peters et al. (2017) train a multilingual sequence-to-sequence translation architecture on grapheme-to-phoneme conversion using more than 300 languages. They report better performance when adding multiple languages, even those which are not present in the test data. Finally, massively multilingual language representations trained on over 900 languages have been shown to resemble language families (Östling & Tiedemann, 2016) and can successfully predict linguistic typology features (Malaviya et al., 2017).

In the vision and language domain, multilingual-multimodal sentence representation learning has been limited so far to two languages. The joint training of models on English and German data has been shown to outperform monolingual baselines on image-sentence ranking and semantic textual similarity tasks (Gella et al., 2017; Calixto et al., 2017b). Recently (Harwath et al., 2018) also showed the benefit of joint bilingual training in the domain of speech-to-image and image-to-speech retrieval using English and Hindi data.

Require: p: task switching probability.
$\mathcal{D}_{c2i}$: datasets $D_1 \ldots D_k$ of image-caption pairs
$< c, i >$ for all k languages.
D_{c2c}: data set of all possible caption pairs
$< c_a, c_b >$ for all k languages.
$\phi(c, \theta_\phi)$: caption encoder
$\psi(i, \theta_\phi)$: image encoder

while not stopping criterion **do**
  $T \sim \mathrm{Bern}(p)$
  if $T = 1$ **then**
    $D_n \sim \mathcal{D}_{c2i}$
    $< c, i > \sim D_n$
    $\mathbf{a} \leftarrow \phi(c, \theta_\phi)$
    $\mathbf{b} \leftarrow \psi(i, \theta_\psi)$
  else
    $< c_a, c_b > \sim D_{c2c}$
    $\mathbf{a} \leftarrow \phi(c_a, \theta_\phi)$
    $\mathbf{b} \leftarrow \phi(c_b, \theta_\phi)$
  end if
  $[\theta_\phi; \theta_\psi] \leftarrow \mathrm{SGD}(\nabla_{[\theta_\phi;\theta_\psi]} \mathcal{J}(\mathbf{a}, \mathbf{b}))$
end while

Figure 6.2: Pseudo-code of the training procedure used to train our multilingual multi-task model.

6.3 Multilingual grounded learning

We train a standard model of grounded language learning which projects images and their textual descriptions into the same space (Kiros et al., 2014; Karpathy & Fei-Fei, 2015). The training procedure is illustrated by the pseudo-code in Figure 6.2. Images i are encoded by a fixed pre-trained CNN followed by a learned affine transformation $\psi(i, \theta_\psi)$, and captions c are encoded by a randomly initialized RNN $\phi(c, \theta_\phi)$. The model learns to minimize the distance between pairs $<a, b>$ using a max-of-hinges ranking loss (Faghri et al., 2017):

$$\begin{aligned}\mathcal{J}(a, b) = \max_{<\hat{a}, b>}&[\max(0, \alpha - s(a, b) + s(\hat{a}, b))] + \\ &\max_{<a, \hat{b}>}[\max(0, \alpha - s(a, b) + s(a, \hat{b}))]\end{aligned}$$

where $< a, b >$ are the true pairs, and $< a, \hat{b} >$ and $< \hat{a}, b >$ are all possible contrastive pairs in the mini-batch. The pairs either consists of image-caption pairs $< i, c >$, where the model solves a caption-image **c2i** ranking task, or pairs of captions in multiple languages belonging to the same image $< c_a, c_b >$, where the model solves a caption-caption **c2c** ranking task (Gella et al., 2017). Our monolingual models are trained to minimize the caption-image ranking objective **c2i** on the training set. The multilingual models are trained to minimize the ranking loss for the set of all languages $\mathcal{L}$ in the collection: at each iteration the model is either updated for the **c2i** objective or the caption-caption **c2c** objective given either $< c^l, i >$ or a $< c_a^k, c_b^m >$ pair in languages $l, k, m, \ldots \in \mathcal{L}$. All models are trained by first selecting a task, either **c2i** or **c2c**. In the **c2i** case, a language is sampled at random followed by sampling a random batch;

in the **c2c** case, all possible $< c_a, c_b >$ pairs across all languages are treated as a single data set. All of the model parameters are shared across all tasks and languages.

Implementation. We build our model on the PyTorch implementation[2] of the VSE++ model (Faghri et al., 2017). Images are represented by the 2048D average-pool features extracted from the ResNet50 architecture (He et al., 2016) trained on ImageNet (Deng et al., 2009); this is followed by a trained linear layer $\mathbf{W}_I \in \mathbb{R}^{2048\times1024}$. Other implementation details follow (Faghri et al., 2017): sentences are represented as the final hidden state of a GRU (Chung et al., 2014) with 1024 units and 300 dimensional word-embeddings trained from scratch. We use a single word embedding matrix containing the union of all words in all considered languages. The similarity function s in the ranking loss is cosine similarity. We ℓ_2 normalize both the caption and image representations. The model is trained with the Adam optimizer (Kingma & Ba, 2014) using default parameters and learning-rate of 2e-4. We train the model with an early stopping criterion, which is to maximise the sum of the image–sentence recall scores R@1, R@5, R@10 on the validation set with patience of 10 evaluations. In the monolingual setting the stopping criterion is evaluated at the end of each epoch, whereas in the multilingual setup it is evaluated every 500 iterations. The probability of switching between the **c2i** and **c2c** tasks is set to 0.5. Batches from all data sets are sampled by shuffling the full dataset, going through each batch and re-shuffling when exhausted. The sentence-pair dataset used to train the **c2c** ranking model for ℓ languages is generated as follows.

[2]Code to reproduce our results is available at `https://github.com/kadarakos/mulisera`.

	En	De	Fr	Cz
En	1.0	0.04	0.06	0.02
De	–	1.0	0.03	0.01
Fr	–	–	1.0	0.01
Cz	–	–	–	1.0

Table 6.1: Vocabulary overlap as measured by the Jaccard coefficient between the different languages on the translation portion of the Multi30K dataset.

For a given image i, a set of languages $1 \cdots \ell$, and a set of captions $C^i_1, \ldots, C^i_\ell$ associated with an image i, we generate the set of all possible combinations of size 2 from caption sets $\mathcal{C}^i$ and add the Cartesian product between all resulting pairs $C^i_m \times C^i_n$ in $\mathcal{C}^i$ to the training set.

6.4 Experimental setup

Datasets. We train and evaluate our models on the *translation* and *comparable* portions of the Multi30K dataset (Elliott et al., 2016, 2017). The translation portion (a low-resource dataset) contains 29K images, each described in one English caption with German, French, and Czech translations. The comparable portion (a higher-resource dataset) contains the same 29K images paired with five English and five German descriptions collected independently. Figure 6.1 presents an example of the translation and comparable portions of the data. We used the preprocessed version of the dataset, in which the text is lowercased, punctuation is normalized, and the text is tokenized[3]. To reduce the vocabulary size of the joint models, we replace all words

[3]https://github.com/multi30k/dataset

occurring fewer than four times with a special "UNK" symbol. Table 6.1 shows the overlap between the vocabularies of the *translation* portion of the Multi30K dataset. The total number of tokens across all four languages is 17,571, and taking the union of the tokens in these four languages results in vocabulary of 16,553 tokens – a 6% reduction in vocabulary size. On the *comparable* portion of the dataset, the total vocabulary between English and German contains 18,337 tokens, with a union of 17,667, which is a 4% reduction in vocabulary size.

Evaluation. We evaluate our models on the 1K images of the 2016 test set of Multi30K either using the 5K captions from the comparable data or the 1K translation pairs. We evaluate on image-to-text (I→ T) and text-to-image (T→ I) retrieval tasks. For most experiments we report Recall at 1 (R@1), 5 (R@5) and 10 (R@10) scores averaged over 10 randomly initialised models. However, in Section 6.6 we only report R@10 due to space limitations and because it has less variance than R@1 or R@5.

6.5 Bilingual Experiments

6.5.1 Reproducing Gella et al. (2017)

We start by attempting to reproduce the findings of Gella et al. (2017). In these experiments we train our multi-task learning model on the *comparable* portion of Multi30K. Our models re-implement their setups used for `VSE` (Monolingual) and bilingual models `Pivot-Sym` (Bilingual) and `Parallel-Sym` (Bilingual + c2c). The `OE`, `Pivot-Asym` and `Parallel-Asym` models are trained using the asymmetric simi-

		I→T			T→I		
		R@1	R@5	R@10	R@1	R@5	R@10
Symmetric	`VSE`	31.6	60.4	72.7	23.3	53.6	**65.8**
	`Pivot-Sym`	31.6	61.2	73.8	23.5	53.4	**65.8**
	`Parallel-Sym`	**31.7**	**62.4**	**74.1**	**24.7**	**53.9**	65.7
Asymmetric	`OE`	**34.8**	**63.7**	74.8	25.8	**56.5**	67.8
	`Pivot-Asym`	33.8	62.8	**75.2**	26.2	56.4	**68.4**
	`Parallel-Asym`	31.5	61.4	74.7	**27.1**	56.2	66.9
	Monolingual	42.4	69.9	79.8	30.5	57.8	67.9
	Bilingual	42.7	70.7	80.1	30.6	58.1	68.3
	+ c2c	**43.8**	**71.8**	**81.4**	**32.3**	**59.9**	**70.2**

Table 6.2: English Image-to-text (I→T) and text-to-image (T→I) retrieval results on the *comparable* part of Multi30K, measured by Recall at 1, 5 at 10. `Typewriter` font shows performance of two sets of symmetric and asymmetric models from Gella et al. (2017).

larity measure introduced for the order-embeddings Vendrov et al. (2016). The main differences between our models and (Gella et al., 2017) is that they use VGG-19 image features, whereas we use ResNet50 features, and we use the max-of-hinges loss instead of the more common sum-of-hinges loss.

Table 6.2 shows the results on the English comparable 2016 test set. Overall our scores are higher than (Gella et al., 2017), which is most likely due to the different image features ((Faghri et al., 2017) also report a large performance gain when they use the ResNet instead of the VGG image features). Nevertheless, our results show a similar trend to the symmetric cosine similarity models from (Gella

		I→T			T→I		
		R@1	R@5	R@10	R@1	R@5	R@10
Symmetric	`VSE`	**29.3**	**58.1**	**71.8**	20.3	**47.2**	**60.1**
	`Pivot-Sym`	26.9	56.6	70.0	20.3	46.4	59.2
	`Parallel-Sym`	28.2	57.7	71.3	**20.9**	46.9	59.3
Asymmetric	`OE`	26.8	57.5	70.9	21.0	48.5	60.4
	`Pivot-Asym`	28.2	**61.9**	**73.4**	**22.5**	49.3	61.7
	`Parallel-Asym`	**30.2**	60.4	72.8	21.8	**50.5**	**62.3**
	Monolingual	34.2	63.0	74.0	23.9	49.5	60.5
	Bilingual	35.2	64.3	75.3	24.6	50.8	62.0
	+ c2c	**37.9**	**66.1**	**76.8**	**26.6**	**53.0**	**64.0**

Table 6.3: German Image-to-text (I→T) and text-to-image (T→I) retrieval results on the *comparable* part of Multi30K, measured by Recall at 1, 5 at 10. `Typewriter` font shows performance of two sets of symmetric and asymmetric models from Gella et al. (2017).

et al., 2017): our best results are achieved with bilingual joint training with the added c2c objective. Their models trained with an asymmetric similarity measure show a different trend: the monolingual model is stronger than the bilingual model, and the c2c loss provides no clear improvement.

Table 6.3 presents the German results. Once again, our implementation outperforms Gella et al. (2017), and this is likely due to the different visual features and max-of-hinges loss. However, our Bilingual model with the additional c2c objective performs the best for German, whereas Gella et al. (2017) reports the overall best results for the monolingual baseline `VSE`. Their models that use the asym-

	English		German	
	I→T	T→I	I→T	T→I
Monolingual	56.3	40.1	39.5	20.9
Bi-translation	67.4	55.1	58.3	44.6
+ c2c	58.2	47.7	51.0	39.6
Bi-comparable	**67.9**	55.7	**62.0**	48.1
+ c2c	67.6	**56.0**	61.9	**49.1**

Table 6.4: R@10 retrieval results on the *comparable* part of Multi30K. Bi-translation is trained on 29K *translation pair* data; bi-comparable is trained by downsampling the *comparable* data to 29K.

metric similarity function are clearly better than the Monolingual `OE` model. In general, the results from Gella et al. (2017) indicate the benefits of bilingual joint training, however, they do not find a clear pattern between the model configurations across languages. In our implementation, we only focused on the symmetric cosine similarity function and found a systematic pattern across both languages: bilingual training improves results on all performance metrics for both languages, and the additional c2c objective always provides further improvements.

6.5.2 Translations vs. independent captions

We now study whether the model can be trained on either translation pairs or independently collected bilingual captions. Gella et al. (2017) only conducted experiments on independently collected captions. However, it is known that humans have equally strong preference for translated or independently collected captions of images (Frank et al., 2018), which has implications for the difficulty and cost of collecting training data. Our baseline is a Monolingual model trained on 29K single-captioned images in the *translation* portion of Multi30K. The Bi-translation model is trained on both German and English, with shared parameters. Table 6.4 shows that there is a substantial improvement in performance for both languages in the bilingual setting. However, the additional c2c loss degrades performance here. This could be because we only have one caption per image in each language and it is easier to find a relationship between these views of the translation pairs.

In the Bi-comparable setting, we randomly select an English and a German sentence for each image in the *comparable* portion of Multi30K. We only find a minor difference in performance between the Bi-translation and Bi-comparable models for English, but the German results are improved. Crucially, it is still better than training on monolingual data. In the Bi-comparable setting, the c2c loss does not have a detrimental effect on model performance, unlike in the Bi-translation experiment. Overall we find that the *comparable* data leads to larger improvements in retrieval performance.

6.5.3 Overlapping vs. non-overlapping images

In a bilingual setting, we can improve an image-sentence ranking model by collecting more data in a second language. This can be achieved in two ways: by collecting captions in a new language for the same overlapping set of images, or by using a disjoint set of images and captions in a new language. We compare these two settings here.

In the Bi-overlap condition, we collect captions for the existing

	English		German	
	I→T	T→I	I→T	T→I
Full Monolingual	79.8	67.9	74.0	60.5
Half Monolingual	73.7	61.6	66.4	53.9
Bi-overlap	73.6	62.2	67.6	54.9
+ c2c	**76.0**	**65.9**	**71.2**	**59.1**
Bi-disjoint	73.1	62.1	67.9	54.9

Table 6.5: R@10 retrieval results on the *comparable* part of Multi30K. Full model trained on the 29K images of the *comparable* part, Half model on 14.5K images using random downsampling. For Bi-overlap, both English and German captions are used for 14.5K images. For Bi-disjoint, 14.5K images are used for English and the remaining 14.5K images for German.

images in a new language, i.e. we use all of the English and German captions paired with a random selection of 50% of the images in *comparable* Multi30K. This results in a training dataset of 14.5K images with 145K bilingual captions. In the Bi-disjoint condition, we collect captions for new images in a new language, i.e. we use all of the English captions from a random selection of 50% of the images, and all of the German captions for the remaining 50% of the images. This results in a training dataset on 29K images with a total of 145K bilingual captions.

Table 6.5 shows the results of this experiment. The upper-bound is to train a Monolingual model on the full *comparable* corpus. For the lower bound, we train Half Monolingual models by randomly sampling half of the 29K images and their associated captions, giving 72.5K captions over 14.5K images. Unsurprisingly, the Half Monolingual models perform worse than the Full Monolingual models. In the Bi-overlap experiment, the German model is improved by collecting captions for the existing images in English. There is no difference in the performance of the English model, echoing the results from Section 6.5.1. The Bi-overlap model also benefits from the added c2c objective. Finally, the Bi-disjoint model performs as well as the Bi-overlap model without the c2c objective. (It was not possible to train the Bi-disjoint model with the additional c2c objective because there are no caption pairs for the same image.)

Overall, these results suggest that it is best to collect additional captions in the original language, but when adding a second language, it is better to collect extra captions for existing images and exploit the additional c2c ranking objective.

6.6 Multilingual experiments

We now turn out attention to multilingual learning using the English, German, French and Czech annotations in the *translation* portion of Multi30K. We only report the text-to-image (T→I) R@10 results due to space limitations.

We did not repeat the overlapping vs. non-overlapping experiments from Section 6.5.3 in a multilingual setting because this would introduce too much data sparsity. In order to conduct this experiment, we would have to downsample the already low-resource French and Czech captions by 50%, or even further for multi-way experiments.

6.6.1 Translation vs. independent captions

Table 6.6 shows the results of repeating the translations vs. comparable captions experiment from Section 6.5.2 with data in four languages. The Multi-translation models are trained on 29K images paired with a single caption in each language. These models perform better than their Monolingual counterparts, and the German, French, and Czech models are further improved with the c2c objective. The Multi-comparable models are trained by randomly sampling one English and one German caption from the *comparable* dataset, alongside the French and Czech translation pairs. These models perform as well as the Multi-translation models, and the c2c objective brings further improvements for all languages in this setting.

These results clearly demonstrate the advantage of jointly training on more than two languages. Text-to-image retrieval performance increases by more than 11 R@10 points for each of the four languages in our experiment.

6.6.2 High-to-low resource transfer

We now examine whether the lower-resource French and Czech models benefit from training with the full complement of the higher-resource English and German comparable data. Therefore we train a joint model on the *translation* as well as *comparable* portions of Multi30K, and examine the performance on French and Czech.

Table 6.7 shows the results of this experiment. We find that the French and Czech models improve by 8.8 and 5.5 R@10 points respectively when they are only trained on the multilingual translation pairs (compared to the monolingual version), and by another 2.2 and 2.8

	En	De	Fr	Cz
Monolingual	50.4	39.5	47.0	42.0
Multi-translation	58.7	51.2	57.0	51.0
+ c2c	56.3	52.2	55.0	51.6
Multi-comparable	59.2	49.6	57.2	50.8
+ c2c	**61.8**	**52.7**	**59.2**	**55.2**

Table 6.6: The Monolingual and joint Multi-translation models trained on *translation pairs*, and the Multi-comparable trained on the downsampled *comparable* set with one caption per image.

points if trained on the extra 155K English and German *comparable* descriptions. We also find that the additional c2c objective improves the Czech model by a further 4.8 R@10 points (this improvement is likely caused by training the model on 46 possible caption pairs). Our results show the impact of jointly training with the larger English and German resources, which demonstrates the benefits of high-to-low resource transfer.

6.6.3 Bilingual vs. multilingual

Finally, we investigate how useful it is to train on four languages instead of two. Figure 6.3 presents the image-to-text and text-to-image retrieval results of training Monolingual, Bilingual, or Multilingual models. The Monolingual and Bilingual models are trained on a random single-caption-image subsample of the *comparable* dataset with the additional c2c objective, as this configuration provided the overall best results in Sections 6.5.2 and 6.6.1. The Multilingual models

	French	Czech
Monolingual	47.0	42.0
Multilingual	56.3	51.3
+ Comparable	58.9	52.4
+ c2c	**61.6**	**57.2**

Table 6.7: Multilingual is trained on all *translation pairs*, + Comparable adds the *comparable* data set.

are trained with the additional French and Czech *translation* data. As can be seen in Figure 6.3, the performance on both tasks and for both languages improves as we move from using data from one to two to four languages.

6.7 Conclusions

We learn multilingual multimodal sentence embeddings and show that multilingual joint training improves over bilingual joint training. We also demonstrate that low-resource languages can benefit from the additional data found in high-resource languages. Our experiments suggest that either translation pairs or independently-collected captions improve the performance of a multilingual model, and that the latter data setting provides further improvements through a caption–caption ranking objective. We also show that when collecting data in an additional language, it is better to collect captions for the existing images because we can exploit the caption–caption objective. Our results lead to several directions for future work. We would like to pin down the mechanism via which multilingual training contributes to

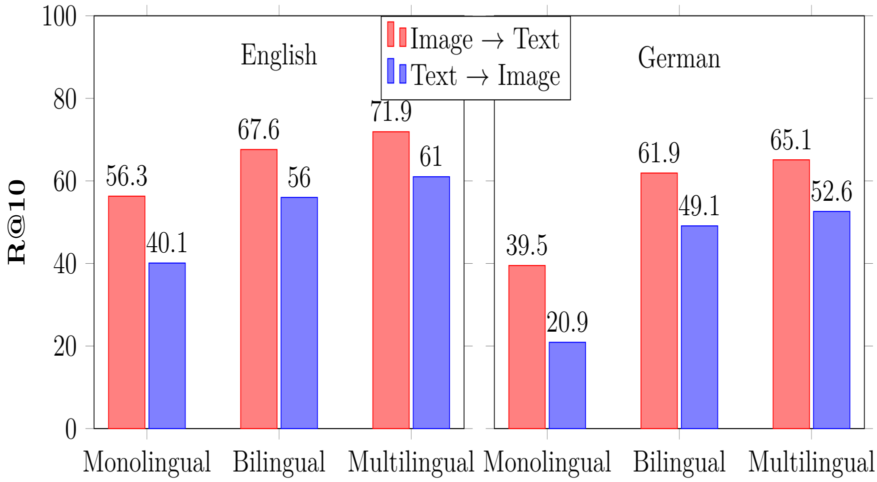


Figure 6.3: Comparing models from the Monolingual, Bilingual and Multilingual settings. The Monolingual and Bilingual models are trained on the downsampled English and German *comparable* sets with additional c2c objective. The Multilingual model uses the French and Czech *translation pairs* as additional data. The results are reported on the full 2016 test set of the *comparable* portion of Multi30K.

improved performance for image-sentence ranking. Additionally, we only consider four languages and show the gain of multilingual over bilingual training only for the English-German language pair. In future work we will incorporate more languages from data sets such as the Chinese Flickr8K (Li et al., 2016c) or Japanese COCO (Miyazaki & Shimizu, 2016).

Acknowledgements

Desmond Elliott was supported by an Amazon Research Award.

7

Synthetic Pairs for Multilingual Grounded Language Learning

abstract Recent work has highlighted the advantage of jointly learning grounded sentence representations from multiple languages. However, the data used in these studies has been limited to an *aligned* scenario, in which the same images are annotated with sentences in multiple languages. In this paper, we focus on the more realistic *disjoint* scenario using English and German image–caption data sets where the images do not overlap between languages. We observe that training with aligned data provides larger gains than training with disjoint data, which may be caused by the lack of coherence between the disjoint data sets. To address the lack of coherence, we propose a *pseudopairing* method, in which we create *synthetically aligned* English–German–Image triplets from the disjoint sets. The

pseudopairs are created in a two-step process: first we train a model on the *disjoint* data, then we create novel pairs across data sets using sentence similarity under the learned model. Experiments show that the pseudopair method improves image–sentence retrieval performance, despite requiring no external training data or models. We do find, however, that using an external machine translation model to generate the synthetic data sets results in better performance.

This chapter is based on Kádár, Á., Elliott, D., Chrupała, G., & Alishahi, A. (2019). Synthetic Pairs for Multilingual Grounded Language Learning. *Submitted to the 2019 Conference on Empirical Methods in Natural Language Processing.*

7.1 Introduction

The perceptual-motor system plays an important role in concept acquisition and representation, and in learning the meaning of linguistic expressions (Pulvermüller, 2005). In natural language processing, many approaches have been proposed that integrate visual information in the learning of word and sentence representations, highlighting the benefits of visually grounded representations (Lazaridou et al., 2015; Baroni, 2016; Kiela et al., 2017; Elliott & Kádár, 2017). In these approaches the visual world is taken as a naturally occurring meaning representation for linguistic utterances, grounding language in perceptual reality.

Recent work has shown that we can learn better visually grounded representations for sentences by jointly training image–sentence ranking models on multiple languages (Gella et al., 2017; Calixto & Liu, 2017b; Kádár et al., 2018). This line of research has focused on training models on datasets where the same images are annotated with sentences in multiple languages. This *alignment* has either been in the form of the translation pairs (e.g. the German, English, French, and Czech data in Multi30K (Elliott et al., 2016)) or independently collected sentences (English and Japanese captions in STAIR (Yoshikawa et al., 2017)).

In this paper, we consider the problem of training an image–sentence ranking model using image-caption collections in different languages with non-overlapping images drawn from different sources. We call these collections *disjoint* datasets, as opposed to *aligned* datasets. Kádár et al. (2018) showed that a model trained on disjoint datasets performs on-par with a model trained on aligned data. However, the disjoint datasets in their paper are artificial because they were formed by randomly splitting the Multi30K dataset into two halves. We examine whether the ranking model can benefit from multilingual supervision when it is trained using disjoint datasets drawn from different sources. In experiments with the Multi30K and COCO datasets, we find substantial benefits from training with these disjoint sources, but the best performance comes from training on aligned datasets.

Given the empirical benefits of training on aligned datasets, we explore two approaches to creating synthetically aligned training data in the *disjoint* scenario. One approach to creating synthetically aligned data is to use an off-the-shelf machine translation system to generate new image-caption pairs by translating the original captions. This approach is very simple, but has the limitation that an external system needs to be trained, which requires additional data.

The second approach is to generate synthetically aligned data that are *pseudopairs*. We assume the existence of image–caption datasets in different languages where the images do not overlap between the datasets. Pseudopairs are created by annotating the images of one dataset with the captions from another dataset. This can be achieved by leveraging the sentence similarities predicted by an image-sentence ranking model trained on the original image–caption datasets. One advantage of this approach is that does not require additional models or datasets because it uses the trained model to create new pairs. The resulting pseudopairs can then be used to re-train or fine-tune the original model.

In experiments on the Multi30K and COCO datasets, we find that the pseudopair approach consistently improves performance compared to training on the disjoint datasets, but the improvements less than using the translation approach. Nevertheless, the results show the potential for model-based data synthesis to improve the learning of visually grounded sentence representations.

7.2 Method

We adopt the model architecture and training procedure of Kádár et al. (2018) for the task of matching images with sentences. This task is defined as learning to rank the sentences associated with an image higher than other sentences in the data set, and vice-versa (Hodosh et al., 2013). The model is comprised of a recurrent neural network language model and a convolutional neural network image encoder. The parameters of the language encoder are randomly initialized, while the image encoder pre-trained, frozen during training and followed by a linear layer which is tuned for the task. The model is trained to make true pairs $< a, b >$ similar to each other, and contrastive pairs $< \hat{a}, b >$ and $< a, \hat{b} >$ dissimilar from each other in a joint embedding space by minimizing the max-violation loss function (Faghri et al., 2017):

$$\begin{aligned}\mathcal{J}(a, b) = &\max_{<\hat{a},b>}[\max(0, \alpha - s(a, b) + s(\hat{a}, b))] + \\ &\max_{<a,\hat{b}>}[\max(0, \alpha - s(a, b) + s(a, \hat{b}))]\end{aligned} \tag{7.1}$$

In our experiments, the $< a, b >$ pairs are either image-caption pairs $< i, c >$ or caption–caption pairs $< c_a, c_b >$ (following Gella et al. (2017); Kádár et al. (2018)). When we train on $< i, c >$ pairs, we sample a batch from an image–caption data set with uniform probability, encode the images and the sentences, and perform an update of the model parameters. For the caption–caption objective, we follow

Kádár et al. (2018) and generate a sentence pair data set by taking all pairs of sentences that belong to the same image and are written in different languages: 5 English and 5 German captions result in 25 English-German pairs. The sentences are encoded and we perform an update of the model parameters using the same loss. When training with both the image–caption and caption–caption (c2c) ranking tasks, we randomly select the task to perform with probability p=0.5.

7.2.1 Generating Synthetic Pairs

We propose two approaches to creating synthetic image–caption pairs to improve image–sentence ranking models when training with disjoint data sets. We assume the existence of datasets $\mathcal{D}_1$: $\mathcal{I}^1$-$\mathcal{C}^{\ell_1}$ and $\mathcal{D}_2$: $\mathcal{I}^2$-$\mathcal{C}^{\ell_2}$ consisting of image–caption pairs $< i_i^1, c_i^{\ell_1} >$ and $< i_i^2, c_i^{\ell_2} >$ in languages ℓ_1 and ℓ_2, where the image sets do not overlap $\mathcal{I}^1 \cap \mathcal{I}^2 = \varnothing$. We seek to extend $\mathcal{I}^2$-$\mathcal{C}^{\ell_2}$ to a bilingual dataset with synthetic captions $\hat{c}_i^{\ell_1} \in \hat{\mathcal{C}}^{\ell_1}$ in language ℓ_1, resulting in a triplet data set $\mathcal{I}^2$-$\hat{\mathcal{C}}^{\ell_1}$-$\mathcal{C}^{\ell_2}$ consisting of triplets $< i_i^2, \hat{c}_i^{\ell_1}, c_i^{\ell_2} >$. We hypothesize that the new dataset will improve model performance because it will be trained to map the images to captions in both languages.

7.2.2 Pseudopairs approach

Given two image-caption corpora $\mathcal{I}^1$-$\mathcal{C}^{\ell_1}$ and $\mathcal{I}^2$-$\mathcal{C}^{\ell_2}$ with pairs $< i_i^1, c_i^{\ell_1} >$ and $< i_i^2, c_i^{\ell_2} >$, we generate a *pseudopair* corpus labeling each image in $\mathcal{I}^2$ with a caption from $\mathcal{C}^{\ell_1}$. In our experiments, we create pseudopairs in only in one direction leading to new image–caption pairs $< i^2, \hat{c}^{\ell_1} >$.

The pseudopairs are generated using the sentence representations of the model trained on both corpora $\mathcal{I}^1$-$\mathcal{C}^{\ell_1}$ and $\mathcal{I}^2$-$\mathcal{C}^{\ell_2}$ jointly. We encode all captions $c_i^{\ell_1} \in \mathcal{C}^{\ell_1}$ and $c_i^{\ell_2} \in \mathcal{C}^{\ell_2}$ and for each $c_i^{\ell_2}$ find the most similar caption $\hat{c}_i^{\ell_1}$ using the cosine similarity between the sentence representations. This leads to pairs $< c_i^{\ell_2}, \hat{c}_i^{\ell_1} >$ and as a result to triplets $< i_i^2, c_i^{\ell_2}, \hat{c}_i^{\ell_1} >$.

Filtering Optionally we filter the resulting pseudopair set $\mathcal{C}^{\ell_1}$, in an attempt to avoid misleading samples with three filtering strategies:

1. No filtering.
2. Keep top: keep items with similarity scores in the 75% percentile; keep top 25%
3. Remove bottom: keep items with similarity scores in the 25%; remove bottom 25%

Fine-tuning vs. restart After the pseudopairs are generated we consider two options: re-train the model from scratch with all previous data sets adding the generated pseudopairs or fine-tunening with same data sets and the additional pseudopairs.

7.2.3 Translation approach

Given a corpus $\mathcal{I}^2$-$\mathcal{C}^{\ell_2}$ with pairs $< i_i^2, c_i^{\ell_2} >$, we use a machine translation system to and translate each caption $c_i^{\ell_2}$ to a language ℓ_1 leading to new image–caption pairs $< i_i^2, \hat{c}_i^{\ell_1} >$. [1] Any off-the-shelf translation system could be used to create the translated captions, e.g. an online service, such as Google Translate, or a pre-trained translation model. Here, we use a pre-trained model as it facilitates reproduction.

7.3 Experimental Protocol

7.3.1 Model

Our implementation, training protocol and parameter settings are based on the existing codebase of Kádár et al. (2018). In all experiments, we use the 2048 dimensional image-features extracted from the last average-pooling layer of a pre-trained[2] ResNet50 CNN (He et al., 2016).

The image representation used in our model is obtained by a single affine transformation that we train from scratch $\mathbf{W}_I \in \mathbb{R}^{2048\times 1024}$. For the sentence encoder we use a uni-directional Gated Recurrent Unit (GRU) network (Cho et al., 2014a) with a single hidden layer with 1024 hidden units and 300 dimensional word embeddings. When training bilingual models we use a single word-embedding for the same word-forms, making no distinction if they come from different languages. Each sentence is represented by the final hidden state of the GRU. For the similarity function in the loss function (Eq. 1) we use cosine similarity and $\alpha = 0.2$ margin parameter.

[1] Li et al. (2016c) used a similar approach to create Chinese captions for images in the Flickr8K dataset, but they used the translations to train a Chinese image captioning model.

[2] Trained on the ILSVRC 2012 1.2M image 1000 class object classification subset ImageNet (Russakovsky et al., 2015)

In all experiments we inspect the model performance on the validation set at every 500 updates and stop training when no improvement is observed for 10 inspections. The performance metric we use as the stopping criterion is the sum of text-to-image (T→I) and image-to-text (I→T) recall at 1, 5 and 10 scores across all languages in the training data. In all experiments we use a batch-size of 128. The models are trained with the Adam optimizer (Kingma & Ba, 2014) using default parameters and an initial learning rate of 2e-4 without applying any learning-rate decay schedule. We apply gradient norm clipping with a value of 2.0.

We use a pre-trained OpenNMT (Klein et al., 2017) English-German model [3] to create the data for translation approach described in Section 7.2.3.

7.3.2 Datasets

The models are trained and evaluated on the bilingual English-German Multi30K dataset (M30K), and we optionally train on the English COCO dataset (Chen et al., 2015). In *monolingual* experiments, the model is trained on a single language from M30K or COCO.

In the *aligned* bilingual experiments, we use the independently collected English and German captions in M30K: The training set consists of 29K images and 145K captions; the validation and test sets have 1K images and 5K captions.

For the *disjoint* experiments, we use the COCO data set with the (Karpathy & Fei-Fei, 2015) splits. This gives 82,783 training, 5,000 validation, and 5,000 test images; each image is paired with

[3] `https://s3.amazonaws.com/opennmt-models/wmt-ende_l2-h1024-bpe32k_release.tar.gz`

five captions. The data set has an additional split containing the 30,504 images from the original validation set of MS-COCO ("rest-val"), which we add to the training set as in previous work (Karpathy & Fei-Fei, 2015; Vendrov et al., 2016; Faghri et al., 2017).

7.3.3 Evaluation

We report results on Multimodal Translation Shared Task 2016 test split of M30K (Specia et al., 2016). Due to space constraints, we only report recall at 1 (R@1) for Image-to-Text (I→T) and Text-to-Image (T→I) retrieval, and the sum of R@1, R@5, and R@10 recall scores across both tasks and languages (Sum).[4]

7.4 Baseline Results

The experiments presented here set the baseline performance for the visually grounded bilingual models and introduces the data settings that we will use in the later sections.

Aligned In these experiments we only use the *aligned* English-German data from M30K. Tables 7.1 and 7.2 present the result for English and German, respectively. The Sum-of-recall scores for both languages show that the best approach is the bilingual model with the c2c loss (En+De+c2c, and De+En+c2c). These results reproduce the findings of Kádár et al. (2018).

[4] This is the criterion we use for early-stopping.

	English		
	I → T	T → I	Sum
En	40.5	28.8	346.4
+ De	41.4	29.9	352.8
+ c2c	42.8	32.1	361.6
COCO	34.4	24.8	304.0
+ En	**46.2**	**33.4**	**374.4**

Table 7.1: Performance on the English M30K 2016 test set in the *aligned setting* for models trained on M30K English (En), both M30K German and English (+De), and with caption ranking (+c2c).

	German		
	I → T	T → I	Sum
De	34.9	24.6	311.2
+ En	**38.6**	26.0	324.6
+ c2c	38.3	**27.7**	**334.0**
+ COCO	36.4	25.7	319.7

Table 7.2: Performance on the German M30K 2016 test set in the *aligned setting* for models trained on M30K German (De), both M30K German and English (+En), and with caption ranking (+c2c).

	English			German		
	I → T	T → I	Sum	I → T	T → I	Sum
En+De+c2c	42.8	28.6	361.6	38.3	27.7	334.0
+ COCO	**46.5**	**34.8**	**378.9**	**40.6**	**28.8**	**344.6**

Table 7.3: Recall @ 1 and Sum-of-Recall-Scores for Image-to-Text (I → T) and Text-to-Image (T → I) baseline results on the English and German M30K 2016 test in the *aligned plus disjoint* setting

Disjoint We now determine the performance of the model when it is trained on data drawn from different data sets with no overlapping images.

First we train two English *monolingual* models: one on the M30K English dataset and one on the English COCO dataset. Both models are evaluated on image–sentence ranking performance on the M30K English test 2016 set. The results in Table 7.1 show that there is a substantial difference in performance in both text-to-image and image-to-text retrieval, depending on whether the model is trained on the M30K or the COCO dataset. The final row of Table 7.1 shows, however, that jointly training on both data sets improves over only using the M30K English training data.

We also conduct experiments in the *bilingual disjoint* setting, where we study whether it is possible to improve the performance of a German model using the out-of-domain English COCO data. Table 7.2 shows that there is an increase in performance when the model is trained on the disjoint sets, as opposed to only the in-domain M30K German (compare De against De+COCO). This result is not too surprising as we have observed both the advantage of joint training on both languages in the *aligned* setting and the overlap between the different datasets.

Finally, we compare the performance of a German model trained in the *aligned* and *disjoint* settings. We find that a model trained in the *aligned* setting (De+En) is better than a model trained in the *disjoint* setting (De+COCO), as shown in Table 7.2. This finding contradicts the conclusion of Kádár et al. (2018), who claimed that the *aligned* and *disjoint* conditions lead to comparable performance. This is most likely because the disjoint setting in Kádár et al. (2018) is artificial, in the sense that they used different 50% subsets of M30K. In our experiments the disjoint image–caption sets are real, in the sense that we trained the models on the two different datasets.

Aligned plus disjoint Our final baseline experiments explore the combination of *disjoint* and *aligned* data settings. We train an English-German bilingual model with the c2c objective on M30K, and we also train on the English COCO data. Table 7.3 shows that adding the *disjoint* data improves performance for both English and German compared to training solely the *aligned* model.

Summary First we reproduced the findings of Kádár et al. (2018) showing that bilingual joint training improves over monolingual and using c2c loss further improves performance. Furthermore, we have found that adding the COCO as additional training data both when only training on German, and training on both German-English from M30K improves performance even if the model is trained on data drawn from a different dataset.

7.5 Training with Pseudopairs

In this section we turn our attention to creating a synthetic English-German *aligned* data set from the English COCO using the pseudopair method (Section 7.2.1). The synthetic data set is used to train an image-sentence ranking model either from scratch or by fine-tuning the original model; in addition, we also explore the effect of using all of the pseudopairs or by filtering the psuedopairs.

	German		
	I $\rightarrow$ T	T $\rightarrow$ I	Sum
De + COCO	36.4	25.7	319.7
+ pseudo	37.3	25.2	319.9
+ fine-tune	**38.0**	25.6	**322.9**
+ pseudo 25%	37.3	**25.9**	320.9
+ fine-tune	37.2	25.7	320.7
+ pseudo 75%	36.8	25.1	316.3
+ fine-tune	36.5	25.5	317.5

Table 7.4: A *disjoint* model is trained on the De+COCO datasets and used to generate pseudopairs. Then the full pseudopair set (+pseudo) or the filtered versions (+pseudo 25% and +pseudo 75%) are used as an extra data set to either re-train the moddel from scratch or fine-tune the original De+COCO model (+fine-tune).

Disjoint We generate pseudopairs using the *disjoint* bilingual model trained on the German M30K and the English COCO. Table 7.4 reports the results when evaluating on the M30K German data. Line 2 shows that using the full pseudopair set and re-training the model does not lead to noticeable improvements. However, line 3 shows that performance increases when we train with all pseudopairs and fine-tuning the original disjoint bilingual model. Filtering the pseudopairs at either the 25% and 75% percentile is detrimental to the final performance.[5]

Aligned plus disjoint We generate pseudopairs using a model trained on M30K English-German data with the c2c objective and the English COCO data set. The results for both English and German are

[5] We did not find any improvements in the *disjoint* setting when training with pseudopairs and the additional c2c loss.

	English			German			
	I → T	T → I	Sum	I → T	T → I	Sum	Sum(Sum)
En+De+COCO+c2c	46.5	34.8	378.9	40.6	28.8	344.6	723.5
+ pseudo	**48.1**	35.6	**382.3**	**41.8**	29.0	345.6	727.8
+ fine-tune	47.0	**35.7**	381.5	40.9	28.7	346.8	**728.2**
+ pseudo 25%	47.5	34.9	380.2	41.5	28.9	345.5	725.7
+ fine-tune	46.1	35.4	379.7	41.6	**29.1**	**347.8**	727.5
+ pseudo 75%	45.9	34.0	373.6	40.3	27.9	339.1	712.7
+ fine-tune	46.2	35.1	378.6	41.0	**29.1**	345.1	723.6

Table 7.5: We train the *aligned plus disjoint* model with c2c loss and add the full pseudopair set (+pseudo) or the filtered versions (+pseudo 25% and +pseudo 75%) is added as an extra data set. The model is either re-trained from scratch or fine-tuned (+fine-tune).

reported in Table 7.5; note that when we train with the pseudopairs we also train with the c2c loss on both data sets. Overall we find that pseudopairs improve performance, however, we do not achieve the best results for English and German in the same conditions. The best results for German are to filter at 25% percentile and apply fine-tuning, while for English the best results are without filtering or fine-tuning. The best overall model is trained with all the pseudopairs with fine-tuning, according to the Sum of the Sum-of-recall scores across both English and German. The performance across both data sets is increased from 723.5 to 728.2 using the pseudopair method.

Summary In both *aligned* and *disjoint* training with the additional pseudopairs improve performance and in both cases the best performance is achieved when applying the fine-tuning strategy and no filtering of the samples.

	German		
	I → T	T → I	Sum
De + COCO	36.4	25.7	319.7
+ Translation	37.7	26.3	327.2
+ c2c	**39.9**	**26.7**	**335.5**

Table 7.6: Results on the German M30K 2016 test set with the *aligned plus disjoint* (En+De+COCO+c2c) model, the additional automatically translated COCO (+Translation) and with the c2c on the synthetic pairs.

	English			German			
	I → T	T → I	Sum	I → T	T → I	Sum	Sum(Sum)
En+De+COCO+c2c	46.5	34.8	378.87	40.6	28.8	344.59	723.45
+ Translation	**47.5**	**36.2**	**384.50**	**43.5**	**30.5**	**357.90**	**742.40**

Table 7.7: Performance measured on the English and German M30K 2016 test set for the *aligned plus disjoint* (En+De+COCO+c2c) model and with the additional automatically translated (+Translation).

7.6 Training with Translations

We now focus on our second approach to creating an English-German *aligned* dataset using the translation method described in Section 7.2.1.

Disjoint We first report the results of disjoint bilingual model trained on the German M30K, the English COCO data, and the translated German COCO in Table 7.6. The results show that retrieval performance is improved when the model is trained on the translated German COCO data in addition to the English COCO data. We find the best performance when we jointly train on the M30K German, the Translated German COCO and the English COCO with the additional c2c objective over the COCO datasets (Line 4). We note that this setup leads to a better model, as measured by the sum-of-recall-scores, than training on the aligned M30K data (compare De+COCO+Translation+c2c in Table 7.6 to De+En+c2c in Table 7.2).

Aligned plus Disjoint In these experiments, we train models with the *aligned* M30K data, the *disjoint* English COCO data, and the translated German COCO data. Table 7.7 presents the results for the English and German evaluation. We find that training on the German Translated COCO data and using the c2c loss over the COCO data results in improvements for both languages.

Summary In both the *disjoint* and *aligned plus disjoint* settings, we find that training with the translations of COCO improves performance over training with only the English COCO data.

7.7 Discussion

7.7.1 Sentence-similarity quality

The core of the proposed pseudopairing method is based on measuring the similarity between sentences, but how well does our model encode similar sentences? Here we analyse the ability of our models to identify translation equivalent sentences using the English-German translation pairs in the Multi30K test 2016 data. This experiment proceeds as follows: (i) we assume a pre-trained image–sentence ranking model, (ii) we encode the German and English sentences using

	EN → DE	DE → EN
BCN	57.9	57.0
IMG_PIVOT	77.2	76.3
DPCCA	82.6	79.1
En + De	82.7	83.4
En + De + c2c	**90.6**	**91.2**
En + De + COCO	82.5	81.0
En + De + COCO + c2c	90.0	90.1
De + COCO	73.4	70.7

Table 7.8: Translation retrieval results (Recall @ 1) on the M30K 2016 test set compared to the state of the art.

the language encoder of the model, (iii) we calculate the model's performance on the task of ranking the correct translation for English sentences, given the German caption, and vice-versa.

Table 7.8 reports the results of this experiment; we compare against the best approaches as reported by Rotman et al. (2018). Our models consistently improve upon the state-of-the-art. The baseline aligned model trained on the Multi30K data slightly outperforms the DPCCA for EN → DE retrieval, and more substantially outperforms DPCAA for DE → EN. If we train the same model with the additional caption–caption retrieval objective, R@1 improves by 8.0 and 12.1 points, respectively. We find that adding more monolingual English data from the external COCO data set slightly degrades retrieval performance, and that performing sentence retrieval using a model trained on the disjoint M30K German and English COCO data sets result in much lower retrieval performance. We conclude that the model that we used to estimate sentence similarity is the best-performing method known for this task on this data set, but there is room for improvement for models trained on disjoint data sets.

7.7.2 Characteristics of the Pseudopairs

We now investigate the properties of the pseudopairs generated by our method. In particular, we focus on pseudpairs generated by an aligned plus disjoint model (En+De+COCO+c2c) and a disjoint model (De+COCO).

The pseudopairs generated by the *aligned plus disjoint* model cover 40% of the German captions in the M30K data set, and the distribution is not uniform, it is Zipfian (Zipf, 1949). Figure 7.1 shows the distribution of transferred annotations against their rank on a log-log scale: 150 captions are used to annotate 20% of the data. We find a similar pattern for the pseudopairs generated by the *disjoint* model: the pseudopairs cover 37% of the M30K data set, and the top 150 captions cover 23% of the data. This is far from using each caption equally in the pseudopair transfer, and may suggest a hubness problem (Dinu et al., 2014). We assessed the stability of the sets of transferred captions using the Jaccard measure in two cases: (i) different random seeds, and (ii) disjoint or aligned plus disjoint. For the *aligned plus disjoint* model, we observe an overlap of 0.53 between different random seeds compared to 0.51 for the *disjoint* model. The overlap between the two types of models is much lower at 0.41. Finally, we find that when a caption is transferred by both models, the overlap of the caption annotating the same COCO image is 0.33 for the *disjoint* model, and 0.34 for the *aligned plus disjoint* model, and the overlap between the models is 0.16. This shows that the

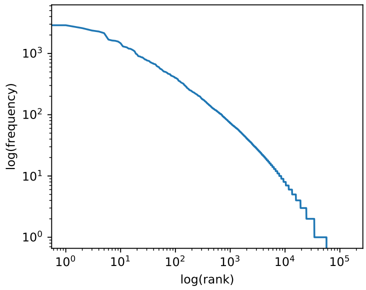


Figure 7.1: Frequency distribution of the German captions transferred by the *disjoint plus aligned* model.

models do not transfer the same captions for the same images.

Figure 7.2 presents examples of the annotations transferred using the pseudopair method. The first example demonstrates the difference between the Multi30K and COCO datasets: there are no giraffes in the former, but there are dogs ("Hund"). In the second example, both captions imply that the man sits *on* the tree not beside it. This shows that even if the datasets are similar, transferring a caption that exactly matches the picture is difficult. The final two examples show semantically accurate and similar sentences are transferred by both models. In the fourth example, both models transfer exactly the same caption.

7.8 Related Work

Image–sentence ranking is the task of retrieving the sentences that best describe an image, and vice-versa (Hodosh et al., 2013). Most recent approaches are based learning to project image representations and sentence representations into a shared space using deep neural networks (Frome et al., 2013; Socher et al., 2014; Vendrov et al., 2016;

1.) Ein hund steht auf einem baumstamm im wald.

2.) Hund im wald.

1.) Mann sitzt im baum.

2.) Der mann der auf einem baum sitzt.

1.) Ein mann sitzt in einem boot auf einem see.

2.) Ein mann sitzt am see auf dem ein boot fährt.

1.) Ein jet jagt steil in die luft, viel rauch kommt aus dem rumpf.

2.) Ein jet jagt steil in die luft, viel rauch kommt aus dem rumpf.

Figure 7.2: Visualisation of the sentences transferred from Multi30K to the COCO data set using the pseudopair method. (1) is transferred from a model trained on De+COCO, whereas (2) is transferred from En+De+COCO.

Faghri et al., 2017, *inter-alia*).

More recently, there has been a focus on solving this task using multilingual data (Gella et al., 2017; Kádár et al., 2018) in the Multi30K dataset (Elliott et al., 2016); an extension of the popular Flickr30K dataset into German, French, and Czech. These works take a multi-view learning perspective in which images and their descriptions in multiple languages are different views of the same concepts. The assumption is that common representations of multiple languages and perceptual stimuli can potentially exploit complementary information between views to learn better representations. For example, (Rotman et al., 2018) improves bilingual sentence representations by incorporating image information as a third view by Deep Partial Canonical Correlation Analysis. More similar to our work (Gella et al., 2017), propose a convolutional-recurrent architecture with both an image–caption and caption–caption loss to learn bilingual visually grounded representations. Their results were improved by the approach presented in Kádár et al. (2018), who also shown the multilingual models outperform bilingual models, and that image–caption retrieval performance in languages with less resources can be improved with data from higher-resource languages. We largely follow Kádár et al. (2018), however, our main interest lies in learning multimodal and bilingual representations in the scenario where the images do not come from the same data set i.e.: the data is presented is two sets of image–caption tuples rather than image–caption–caption triples.

Taking a broader perspective, images have been used as pivots in multilingual multimodal language processing. On the word level this intuition is applied to visually grounded bilingual lexicon induction, which aims to learn cross-lingual word representations without aligned text using images as pivots (Bergsma & Van Durme, 2011; Kiela et al., 2015; Vulić et al., 2016; Hartmann & Søgaard, 2017; Hewitt et al., 2018). Images have been used as pivots to learn translation models only from image–caption data sets, without parallel text (Hitschler et al., 2016; Nakayama & Nishida, 2017; Lee et al., 2017b; Chen et al., 2018).

7.9 Conclusions

Previous work has demonstrated improved image–sentence ranking performance when training models jointly on multiple languages (Gella et al., 2017; Kádár et al., 2018). In this paper, we presented a study where we learn multimodal and multilingual representations in the *disjoint setting*, where images between languages do not overlap. We found that learning visually grounded sentence embeddings in this setting is more challenging then when the images are *aligned* between languages. To close the gap, we developed a *pseudopairing* technique that creates synthetic pairs by annotating the images of one of the data set with the image–descriptions of the other using the sentence similarities of the model trained on both. We showed that training with the pseudopairs improves the performance of the model without the need to augment training from additional data sources or other pipeline components. However, our technique is outperformed by creating synthetic pairs using an off-the-shelf automatic machine translation system. As such our results suggest that it is better to use translation, when a good translation system is available, however, in its absence, pseudopairs offer consistent improvements. The pseudopairing method only transfers annotations from a small number of images; in the future we plan to substitute our naive matching algorithms with approaches developed to mitigate this hubness issue (Radovanović et al., 2010).

8 Discussion and Conclusion

Traditional techniques to learn representations of words and sentences only consider linguistic context as a source of information and focus on the representation of a single language. The aim of this thesis was to make advances towards learning and understanding *visually grounded* and *multilingual* representations. This final chapter is dedicated not only to briefly summarize our main findings, but also to point to some of the limitations of our work and towards future directions.

8.1 Visually grounded word representations

The starting point of the thesis is visual grounding on the word level for a single language. Chapter 3 is dedicated to learning word-representations based on the co-occurrences between words and high-level global image features. We implemented a model taking inspiration from the cross-situational account of word acquisition: a framework of lexical development based on the process of children repeatedly being exposed to pairs of linguistic utterances and perceptual stimuli. Throughout this process they learn a mapping between linguistic units and their referents updating their hypothesis based on a stream of co-occurrence statistics between modalities as evidence.

We departed in two major ways from the canonical computational cross-situational models (Siskind, 1996; Fontanari et al., 2009; Fazly et al., 2010a; Kachergis et al., 2012; Matusevych et al., 2013; Yu & Siskind, 2013): 1.) we use large image–description benchmark datasets developed for machine learning purposes, which naturally present a high level of referential uncertainty and ambiguity and 2.) continuous image representations rather than artificial symbolic scene descriptors.

We adapted the cross-situational word learning model of Fazly et al. (2010a) – an online version of the IBM Model 1 word-translation model (Brown et al., 1993) – to take as input scene representations convolutional neural network activations. Even more unorthodox in the cross-situational literature our model does not learn a mapping between words and objects (Fazly et al., 2010a; Lazaridou et al., 2016), but the strength of the relationship between words and high-level global image feature descriptors. As such our main experimental protocol is related, but departs from the standard word-object retrieval evaluation performed on the CHILDES data set (Goodman et al., 2008; Kievit-Kylar et al., 2013; Lazaridou et al., 2016) and is more similar to performing image–tagging (Weston et al., 2010). Through retrieval experiments we find that our model tags images with relevant concepts and as such learns to ground the meaning of words in visual scenes.

In another experiment we measure the correlation between word similarities under our model and according to human participants and compare the results to a state-of-the-art distributional model trained on the same text corpus. Word-similarity benchmarks were the standard method at the time to compare computational models of word meaning (Faruqui & Dyer, 2014), however, we do note that concerns about the validity and consistency of this methodology has been raised (Faruqui et al., 2016).

Using this protocol allowed us to directly compare a cross-situational and visually grounded computational model of child word learning to the state of the art distributional word2vec approach of the time. The models perform on the same level, however, by creating *abstract* and *concrete* portions of the word similarity benchmarks we find that the distributional word2vec is better on the *abstract*, while the cross-situational model on the *concrete* portion. This result highlights the complementarity between these different sources of learning signals.

Even though our algorithm introduced in Chapter 3 is incremental and is based on a well established model of word-learning (Fazly et al., 2010a) it is not intended as a complete account of child lexical development. Consider the lack of integration of social cues or parent directed attention, both of which have been shown to boost learning in children (Gleitman, 1990; Tomasello & Akhtar, 1995) and in computational cross situational models of word learning (Yu & Ballard, 2007; Lazaridou et al., 2016).

8.2 Visually grounded sentence representations

Moving from representations of words to full sentences in Chapter 4 we continue our comparative study between distributional and visually grounded representations. Concretely, in Chapter 4 we analyze the patterns of hidden activations in our recurrent visually-grounded language learning model IMAGINET (Chrupała et al., 2015).

IMAGINET consists of two separate recurrent networks coupled through a shared word embedding matrix. The VISUAL pathway performs visually grounded learning through an image–sentence ranking objective, while TEXTUAL is trained as a language model to maximize the likelihood of the following word given the preceding context.

Chapter 4 is dedicated to an in-depth investigation of the kind of linguistic structure that is encoded in the hidden activations IMAGINET. Crucially, we perform a controlled comparison between the grounded VISUAL and the distributional TEXTUAL pathways: both models have the same architecture and share word-embeddings. As the linguistic interpretation of representations learned by recurrent networks did not enjoy a large amount of attention at the time we introduced two novel techniques.

Firstly, we introduced the *omission score*, an input-perturbation-based saliency metric assigning a real-valued importance score to each input token signaling how much impact they have on the final sentence representation. Through computing the log ratios of the omission score distribution over part-of-speech tags and dependency labels between VISUAL and TEXTUAL we find that the representations of the former is more impacted by words belonging to categories usually filled with semantic content, while for TEXTUAL the distribution is fairly uniform.

To disentangle the impact of word forms and their functions on the representations we fit ridge regression models to predict omission scores for tokens using the word identity as predictor or adding dependency relation (+DEP), sentence position variables or both and their interactions. We find that the word-identity is much less predictive of omission score for TEXTUAL than it is for VISUAL and that position is most informative for TEXTUAL, while for VISUAL dependency relation and position variables provide similar increases in R^2.

In a more fine-grained analysis we examined the omission distributions computed for the VISUAL pathway of words for which the increase in R^2 from the word-identity regression model to the +DEP model is the highest. We find a general pattern that words generally produce the highest omission scores when they fill the noun subject or root function in a sentence and smallest when they appear as conjuncts.

Lastly we turn to contrasting what the individual dimensions of the two pathways encode. To do so we compute the mutual information between the binned activation values and different type of contexts. For these contexts we consider word and dependency relation n-grams up to order three. This results in distributions of mutual information values between activations and contexts. We take the median of these distributions to compare the types of contexts that are more related to the activation patters in VISUAL than TEXTUAL. We find a significant difference between the pathways: the features encoded by TEXTUAL are more related to dependency labels while that of VISUAL are more related to words.

We can also use these mutual information scores to find for each dimension of the hidden states of the recurrent networks, which contexts they are most related to. Through visual inspection we find these contexts for both VISUAL and TEXTUAL to be a combination of semantic/syntactic constructions. Consider for example two of these top contexts for one of the hidden units of TEXTUAL: *male on a*, *person rides a*. In general for VISUAL we find contexts to be more topically than syntactically related.

Through the development and application of the omission score and mutual information-based interpretation techniques we have found similarities and differences between the sentence representations learned by a image–sentence ranking model and a language model.

Since our article was published there have been a large number of input token saliency measures introduced in the literature as discussed in Section 2.7. Shrikumar et al. (2017) points out that perturbation-based techniques like our omission score tend to struggle due to the limited perturbation window i.e.: the important context might be larger than the scope of the perturbation function. On the explanation evaluation benchmark experiments of Poerner et al. (2018b) perturbation based methods (including ours) lead to inconsistent results and backpropagation style methods (Bach et al., 2015; Shrikumar et al., 2017) have a stronger performance in general.

To find typical inputs to hidden units Poerner et al. (2018a) develop an alternative to our mutual-information-based approach and apply it to our IMAGINET architecture. They perform gradient ascent on maximizing the activation values for particular dimensions.

Even though their technique is quite different, their results reproduce our findings that the hidden units of TEXTUAL tend to be more syntax-aware than that of VISUAL.

Interpretation of deep models of language has become a subfield of NLP in its own right with dedicated venues such as the first BlackboxNLP Workshop at EMNLP 2018 (Alishahi et al., 2019). Future work in the field of linguistic representation learning can benefit from these analysis methods by shedding more light onto the differences between learning from different cues such as the linguistic versus perceptual contexts we considered in this thesis.

8.3 Improving translation with visual grounding

Multi-modal machine translation has been introduced as a shared task in the First Conference of Machine Translation (WMT16) (Specia et al., 2016) initially with German and English data (Elliott et al., 2016) and later with added French and Czech data (Elliott et al., 2017). The task is machine translation with added side information in the form of images i.e. the generation of the target sentence is conditioned both on the source and the corresponding image. Through a review of the state-of-the-art we observe in Chapter 5 that the multi-modal and text-only versions of the top performing systems during the first competition have very similar performances (Specia et al., 2016). Furthermore, when multi-modal systems do perform better it is unclear whether they successfully use images as context to aid translation or they benefit from learning better representations though visual grounding.

In Chapter 5 rather than conditioning translation on the additional image context we introduce an architecture to learn visually grounded representations jointly with translation. More precisely we train a shared encoder which provides input representation to an image–sentence ranking and an attentional translation decoder. The two sub-models are trained jointly with multi-task learning, much like IMAGINET in Chrupała et al. (2015).

We considered two setups: 1.) *aligned* setting where the input is made up of triplets of an English sentence, its German translation and a corresponding image, 2.) *disjoint* where we have a separate English-German parallel corpus and an image–sentence corpus in English. We find that the multi-task model significantly outperforms the text-only baseline with no significant difference in performance between the *aligned* and *disjoint* settings. This result provides evidence that visual grounding can provide a useful inductive bias to improve translation quality. We also performed a follow up experiment where we set a stronger baseline by improving our text-only model performance with training on an additional English-German parallel corpus. We observed that even when extra translation data was available visual grounding still provided improvements in performance. Finally, our full system with the extra parallel corpus and an additional image–sentence data set achieved 2nd place according the human judgments in the WMT Shared Task on Multimodal Translation and Multilingual Image Description (Elliott et al., 2017) on the unconstrained task out of 16 submissions.

On this Shared Task for some systems the multi-modal variant achieved better performance (Caglayan et al., 2017), while the text-only was better for others (Ma et al., 2017). The results of the third shared task are in line with that of the previous years: adding images as extra context to translation systems resulted in marginal differences in translation quality (Barrault et al., 2018).

Elliott (2018) investigates the issue further and introduces a measure of the image awareness of multi-modal translation models and conclude that the current version of the Multi30K data set probably does not contain many training samples where the models need to take the visual modality into account for translation. The role of visual context in translation remains an active area of research at the time of writing this thesis: the "Best Short Paper" title at the North American Chapter of Association for Computational Linguistics 2019 was awarded to Caglayan et al. (2019) who show that multi-modal systems are insensitive to images in general, however, the visual contexts can help models recover from corruptions to the source sentences.

8.4 Multilingual visually grounded sentence representations

We show in Chapter 5 that visually grounded sentence representation learning can provide a useful inductive bias for machine translation. Kiros et al. (2018) also report improved translation performance on the English–German and English–French Multi30K and IWSLT 2014 German-English (Cettolo et al., 2014) translation tasks, when initializing the translation models with their visually grounded Picturebook word-embeddings.

They also show that combining pre-trained GloVe (Pennington et al., 2014) and Picturebook embeddings leads to improvements on natural language inference on the SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017) data sets and on predicting semantic relatedness on the SICK data set (Marelli et al., 2014a). Kiela et al. (2017) through transfer experiments on entailment and sentence classification tasks find that the representations learned by an image–sentence ranking model trained on COCO – similar to the ones trained in Chapters 5, 6, 7 – outperform that of their SkipThought (Kiros et al., 2015) re-implementation. They conclude that visual grounding leads to qualitatively different representations, which can be beneficial to many tasks.

Given the promising results on the benefit of visual grounding on sentence level tasks it seems to be an important avenue to explore techniques to learn better quality and more general multi-modal representations. To this end in Chapters 6 and 7 we turned to learning grounded representations for multipe languages jointly. The benefit of multilingual joint learning has been demonstrated for example in dependency parsing (Ammar et al., 2016), machine translation (Johnson et al., 2016) and grapheme-to-phoneme conversion (Peters et al., 2017). In the visually grounded representation learning Gella et al. (2017) showed improved performance in some cases for image–sentence ranking and semantic textual similarity when training on two languages jointly.

In Chapter 6 we improve and expand the results of Gella et al. (2017). Firstly, contrary to their results we report consistent performance gains in the bilingual setting in our setup: bilingual joint training improves for both English and German image–sentence ranking experiments and the cross–lingual caption–caption ranking objective provides further benefits in both cases. Furthermore, we show that the performance of the bilingual model can be improved by training with two additional languages: French and Czech. We further show that the image–sentence ranking results on the lower resource French and Czech data sets can be improved by jointly training with the larger English and German data sets. Our experiments suggest the clear benefits of multilingual visually grounded representations over their mono- and bilingual counterparts.

These results are all based on a setting where the same images are annotated with captions in different languages. To test whether this *alignment* is crucial, we generate a synthetic *disjoint* set using the data set that we used for the alignment experiments. We find that the models perform equally well in both setups.

Chapter 7 is dedicated to explore the *disjoint* setting under more realistic circumstances, where we have two data sets that were collected separately. Specifically we use the English-German Multi30K and the English COCO for our experiments. Training on COCO resulted in a low test performance on the English Multi30K leading us to conclude that though the two data sets appear fairly similar there is a considerable shift in domains. When training the *disjoint* model on the German Multi30K and English COCO we find that the test performance is lower on the German Multi30K test set compared to training the *aligned model* i.e.: training on the English-German Multi30K. This difference might be partially due to domain-shift. Alternatively, the lack of *alignment* makes it challenging to find correspondences between languages that is crucial to exploit complementary information between them.

The motivation for our *pseudopairs* method in Chapter 7 was to improve the models' capacity to learn from *disjoint* data sets by creating a *synthetic alignment*. In the fully *disjoint* experiment our method starts by training a bilingual model on the German Multi30K and the English COCO. Then we use the sentence similarities under this *parent model* to generate a synthetic English-German COCO: For each English sentence in COCO we find the most similar sentence in Multi30k and use it as the German description for that image. We add an extra experiment we call *aligned plus disjoint* setting where both the bilingual Multi30k and the additional larger English COCO is available. In both cases we find improvements when we use the pseudopair method.

We demonstrated in Chapter 6 that using our setup we can learn better visually grounded representations when training on multiple languages. Our results suggest that the more languages are used for training the larger the gain, however, we only considered the same images annotated with four languages. Future work can shed more light on under what circumstances this observation holds by running experiments with different resources such as the Chinese Flickr8k (Li et al., 2016c) or the Japanese COCO (Miyazaki & Shimizu, 2016).

Even though the pseudopair method presented in Chapter 7 provides consistent improvements we find that using an off-the-shelf model to translate the captions to the desired language to create synthetic pairs leads to larger improvements. In future work we will focus on narrowing this gap by applying better matching algorithms to generate pseudopairs that mitigate the hubness problem (Radovanović et al., 2010; Tomašev et al., 2011a,b; Dinu et al., 2014).

8.5 Future directions and limitations

8.5.1 Embedding geometry

The visually grounded sentence representation learning approaches we presented in this thesis learn to represent sentences as points in an embedding space through minimizing symmetric distance functions. An alternative we have not explored is the asymmetric order-embeddings (Vendrov et al., 2016) which was shown to outperform cosine-similarity-based image–sentence ranking models (Faghri et al., 2017). Richer semantic representations could be also explored in the future such as Gaussian distributions over a latent embedding space (Vilnis & McCallum, 2014), which allow for the modeling of asymmetric relationship between sentences such as entailment, specificity/inclusion and uncertainty. Another lesser known approach in the literature is to embed objects in a hyperbolic space (Nickel & Kiela, 2017), which naturally represents hierarchical relationships and can be thought of as a continuous version of trees (Krioukov et al., 2010). In a similar setup to ours Tay et al. (2018) trains a simple hyperbolic sentence embedding architecture for question-answer ranking using a pairwise ranking loss without any attention/interaction layer, which performs at state-of-the-art level. Future work in visually grounded and multilingual representation learning can benefit from exploring spaces of various topology for the embedding of sentences.

8.5.2 Multi-task learning

Progress towards learning better visually grounded and multilingual representations can be accelerated by exploring different multi-task learning strategies. In chapters 4, 5, 6 and 7 we used hard parameter sharing (Caruana, 1997; Collobert et al., 2011) and share the whole encoder across all languages. This strategy is expected to work when there is a close relationship between tasks (Baxter, 2000) – or languages in our case – and worse results are expected as the distance between tasks grow (Maurer, 2006). The other most common approach to parameter sharing in neural networks is soft parameter sharing: each task has its own set of parameters, but a regularization penalty is added forcing these parameter sets to be similar (Duong et al., 2015; Yang & Hospedales, 2016).

Furthermore, in chapters 6 and 7 we follow the common practice of pre-defining uniform task sampling probabilities before training time and keep them fixed during training (Alonso & Plank, 2016; Bingel & Søgaard, 2017). However, Kiperwasser & Ballesteros (2018) observe benefits from applying fixed schedules to the task sampling probabilities, while Sanh et al. (2018) show the benefits of adjusting the sampling probabilities to the size of the training sets of different tasks. Recently Ruder et al. (2017) presented a method to jointly learn the task-weighting or sampling probabilities and which parameters to share.

8.5.3 Local image descriptors

The approaches presented in the thesis extracted global image and sentence features through separate encoders and learned to associate them. Another line of work focuses on learning latent alignments between sentence fragments – usually words – and image regions. The architecture presented in (Karpathy & Fei-Fei, 2015) first encodes the sentence with a bidirectional recurrent network and extracts image region features from a pre-trained convolutional neural network. The dot product between the hidden states of the recurrent network and local region features are interpreted as the similarity between regions and words. This formulation offered a blue-print for later models that define various more sophisticated attention mechanisms to compute region–word interactions in a multi-step fashion (Nam et al., 2017; Huang et al., 2017).

8.6 Conclusion

Chapters 3 and 4 focused on models that learn word and sentence embeddings, using the visual world as a naturally occuring meaning representation for linguistic utterrances. In these chapters we developed and contrasted such visually grounded linguistic representation learning models with traditional distributional approaches.

In Chapter 3 we developed a novel model of cross-situational word learning, which learns to represent each word as a vector of probabilities indicating the strength of association between the word and high-level convolutional image features descriptors. We demonstrated that, while on human similarity benchmarks our model and a state-of-the-art distributional model perform on par, the distributional model captures the semantics of *abstract* and our visually grounded model captures that of the *concrete* words more accurately. We also showed that our model tags images with relevant concepts.

Chapter 4 focused on visually grounded representations on the sentence level. We developed novel techniques to analyze and understand the representations learned by recurrent neural network models of language and used these methods to compare a language model and an image–sentence ranking model i.e. distributional versus visually grounded models of sentence representation learning. We have found that the representations learned by the visually grounded model are mostly impacted by semantically contentful words, while for the language model this saliency score over word categories is more uniform. Furthermore, we have both quantitatively and qualitatively identified differences between the features encoded by the hidden dimensions of grounded and distributional models. We presented evidence that the dimensions of the latter are more associated with syntactic patterns, while that of the former seems to be more lexicalized and focused on topically related word sequences.

Chapters 5, 6 and 7 took the concept of visual grounding a step further and utilized images to bridge between multiple languages. We applied visual grounding to improve machine translation and studied multilingual visually grounded sentence representations.

Chapter 5 introduced our first multilingual experiment and showed that visually grounded sentence representation learning can improve machine translation. We applied multi-task learning to train an English sentence encoder to provide useful representations for both image–sentence matching and translation. Compared to the text-only baseline we observed performance gains both when training on English-German-Image translation pairs and corresponding scene triplets and when training on two separate sets: an English-German parallel corpus and an English image–description data set. Furthermore, we also showed that visually grounded representation learning results in orthogonal improvements to having access to additional parallel corpora.

Finally, Chapters 6 and 7 focused on learning better quality and more general visually grounded sentence representations by learning to represent multiple languages jointly.

Chapter 6 presents several results on the benefits of multilingual joint learning on image–sentence ranking. Building on the experiments of Gella et al. (2017) we presented a multi-task learning setup combining an image–caption and a cross-lingual caption–caption loss functions. Using this setup we showed that multilingual representations consistently outperform bilingual and monolingual grounded sentence representations in image–sentence ranking experiments. Furthermore, we provided evidence that the performance on lower-resource languages can be improved by training jointly with higher-resource languages. In all these cases the caption–caption objective provided consistent improvements. A limitation, however, of most of oure experiments in Chapter 6 is that they are based on a specific data configuration, where the same images are annotated with multiple languages. To assess the impact of the existence of this *alignment* on the performance we generate synthetically a *disjoint* data set from the *aligned* corpus. In Chapter 6 we find that the *disjoint* and *aligned* settings results in comparable performance improvements compared to monolingual training.

Chapter 7 explores the issue of *alignment* in a more realistic scenario, where seperate image–sentence corpora are available for different languages. We do find results contradictory to that of in Chapter 6, namely that the *aligned* setting offers more performance benefit compared to the *disjoint* setting. To help close the gap between the aligned and disjoint conditions we develop a technique we call *pseudopairs* to generate a synthetic *aligned* data set given two *disjoint* sets, without requiring any extra data or pipeline elements. We find that our technique improves performance, however, lags behind using an automatic machine translation system to create a synthetic aligned set in the desired language.

To improve performance and understanding in visually grounded and multilingual representation learning we see several avenues for future research. One of our main interest for future work is to pinpoint some of the mechanisms that lead to improvements when learning multilingual as opposed to monolingual grounde representations. For better results we consider replacing our naive multi-task learning and parameter sharing algorithms with more principled approaches. Similarly, in future work we will focus on improving our pseudopair generation algorithm, replacing our initial implementation with more sophisticated matching algorithms optimizing not only for the quality, but also diversity of the resulting synthetic sets. A further natural extension to the work presented here is taking advantage of image–sentence data sets in all available languages, rather than just focusing on the ones we considered in the thesis. Our result in Chapter 6 suggest that training on more language lead to improved results, however, it is unclear how general our findings are due to the limited number of languages we considered. Finally, exploring embedding spaces with different topologies – such as hyperpolic geometry – we think is an exciting future avenue for visually grounded and multilingual representations learning.

References

Adi, Y., Kermany, E., Belinkov, Y., Lavi, O., & Goldberg, Y. (2017). Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations (ICLR)*.

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., et al. (2015). Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)* (pp. 252–263).

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., & Wiebe, J. (2014). Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 81–91).

Akaho, S. (2006). A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*.

Alishahi, A. & Chrupała, G. (2012). Concurrent acquisition of word meaning and lexical categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 643–654).: Association for Computational Linguistics.

Alishahi, A., Chrupała, G., & Linzen, T. (2019). Analyzing and interpreting neural networks for nlp: A report on the first blackboxnlp workshop. *arXiv preprint arXiv:1904.04063*.

Alonso, H. M. & Plank, B. (2016). When is multitask learning effective? semantic sequence prediction under varying data conditions. *arXiv preprint arXiv:1612.02251*.

Ammar, W., Mulcaire, G., Ballesteros, M., Dyer, C., & Smith, N. A. (2016). Many languages, one parser. *arXiv preprint arXiv:1602.01595*.

Andrew, G., Arora, R., Bilmes, J., & Livescu, K. (2013). Deep canonical correlation analysis. In *International conference on machine learning* (pp. 1247–1255).

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review*, 116(3), 463.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., & Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2425–2433).

AP, S. C., Lauly, S., Larochelle, H., Khapra, M., Ravindran, B., Raykar, V. C., & Saha, A. (2014). An autoencoder approach to learning bilingual word representations. In *Advances in Neural Information Processing Systems* (pp. 1853–1861).

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7), e0130140.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bansal, T., Belanger, D., & McCallum, A. (2016). Ask the gru: Multi-task learning for deep text recommendations. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 107–114).: ACM.

Barman, U., Das, A., Wagner, J., & Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the first workshop on computational approaches to code switching* (pp. 13–23).

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1), 3–13.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014a). Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, 9.

Baroni, M., Dinu, G., & Kruszewski, G. (2014b). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1 (pp. 238–247).

Baroni, M. & Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, 20(1), 55–88.

Barrault, L., Frank, S., Elliott, D., & Specia, L. (2018). Findings of the third shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Third Conference on Machine Translation, Volume 2: Shared Task Papers* Brussels, Belgium.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2), 84–91.

Baxter, J. (2000). A model of inductive bias learning. *Journal of artificial intelligence research*, 12, 149–198.

Beekhuizen, B., Fazly, A., Nematzadeh, A., & Stevenson, S. (2013). Word learning in the wild: What natural data can tell us. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society.*

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

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The Journal of Machine Learning Research*, 3, 1137–1155.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2), 157–166.

Bergsma, S. & Van Durme, B. (2011). Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22 (pp. 1764).

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55, 409–442.

Bernardy, J.-P. & Lappin, S. (2017). Using deep neural networks to learn syntactic agreement. *LiLT (Linguistic Issues in Language Technology)*, 15.

Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., & Samek, W. (2016). Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks* (pp. 63–71).: Springer.

Bingel, J. & Søgaard, A. (2017). Identifying beneficial task relations for multi-task learning in deep neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 164–169).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 993–1022.

Bock, K. & Miller, C. A. (1991). Broken agreement. *Cognitive psychology*, 23(1), 45–93.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146.

Bornstein, M. H., Cote, L. R., Maital, S., Painter, K., Park, S.-Y., Pascual, L., Pêcheux, M.-G., Ruel, J., Venuti, P., & Vyt, A. (2004). Cross-linguistic analysis of vocabulary in young children: Spanish, dutch, french, hebrew, italian, korean, and american english. *Child Development*, 75(4), 1115–1139.

Bowman, S. R. (2016). *Modeling natural language semantics in learned representations.* PhD thesis, Ph. D. thesis, Stanford University.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., & Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2), 263–311.

Bruni, E., Boleda, G., Baroni, M., & Tran, N.-K. (2012). Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1* (pp. 136–145).: Association for Computational Linguistics.

Bruni, E., Tran, G. B., & Baroni, M. (2011). Distributional semantics from text and images. In *Proceedings of the GEMS 2011 workshop on geometrical models of natural language semantics* (pp. 22–32).: Association for Computational Linguistics.

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res(JAIR)*, 49, 1–47.

Bullinaria, J. A. & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior research methods*, 39(3), 510–526.

Bullinaria, J. A. & Levy, J. P. (2012). Extracting semantic representations from word co-occurrence statistics: stop-lists, stemming, and svd. *Behavior research methods*, 44(3), 890–907.

Caglayan, O., Aransa, W., Bardet, A., García-Martínez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., & Van de Weijer, J. (2017). Lium-cvc submissions for wmt17 multimodal translation task. *arXiv preprint arXiv:1707.04481*.

Caglayan, O., Barrault, L., & Bougares, F. (2016). Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Caglayan, O., Madhyastha, P., Specia, L., & Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.

Calixto, I., Elliott, D., & Frank, S. (2016). DCU-UvA Multimodal MT System Report. In *Proceedings of the First Conference on Machine Translation* (pp. 634–638).

Calixto, I. & Liu, Q. (2017a). Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1003–1014).

Calixto, I. & Liu, Q. (2017b). Sentence-level multilingual multi-modal embedding for natural language processing. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017* (pp. 139–148).: INCOMA Ltd.

Calixto, I., Liu, Q., & Campbell, N. (2017a). Doubly-Attentive Decoder for Multi-modal Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1913–1924).

Calixto, I., Liu, Q., & Campbell, N. (2017b). Multilingual multi-modal embeddings for natural language processing. *arXiv preprint arXiv:1702.01101*.

Carey, S. (1978). The child as word learner. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic Theory and Psychological Reality*. The MIT Press.

Caruana, R. (1997). Multitask learning. *Machine learning*, 28(1), 41–75.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730).: ACM.

Cauchy, A. (1847). Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847), 536–538.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., & Federico, M. (2014). Report on the 11th iwslt evaluation campaign, iwslt 2014. In *Proceedings of the International Workshop on Spoken Language Translation, Hanoi, Vietnam* (pp.57).

Chandar, S., Khapra, M. M., Larochelle, H., & Ravindran, B. (2016). Correlational neural networks. *Neural computation*, 28(2), 257–285.

Chen, D., Socher, R., Manning, C. D., & Ng, A. Y. (2013). Learning new facts from knowledge bases with neural tensor networks and semantic word vectors. *arXiv preprint arXiv:1301.3618*.

Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325.

Chen, X. & Zitnick, C. L. (2014). Learning a recurrent visual representation for image caption generation. *arXiv preprint arXiv:1411.5654*.

Chen, Y., Liu, Y., & Li, V. O. (2018). Zero-resource neural machine translation with multi-agent communication game. *arXiv preprint arXiv:1802.03116*.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014a). On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation* (pp. 103–111).

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014b). Learning phrase representations using RNN encoder-decoder for statistical machine translation. (pp. 1724–1734).

Chomsky, N. (1957). *Syntactic structures.*

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1 (pp. 613–622).

Chrupała, G., Gelderloos, L., Kádár, Á., & Alishahi, A. (2019). On the difficulty of a distributional semantics of spoken language. *Proceedings of the Society for Computation in Linguistics*, 2(1), 167–173.

Chrupała, G., Kádár, À., & Alishahi, A. (2015). Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2 (pp. 112–118).

Chung, J., Cho, K., & Bengio, Y. (2016). A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Deep Learning and Representation Learning Workshop*.

Clark, S. & Pulman, S. (2007). Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction* (pp. 52–55).

Clarke, D. (2012). A context-theoretic framework for compositionality in distributional semantics. *Computational Linguistics*, 38(1), 41–71.

Coecke, B., Sadrzadeh, M., & Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.

Collell, G., Zhang, T., & Moens, M.-F. (2017). Imagined visual representations as multimodal embeddings. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)* (pp. 4378–4384).

Collobert, R. & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning* (pp. 160–167).: ACM.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug), 2493–2537.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 670–680). Copenhagen, Denmark: Association for Computational Linguistics.

Conneau, A., Schwenk, H., Barrault, L., & Lecun, Y. (2016). Very deep convolutional networks for natural language processing. *arXiv preprint*.

Côté, M.-A., Kádár, Á., Yuan, X., Kybartas, B., Barnes, T., Fine, E., Moore, J., Hausknecht, M., El Asri, L., Adada, M., et al. (2018). Textworld: A learning environment for text-based games. In *Proceedings of the Computer Games Workshop at ICML/IJCAI 2018* (pp. 1–29).

Cruse, D. A. & Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1 (pp. 1–2).: Prague.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, (pp. 1–38).

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (pp. 248–255).: IEEE.

Denkowski, M. & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dinu, G., Lazaridou, A., & Baroni, M. (2014). Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

Dolan, B., Quirk, C., & Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on Computational Linguistics* (pp. 350).: Association for Computational Linguistics.

Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2014a). Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014b). Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning* (pp. 647–655).

Dong, D., Wu, H., He, W., Yu, D., & Wang, H. (2015). Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (pp. 1723–1732).

Dosovitskiy, A. & Brox, T. (2015). Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dumais, S. T. (2004). Latent semantic analysis. *Annual review of information science and technology*, 38(1), 188–230.

Duong, L., Cohn, T., Bird, S., & Cook, P. (2015). Low resource dependency parsing: Cross-lingual parameter sharing in a neural network parser. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2 (pp. 845–850).

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, volume 38. SIAM.

Eigen, D., Rolfe, J., Fergus, R., & LeCun, Y. (2014). Understanding deep architectures using a recursive convolutional network. In *International Conference on Learning Representations (ICLR)*.

Eisenschtat, A. & Wolf, L. (2017). Linking image and text with 2-way nets. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4601–4611).

Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2974–2978).

Elliott, D., Frank, S., Barrault, L., Bougares, F., & Specia, L. (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*.

Elliott, D., Frank, S., & Hasler, E. (2015). Multi-language image description with neural sequence models. *CoRR*, abs/1510.04709.

Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language* (pp. 70–74).: Association for Computational Linguistics.

Elliott, D. & Kádár, À. (2017). Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1 (pp. 130–141).

Elliott, D. & Keller, F. (2014). Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2 (pp. 452–457).

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, 14(2), 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine learning*, 7(2-3), 195–225.

Enguehard, E., Goldberg, Y., & Linzen, T. (2017). Exploring the syntactic abilities of rnns with multi-task learning. *arXiv preprint arXiv:1706.03542*.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. In *International Conference on Machine Learning (ICML) Workshop on Learning Feature Hierarchies*, volume 1341.

Evert, S. (2005). The statistics of word cooccurrences: word pairs and collocations.

Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2017). Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612*, 2(7), 8.

Fang, H., Gupta, S., Iandola, F., Srivastava, R., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Platt, J., et al. (2014). From captions to visual concepts and back. *arXiv preprint arXiv:1411.4952*.

Faruqui, M. & Dyer, C. (2014). Community evaluation and exchange of word vectors at wordvectors. org. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 19–24).

Faruqui, M., Tsvetkov, Y., Rastogi, P., & Dyer, C. (2016). Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.

Fazly, A., Ahmadi-fakhr, F., Alishahi, A., & Stevenson, S. (2010a). Cross-situational learning of low frequency words: The role of context familiarity and age of exposure. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2615–20).

Fazly, A., Alishahi, A., & Stevenson, S. (2010b). A probabilistic computational model of cross-situational word learning. *Cognitive Science: A Multidisciplinary Journal*, 34(6), 1017–1063.

Fellbaum, C. (1998). *WordNet*. Wiley Online Library.

Feng, Y. & Lapata, M. (2010). Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 91–99).: Association for Computational Linguistics.

Ferreira, T. C., Moussallem, D., Kádár, Á., Wubben, S., & Krahmer, E. (2018). Neuralreg: An end-to-end approach to referring expression generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1959–1969).

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001). Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web* (pp. 406–414).: ACM.

Firat, O., Cho, K., & Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 866–875).

Fleischman, M. & Roy, D. (2005). Intentional context in situated language learning. In *9th Conference on Computational Natural Language Learning*.

Fontanari, J. F., Tikhanoff, V., Cangelosi, A., Ilin, R., & Perlovsky, L. I. (2009). Cross-situational learning of object–word mapping using neural modeling fields. *Neural Networks*, 22(5-6), 579–585.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In *Advances in Neural Information Processing Systems*, volume 20.

Frank, S., Elliott, D., & Specia, L. (2018). Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3), 393–413.

Freitag, M. & Al-Onaizan, Y. (2016). Fast domain adaptation for neural machine translation. *CoRR*, abs/1612.06897.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al. (2013). Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems* (pp. 2121–2129).

Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., & Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Funaki, R. & Nakayama, H. (2015). Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 585–590).

Gal, Y. & Ghahramani, Z. (2016). A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems 29* (pp. 1019–1027).

Gan, Z., Pu, Y., Henao, R., Li, C., He, X., & Carin, L. (2016). Unsupervised learning of sentence representations using convolutional neural networks. *arXiv preprint arXiv:1611.07897*.

Gelderloos, L. & Chrupała, G. (2016). From phonemes to images: levels of representation in a recurrent neural model of visually-grounded language learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics* (pp. 1309–1319).

Gella, S., Sennrich, R., Keller, F., & Lapata, M. (2017). Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2839–2845).: Association for Computational Linguistics.

Gers, F. A., Schmidhuber, J., & Cummins, F. (1999). Learning to forget: Continual prediction with lstm.

Giles, C. L., Miller, C. B., Chen, D., Sun, G., Chen, H., & Lee, Y. (1991). Extracting and learning an unknown grammar with recurrent neural networks. In *Advances in Neural Information Processing Systems* (pp. 317–324).

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580–587).

Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, 1(1), 3–55.

Goller, C. & Kuchler, A. (1996). Learning task-dependent distributed representations by backpropagation through structure. *Neural Networks*, 1, 347–352.

Gong, Y., Wang, L., Hodosh, M., Hockenmaier, J., & Lazebnik, S. (2014). Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision* (pp. 529–545).: Springer.

Goodman, N., Tenenbaum, J. B., & Black, M. J. (2008). A bayesian framework for cross-situational word-learning. In *Advances in neural information processing systems* (pp. 457–464).

Graham, Y., Baldwin, T., Moffat, A., & Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1), 3–30.

Gregor, K., Danihelka, I., Graves, A., & Wierstra, D. (2015). Draw: A recurrent neural network for image generation. In *International Conference on Machine Learning (ICML)*.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Haghighi, A., Liang, P., Berg-Kirkpatrick, T., & Klein, D. (2008). Learning bilingual lexicons from monolingual corpora. *Proceedings of ACL-08: Hlt*, (pp. 771–779).

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

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), 146–162.

Hartmann, M. & Søgaard, A. (2017). Limitations of cross-lingual learning from image search. *arXiv preprint arXiv:1709.05914*.

Harwath, D., Chuang, G., & Glass, J. (2018). Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. *arXiv preprint arXiv:1804.03052*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hewitt, J., Ippolito, D., Callahan, B., Kriz, R., Wijaya, D. T., & Callison-Burch, C. (2018). Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1 (pp. 2566–2576).

Hill, F., Cho, K., & Korhonen, A. (2016). Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4), 665–695.

Hitschler, J., Schamoni, S., & Riezler, S. (2016). Multimodal Pivots for Image Caption Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 2399–2409).

Hochreiter, S. & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

Hodosh, M., Young, P., & Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47, 853–899.

Horst, P. (1961). *Generalized canonical correlations and their applications to experimental data*. Technical report, WASHINGTON UNIV SEATTLE.

Howard, J. & Ruder, S. (2018). Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1 (pp. 328–339).

Huang, P.-Y., Liu, F., Shiang, S.-R., Oh, J., & Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation* (pp. 639–645).

Huang, Y., Wang, W., & Wang, L. (2017). Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2310–2318).

Jabri, A., Joulin, A., & van der Maaten, L. (2016). Revisiting visual question answering baselines. In *European conference on computer vision* (pp. 727–739).

Jernite, Y., Bowman, S. R., & Sontag, D. (2017). Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., et al. (2016). Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential network. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*.

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Kachergis, G., Yu, C., & Shiffrin, R. M. (2012). An associative model of adaptive inference for learning word–referent mappings. *Psychonomic bulletin & review*, 19(2), 317–324.

Kádár, Á., Alishahi, A., & Chrupała, G. (2015a). Learning word meanings from images of natural scenes. *Traitement Automatique des Langues*, 55(3).

Kádár, A., Chrupała, G., & Alishahi, A. (2015b). Lingusitic analysis of multimodal recurrent neural networks. In *Proceedings of the Fourth Workshop on Vision and Language* (pp. 8–9).

Kádár, A., Chrupała, G., & Alishahi, A. (2017). Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4), 761–780.

Kádár, Á., Côté, M.-A., Chrupała, G., & Alishahi, A. (2018a). Revisiting the hierarchical multiscale lstm. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3215–3227).

Kádár, Á., Elliott, D., Côtė, M.-A., Chrupała, G., & Alishahi, A. (2018b). Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning* (pp. 402–412).

Kahou, S. E., Atkinson, A., Michalski, V., Kádár, Á., Trischler, A., & Bengio, Y. (2018). Figureqa: An annotated figure dataset for visual reasoning.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188*.

Karpathy, A. & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3128–3137).

Karpathy, A., Johnson, J., & Li, F.-F. (2016). Visualizing and understanding recurrent networks. In *International Conference on Learning Representations (ICLR) Workshop*.

Karpathy, A., Joulin, A., & Fei-Fei, L. F. (2014). Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems* (pp. 1889–1897).

Kenter, T., Borisov, A., & de Rijke, M. (2016). Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*.

Kiela, D. & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014*

Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 36–45).

Kiela, D., Conneau, A., Jabri, A., & Nickel, M. (2017). Learning visually grounded sentence representations. *arXiv preprint arXiv:1707.06320*.

Kiela, D., Hill, F., Korhonen, A., & Clark, S. (2014). Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2 (pp. 835–841).

Kiela, D., Verő, A. L., & Clark, S. (2016). Comparing Data Sources and Architectures for Deep Visual Representation Learning in Semantics. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-16)* (pp. 447–456).

Kiela, D., Vulic, I., & Clark, S. (2015). Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*: ACL.

Kievit-Kylar, B., Kachergis, G., & Jones, M. (2013). Naturalistic word-concept pair learning with semantic spaces. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 35.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Kingma, D. P. & Ba, J. (2014). Adam: A method for stochastic optimization. In *International Conference for Learning Representations*.

Kiperwasser, E. & Ballesteros, M. (2018). Scheduled multi-task learning: From syntax to translation. *Transactions of the Association of Computational Linguistics*, 6, 225–240.

Kiros, J. & Chan, W. (2018). Inferlite: Simple universal sentence representations from natural language inference data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4868–4874).

Kiros, J., Chan, W., & Hinton, G. (2018). Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 922–933).

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Skip-thought vectors. In *Advances in Neural Information Processing Systems* (pp. 3276–3284).

Klein, B. E., Lev, G., Sadeh, G., & Wolf, L. (2015). Associating neural word embeddings with deep image representations using fisher vectors. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 4437–4446).

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A. M. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*.

Klejch, O., Avramidis, E., Burchardt, A., & Popel, M. (2015). Mt-compareval: Graphical evaluation interface for machine translation development. *The Prague Bulletin of Mathematical Linguistics*, 104(1), 63–74.

Klerke, S., Goldberg, Y., & Søgaard, A. (2016). Improving sentence compression by learning to predict gaze. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1528–1533).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual meeting of Association for Computational Linguistics* (pp. 177–180).

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., & Boguná, M. (2010). Hyperbolic geometry of complex networks. *Physical Review E*, 82(3), 036106.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kádár, ., Elliott, D., Côté, M.-A., Chrupała, G., & Alishahi, A. (2018). Lessons learned in multilingual grounded language learning. In *Proceedings of the 21st Conference on Computational Natural Language*.

Landau, B., Smith, L., & Jones, S. (1998). Object perception and object naming in early development. *Trends in cognitive sciences*, 2(1), 19–24.

Landauer, T. K. & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.

Lazaridou, A., Chrupała, G., Fernández, R., & Baroni, M. (2016). Multimodal semantic learning from child-directed input. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 387–392).

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive science*, 41, 677–705.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Le, P. & Zuidema, W. (2015). Compositional distributional semantics with long short term memory. *arXiv preprint arXiv:1503.02510*.

Lee, J., Cho, K., & Hofmann, T. (2017a). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5, 365–378.

Lee, J., Cho, K., Weston, J., & Kiela, D. (2017b). Emergent translation in multi-agent communication. *arXiv preprint arXiv:1710.06922*.

Leong, C. W. & Mihalcea, R. (2011a). Going beyond text: A hybrid image-text approach for measuring word relatedness. In *Proceedings of 5th International Joint Conference on Natural Language Processing* (pp. 1403–1407).

Leong, C. W. & Mihalcea, R. (2011b). Measuring the semantic relatedness between words and images. In *Proceedings of the Ninth International Conference on Computational Semantics* (pp. 185–194).: Association for Computational Linguistics.

Levy, O. & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185).

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3, 211–225.

Li, H. (2011). Learning to rank for information retrieval and natural language processing. *Synthesis Lectures on Human Language Technologies*, 4(1), 1–113.

Li, J., Chen, X., Hovy, E., & Jurafsky, D. (2016a). Visualizing and understanding neural models in NLP. In *North American Chapter of the Association for Computational Linguistic (NAACL)*.

Li, J., Luong, M.-T., Jurafsky, D., & Hovy, E. (2015). When are tree structures necessary for deep learning of representations? *arXiv preprint arXiv:1503.00185*.

Li, J., Monroe, W., & Jurafsky, D. (2016b). Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical development in a self-organizing neural network. *Neural Networks*, 17, 1345–1362.

Li, X., Lan, W., Dong, J., & Liu, H. (2016c). Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval* (pp. 271–275).: ACM.

Li, Y., Yosinski, J., Clune, J., Lipson, H., & Hopcroft, J. (2016d). Convergent learning: Do different neural networks learn the same representations? In *International Conference on Learning Representation (ICLR)*.

Libovicky, J. & Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 196–202). Stroudsburg, PA, USA: Association for Computational Linguistics.

Libovický, J., Helcl, J., Tlustý, M., Bojar, O., & Pecina, P. (2016). Cuni system for wmt16 automatic post-editing and multimodal translation tasks. In *Proceedings of the First Conference on Machine Translation* (pp. 646–654).

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European conference on computer vision* (pp. 740–755).: Springer.

Lin, X. & Parikh, D. (2015). Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2984–2993).

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4, 521–535.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2), 273–302.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2 (pp. 1150–1157).: Ieee.

Lund, K. & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior research methods, instruments, & computers*, 28(2), 203–208.

Luong, M., Le, Q. V., Sutskever, I., Vinyals, O., & Kaiser, L. (2016). Multi-task sequence to sequence learning. In *ICLR*.

Ma, M., Li, D., Zhao, K., & Huang, L. (2017). Osu multimodal machine translation system report. *arXiv preprint arXiv:1710.02718*.

MacWhinney, B. (2014). *The CHILDES Project: Tools for Analyzing Talk, Volume I: Transcription Format and Programs*. Psychology Press.

Mahendran, A. & Vedaldi, A. (2015). Understanding deep image representations by inverting them. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 5188–5196).

Mahendran, A. & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, 120(3), 233–255.

Malaviya, C., Neubig, G., & Littell, P. (2017). Learning language representations for typology prediction. *arXiv preprint arXiv:1707.09569*.

Manjavacas, E., Kádár, Á., & Kestemont, M. (2019). Improving lemmatization of non-standard languages with joint learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2014a). Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*.

Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L. (2014b). Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*.

Marcheggiani, D. & Titov, I. (2017). Encoding sentences with graph convolutional networks for semantic role labeling. *arXiv preprint arXiv:1703.04826*.

Marelli, M., Bentivogli, L., Baroni, M., Bernardi, R., Menini, S., & Zamparelli, R. (2014a). Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)* (pp. 1–8).

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014b). A sick cure for the evaluation of compositional distributional semantic models. In *LREC* (pp. 216–223).

Martens, J. & Sutskever, I. (2011). Learning recurrent neural networks with hessian-free optimization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 1033–1040).: Citeseer.

Matusevych, Y., Alishahi, A., & Vogt, P. (2013). Automatic generation of naturalistic child–adult interaction data. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society. Austin, TX: Cognitive Science Society* (pp. 2996–3001).

Maurer, A. (2006). Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7(Jan), 117–139.

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex*, 48(7), 788–804.

Meteyard, L. & Vigliocco, G. (2008). The role of sensory and motor information in semantic representation: A review. *Handbook of cognitive science: An embodied approach*, (pp. 293–312).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Karafiát, M., Burget, L., Černockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Miller, G. A. & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1), 1–28.

Mitchell, J. & Lapata, M. (2008). Vector-based models of semantic composition. *proceedings of ACL-08: HLT*, (pp. 236–244).

Miyazaki, T. & Shimizu, N. (2016). Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1780–1790).: Association for Computational Linguistics.

Montague, R. (1970). English as a formal language. *Linguaggi nella Societa e nella Tecnica*.

Nakayama, H. & Nishida, N. (2017). Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2), 49–64.

Nam, H., Ha, J.-W., & Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 299–307).

Nelson, D., McEvoy, C., & Schreiber, T. (1998). The university of south florida word association, rhyme, and word fragment norms. 1998 http://www. usf. edu. *FreeAssociation.[PubMed]*.

Nevin, B. E. & Johnson, S. M. (2002). *The legacy of Zellig Harris: language and information into the 21st century*, volume 1. John Benjamins Publishing.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 689–696).

Nguyen, A., Yosinski, J., & Clune, J. (2016). Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. In *Visualization for Deep Learning workshop at International Conference on Machine Learning (ICML)*.

Nickel, M. & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems* (pp. 6338–6347).

Nivre, J., Agić, Ž., Aranzabe, M. J., Asahara, M., Atutxa, A., Ballesteros, M., Bauer, J., Bengoetxea, K., Bhat, R. A., Bosco, C., et al. (2015). Universal dependencies 1.2.

Oquab, M., Bottou, L., Laptev, I., & Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1717–1724).

Östling, R. & Tiedemann, J. (2016). Continuous multilinguality with language vectors. *arXiv preprint arXiv:1612.07486*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics* (pp. 311–318).

Pasunuru, R. & Bansal, M. (2017). Multi-Task Video Captioning with Video and Entailment Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1273–1283).

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

Perfetti, C. A. (1998). The limits of co-occurrence: Tools and theories in language research.

Peters, B., Dehdari, J., & van Genabith, J. (2017). Massively multilingual neural grapheme-to-phoneme conversion. In *Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems* (pp. 19–26).

Peters, M., Ruder, S., & Smith, N. A. (2019). To tune or not to tune? adapting pretrained representations to diverse tasks. *arXiv preprint arXiv:1903.05987*.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, MA: MIT Press.

Plank, B. (2016). Keystroke dynamics as signal for shallow syntactic parsing. In *26th International Conference on Computational Linguistics* (pp. 609–619).

Poerner, N., Roth, B., & Schütze, H. (2018a). Interpretable textual neuron representations for nlp. *arXiv preprint arXiv:1809.07291*.

Poerner, N., Schütze, H., & Roth, B. (2018b). Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1 (pp. 340–350).

Pratt, L. Y. (1993). Discriminability-based transfer between neural networks. In *Advances in neural information processing systems* (pp. 204–211).

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature reviews neuroscience*, 6(7), 576.

Quine, W. (1960). *Word and Object*. Cambridge, MA: Cambridge University Press.

Radinsky, K., Agichtein, E., Gabrilovich, E., & Markovitch, S. (2011). A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web* (pp. 337–346).: ACM.

Radovanović, M., Nanopoulos, A., & Ivanović, M. (2010). On the existence of obstinate results in vector space models. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval* (pp. 186–193).: ACM.

Rajendran, J., Khapra, M. M., Chandar, S., & Ravindran, B. (2015). Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519*.

Rao, B. R. (1969). Partial canonical correlations. *Trabajos de estadistica y de investigación operativa*, 20(2), 211–219.

Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (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* (pp. 139–147).: Association for Computational Linguistics.

Recchia, G. & Jones, M. N. (2012). The semantic richness of abstract concepts. *Frontiers in Human Neuroscience*, 6.

Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., & Lee, H. (2016). Generative adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive Science: A Multidisciplinary Journal*, 29, 819–865.

Ribeiro, M. T., Singh, S., & Guestrin, C. (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* (pp. 1135–1144).: ACM.

Riordan, B. & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2), 303–345.

Robinson, A. & Fallside, F. (1987). *The utility driven dynamic error propagation network*. University of Cambridge Department of Engineering.

Rocktäschel, T., Grefenstette, E., Hermann, K. M., Kočiskỳ, T., & Blunsom, P. (2016). Reasoning about entailment with neural attention. In *International Conference on Learning Representations (ICLR)*.

Rotman, G., Vulić, I., & Reichart, R. (2018). Bridging languages through images with deep partial canonical correlation analysis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1 (pp. 910–921).

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

Rubenstein, H. & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10), 627–633.

Ruder, S., Bingel, J., Augenstein, I., & Søgaard, A. (2017). Sluice networks: Learning what to share between loosely related tasks. *arXiv preprint arXiv:1705.08142*.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation.* Technical report, California Univ San Diego La Jolla Inst for Cognitive Science.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3), 211–252.

Saha, A., Khapra, M. M., Chandar, S., Rajendran, J., & Cho, K. (2016). A correlational encoder decoder architecture for pivot based sequence generation. In *26th International Conference on Computational Linguistics: Technical Papers* (pp. 109–118).

Sanh, V., Wolf, T., & Ruder, S. (2018). A hierarchical multi-task approach for learning embeddings from semantic tasks. *arXiv preprint arXiv:1811.06031*.

Schafer, C. & Yarowsky, D. (2002). Inducing translation lexicons via diverse similarity measures and bridge languages. In *proceedings of the 6th conference on Natural language learning-Volume 20* (pp. 1–7).: Association for Computational Linguistics.

Schuster, M. & Nakajima, K. (2012). Japanese and korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5149–5152).

Schuster, M. & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11), 2673–2681.

Schwenk, H. & Gauvain, J.-L. (2005). Training neural network language models on very large corpora. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 201–208).: Association for Computational Linguistics.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Valerio Miceli Barone, A., Mokry, J., & Nădejde, M. (2017). Nematus: a Toolkit for Neural Machine Translation. (pp. 65–68).

Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 86–96).

Sennrich, R. & Kunz, B. (2014). Zmorge: A german morphological lexicon extracted from wiktionary. In *Language Resources and Evaluation Conference* (pp. 1063–1067).

Shah, K., Wang, J., & Specia, L. (2016). Shef-multimodal: Grounding machine translation on images. In *Proceedings of the First Conference on Machine Translation* (pp. 660–665).

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 806–813).

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *arXiv preprint arXiv:1704.02685*.

Silberer, C., Ferrari, V., & Lapata, M. (2017). Visually grounded meaning representations. *IEEE transactions on pattern analysis and machine intelligence*, 39(11), 2284–2297.

Silberer, C. & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1 (pp. 721–732).

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *International Conference on Learning Representation (ICLR) Workshop*.

Simonyan, K. & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2), 39–91.

Skocaj, D., Kristan, M., Vrecko, A., Mahnic, M., Janicek, M., Kruijff, G.-J. M., Hanheide, M., Hawes, N., Keller, T., Zillich, M., et al. (2011). A system for interactive learning in dialogue with a tutor. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on* (pp. 3387–3394).: IEEE.

Smith, L. B. & Yu, C. (2008). Infants rapidly learn word–referent mappings via cross-situational statistics. *Cognition*, 106(3), 1558–1568.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., & Ng, A. Y. (2014). Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2(1), 207–218.

Socher, R., Lin, C. C., Manning, C., & Ng, A. Y. (2011). Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)* (pp. 129–136).

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1631 (pp. 1642).

Specia, L., Frank, S., Sima'an, K., & Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation* (pp. 543–553).

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning* (pp. 843–852).

Subramanian, S., Trischler, A., Bengio, Y., & Pal, C. J. (2018). Learning general purpose distributed sentence representations via large scale multi-task learning. *arXiv preprint arXiv:1804.00079*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Advances in neural information processing systems* (pp. 3104–3112).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2014). Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2015). Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567.

Tai, K. S., Socher, R., & D. Manning, C. (2015). Improved semantic representations from tree-structured long short-term memory networks. In *Association for Computational Linguistic (ACL)*.

Tay, Y., Tuan, L. A., & Hui, S. C. (2018). Hyperbolic representation learning for fast and efficient neural question answering. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 583–591).: ACM.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Eight International Conference on Language Resources and Evaluation (LREC'12)*.

Tomasello, M. & Akhtar, N. (1995). Two-year-olds use pragmatic cues to differentiate reference to objects and actions. *Cognitive Development*, 10(2), 201–224.

Tomašev, N., Brehar, R., Mladenić, D., & Nedevschi, S. (2011a). The influence of hubness on nearest-neighbor methods in object recognition. In *2011 IEEE 7th International Conference on Intelligent Computer Communication and Processing* (pp. 367–374).: IEEE.

Tomašev, N., Radovanovic, M., Mladenic, D., & Ivanovic, M. (2011b). A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian knn. In *Proc. 20th ACM Int. Conf. on Information and Knowledge Management (CIKM)* (pp. 2173–2176).

Toyama, J., Misono, M., Suzuki, M., Nakayama, K., & Matsuo, Y. (2016). Neural machine translation with latent semantic of image and text. *CoRR*, abs/1611.08459.

Turing, A. M. (1950). I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236), 433–460.

Turney, P. D., Neuman, Y., Assaf, D., & Cohen, Y. (2011). Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on the Empirical Methods in Natural Language Processing* (pp. 680–690).

van Miltenburg, E., Kádar, A., Koolen, R., & Krahmer, E. (2018). Didec: The dutch image description and eye-tracking corpus. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3658–3669).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Vendrov, I., Kiros, R., Fidler, S., & Urtasun, R. (2016). Order-embeddings of images and language. *ICLR*.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2014). Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*.

Vilnis, L. & McCallum, A. (2014). Word representations via gaussian embedding. *arXiv preprint arXiv:1412.6623*.

Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I., & Hinton, G. (2015a). Grammar as a foreign language. In *Advances in Neural Information Processing Systems* (pp. 2755–2763).

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015b). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (pp. 3156–3164).: IEEE.

Visin, F., Kastner, K., Courville, A., Bengio, Y., Matteucci, M., & Cho, K. (2016). Reseg: A recurrent neural network for object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Vouloumanos, A. (2008). Fine-grained sensitivity to statistical information in adult word learning. *Cognition*, 107, 729–742.

Vouloumanos, A. & Werker, J. F. (2009). Infants' learning of novel words in a stochastic environment. *Developmental Psychology*, 45, 1611–1617.

Vulić, I., Kiela, D., Clark, S., & Moens, M.-F. (2016). Multi-modal representations for improved bilingual lexicon learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 188–194).: ACL.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. J. (1990). Phoneme recognition using time-delay neural networks. In *Readings in speech recognition* (pp. 393–404). Elsevier.

Wang, S. & Manning, C. D. (2012). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2* (pp. 90–94).: Association for Computational Linguistics.

Wang, S., Zhang, J., & Zong, C. (2018). Associative multichannel autoencoder for multimodal word representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 115–124).

Wang, W., Arora, R., Livescu, K., & Bilmes, J. (2015). On deep multi-view representation learning. In *International Conference on Machine Learning* (pp. 1083–1092).

Werbos, P. J. (1988). Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1(4), 339–356.

Weston, J., Bengio, S., & Usunier, N. (2010). Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1), 21–35.

Williams, A., Nangia, N., & Bowman, S. R. (2017). A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Williams, R. J. & Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. *Backpropagation: Theory, architectures, and applications*, 1, 433–486.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* (pp. 2048–2057).

Yang, D. & Powers, D. M. (2006). *Verb Similarity on the Taxonomy of WordNet*. Citeseer.

Yang, Y. & Hospedales, T. M. (2016). Trace norm regularised deep multi-task learning. *arXiv preprint arXiv:1606.04038*.

Yoo, K. M., Shin, Y., & Lee, S.-g. (2017). Improving visually grounded sentence representations with self-attention. *arXiv preprint arXiv:1712.00609*.

Yoshikawa, Y., Shigeto, Y., & Takeuchi, A. (2017). Stair captions: Constructing a large-scale japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 417–421). Vancouver, Canada: Association for Computational Linguistics.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. In *International Conference on Machine Learning (ICML)*.

Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (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.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: A computational study. *Connection science*, 17(3-4), 381–397.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language Learning and Development*, 4(1), 32–62.

Yu, C. & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13), 2149–2165.

Yu, C. & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5), 414.

Yu, H. & Siskind, J. M. (2013). Grounded language learning from video described with sentences. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1 (pp. 53–63).

Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2015). Video paragraph captioning using hierarchical recurrent neural networks. In *Describing and Understanding Video & The Large Scale Movie Description Challenge (LSMDC) at Internation Conference on Computer Vision (ICCV).*

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. In *Advances in neural information processing systems* (pp. 649–657).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015). Object detectors emerge in deep scene cnns. In *International Conference on Learning Representations (ICLR).*

Zipf, G. K. (1949). Human behavior and the principle of least effort.

Zoph, B., Yuret, D., May, J., & Knight, K. (2016). Transfer learning for low-resource neural machine translation. *arXiv preprint arXiv:1604.02201.*

TiCC PhD Series

1. Pashiera Barkhuysen. Audiovisual Prosody in Interaction. Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. Dendritic Morphology: Function Shapes Structure. Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. A Framework for Evidence-based Policy Making Using IT. Promotor: H.J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. Dialogue Act Recognition and Prediction. Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. Structured Prediction for Natural Language Processing. Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. New Architectures in Computer Chess. Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. Feature Extraction from Visual Data. Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. Multinomial Language Learning. Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. Digital Analysis of Paintings. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. Recommender Systems for Social Bookmarking. Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. Rapid Adaptation of Video Game AI. Promotor: H.J. van den Herik. Co- promotor: P. Spronck. Tilburg, 3 March 2010.

12. Maria Mos. Complex Lexical Items. Promotor: A.P.J. van den Bosch. Co-promotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).
13. Marieke van Erp. Accessing Natural History. Discoveries in data cleaning, structuring, and retrieval. Promotor: A.P.J. van den Bosch. Co-promotor: P.K. Lendvai. Tilburg, 30 June 2010.
14. Edwin Commandeur. Implicit Causality and Implicit Consequentiality in Language Comprehension. Promotores: L.G.M. Noordman, W. Vonk. Co-promotor: R. Cozijn. Tilburg, 30 June 2010.
15. Bart Bogaert. Cloud Content Contention. Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.
16. Xiaoyu Mao. Airport under Control. Promotores: H.J. van den Herik, E.O. Postma. Co- promotores: N. Roos, A. Salden. Tilburg, 25 May 2011.
17. Olga Petukhova. Multidimensional Dialogue Modelling. Promotor: H. Bunt. Tilburg, 1 September 2011.
18. Lisette Mol. Language in the Hands. Promotores: E.J. Krahmer, A.A. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (cum laude).
19. Herman Stehouwer. Statistical Language Models for Alternative Sequence Selection. Promotores: A.P.J. van den Bosch, H.J. van den Herik. Co-promotor: M.M. van Zaanen. Tilburg, 7 December 2011.
20. Terry Kakeeto-Aelen. Relationship Marketing for SMEs in Uganda. Promotores: J. Chr. van Dalen, H.J. van den Herik. Co-promotor: B.A. Van de Walle. Tilburg, 1 February 2012.
21. Suleman Shahid. Fun & Face: Exploring non-verbal expressions of emotion during playful interactions. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.
22. Thijs Vis. Intelligence, Politie en Veiligheidsdienst: Verenigbare Grootheden? Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).
23. Nancy Pascall. Engendering Technology Empowering Women. Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.

24. Agus Gunawan. Information Access for SMEs in Indonesia. Promotor: H.J. van den Herik. Co- promotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.
25. Giel van Lankveld. Quantifying Individual Player Differences. Promotores: H.J. van den Herik, A.R. Arntz. Co-promotor: P. Spronck. Tilburg, 27 February 2013.
26. Sander Wubben. Text-to-text Generation Using Monolingual Machine Translation. Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.
27. Jeroen Janssens. Outlier Selection and One-Class Classification. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.
28. Martijn Balsters. Expression and Perception of Emotions: The Case of Depression, Sadness and Fear. Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.
29. Lisanne van Weelden. Metaphor in Good Shape. Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 28 June 2013.
30. Ruud Koolen. "Need I say More? On Overspecification in Definite Reference." Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 20 September 2013.
31. J. Douglas Mastin. Exploring Infant Engagement. Language Socialization and Vocabulary. Development: A Study of Rural and Urban Communities in Mozambique. Promotor: A.A. Maes. Co-promotor: P.A. Vogt. Tilburg, 11 October 2013.
32. Philip C. Jackson. Jr. Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language. Promotores: H.C. Bunt, W.P.M. Daelemans. Tilburg, 22 April 2014.
33. Jorrig Vogels. Referential choices in language production: The Role of Accessibility. Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 23 April 2014.
34. Peter de Kock. Anticipating Criminal Behaviour. Promotores: H.J. van den Herik, J.C. Scholtes. Co-promotor: P. Spronck. Tilburg, 10 September 2014.

35. Constantijn Kaland. Prosodic marking of semantic contrasts: do speakers adapt to addressees? Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 1 October 2014.
36. Jasmina Marić. Web Communities, Immigration and Social Capital. Promotor: H.J. van den Herik. Co-promotores: R. Cozijn, M. Spotti. Tilburg, 18 November 2014.
37. Pauline Meesters. Intelligent Blauw. Promotores: H.J. van den Herik, T.A. de Roos. Tilburg, 1 December 2014.
38. Mandy Visser. Better use your head. How people learn to signal emotions in social contexts. Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 10 June 2015.
39. Sterling Hutchinson. How symbolic and embodied representations work in concert. Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2015.
40. Marieke Hoetjes. Talking hands. Reference in speech, gesture and sign. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 7 October 2015.
41. Elisabeth Lubinga. Stop HIV. Start talking? The effects of rhetorical figures in health messages on conversations among South African adolescents. Promotores: A.A. Maes, C.J.M. Jansen. Tilburg, 16 October 2015.
42. Janet Bagorogoza. Knowledge Management and High Performance. The Uganda Financial Institutions Models for HPO. Promotores: H.J. van den Herik, B. van der Walle, Tilburg, 24 November 2015.
43. Hans Westerbeek. Visual realism: Exploring effects on memory, language production, comprehension, and preference. Promotores: A.A. Maes, M.G.J. Swerts. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 10 Februari 2016.
44. Matje van de Camp. A link to the Past: Constructing Historical Social Networks from Unstructured Data. Promotores: A.P.J. van den Bosch, E.O. Postma. Tilburg, 2 Maart 2016.
45. Annemarie Quispel. Data for all: Data for all: How professionals and non-professionals in design use and evaluate information visualizations. Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 15 Juni 2016.

46. Rick Tillman. Language Matters: The Influence of Language and Language Use on Cognition Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 Juni 2016.
47. Ruud Mattheij. The Eyes Have It. Promoteres: E.O. Postma, H. J. Van den Herik, and P.H.M. Spronck. Tilburg, 5 October 2016.
48. Marten Pijl, Tracking of human motion over time. Promotores: E. H. L. Aarts, M. M. Louwerse Co-promotor: J. H. M. Korst. Tilburg, 14 December 2016.
49. Yevgen Matusevych, Learning constructions from bilingual exposure: Computational studies of argument structure acquisition. Promotor: A.M. Backus. Co-promotor: A.Alishahi. Tilburg, 19 December 2016.
50. Karin van Nispen. What can people with aphasia communicate with their hands? A study of representation techniques in pantomime and co-speech gesture. Promotor: E.J. Krahmer. Co-promotor: M. van de Sandt-Koenderman. Tilburg, 19 December 2016.
51. Adriana Baltaretu. Speaking of landmarks. How visual information inuences reference in spatial domains. Promotores: A.A. Maes and E.J. Krahmer. Tilburg, 22 December 2016.
52. Mohamed Abbadi. Casanova 2, a domain specific language for general game development. Promotores: A.A. Maes, P.H.M. Spronck and A. Cortesi. Co-promotor: G. Maggiore. Tilburg, 10 March 2017.
53. Shoshannah Tekofsky. You Are Who You Play You Are. Modelling Player Traits from Video Game Behavior. Promotores: E.O. Postma and P.H.M. Spronck. Tilburg, 19 Juni 2017.
54. Adel Alhuraibi, From IT-Business Strategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT. Promotores: H.J. van den Herik and Prof. dr. B.A. Van de Walle. Co-promotor: Dr. S. Ankolekar. Tilburg, 26 September 2017.
55. Wilma Latuny. The Power of Facial Expressions. Promotores: E.O. Postma and H.J. van den Herik. Tilburg, 29 September 2017.
56. Sylvia Huwaë, Different Cultures, Different Selves? Suppression of Emotions and Reactions to Transgressions across Cultures. Promotores: E.J. Krahmer and J. Schaafsma. Tilburg, 11 October, 2017.

57. Mariana Serras Pereira, A Multimodal Approach to Children's Deceptive Behavior. Promotor: M. Swerts. Co-promotor: S. Shahid Tilburg, 10 January, 2018.
58. Emmelyn Croes, Meeting Face-to-Face Online: The Effects of Video-Mediated Communication on Relationship Formation. Promotores: E.J. Krahmer and M. Antheunis. Co-promotor A.P. Schouten. Tilburg, 28 March 2018.
59. Lieke van Maastricht, Second Language Prosody: Intonation and Rhythm in Production and Perception. Promotores: E.J. Krahmer and M. Swerts. Tilburg, 9 May 2018.
60. Nanne van Noord, Learning visual representations of style. Promotores: E.O. Postma and M. Louwerse. Tilburg, 16 May 2018.
61. Ingrid Masson Carro, Handmade: On the Cognitive Origins of Gestural Representations. Promotores: E.J. Krahmer and M. Goudbeek. Tilburg, 25 June 2018.
62. Bart Joosten, Detecting Social Signals with Spatiotemporal Gabor Filters. Promotores: E.J. Krahmer and E.O. Postma. Tilburg, 29 June 2018.
63. Yan Gu, Chinese hands of time: The effects of language and culture on temporal gestures and spatio-temporal reasoning. Promotor: M. Swerts. Co-promotores: M.W. Hoetjes, R. Cozijn. Tilburg, 5 June 2018.
64. Thiago Castro Ferreira, Advances in Natural Language Generation: Generating Varied Outputs from Semantice Inputs. Promotor: E.J. Krahmer. Co-promotor: S. Wubben. Tilburg, 19 September 2018.