Multilingual Grounded learning

Ákos Kádár

Multilingual Grounded learning

Ákos Kádár
PhD Thesis
Tilburg University, 2018

TiCC PhD Series No. 64

Financial Support was received from NWO

Cover design: Reka Kadar
Design:
Print:

Multilingual Grounded learning

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

Commissieleden

Prof. Dr. A. Gatt

Prof. Dr. A. van den Bosch

Dr. C. Gardent

Dr. F. Schilder

Dr. M.B. Goudbeek

"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

Acknowledgements

Contents

1 Introduction

Traditional distributional representations of linguistic units consider features extracted from large text-corpora in a single language. The aim of the presented work is to learn representations of words, phrases up to sentences from multiple sources of information. Specifically I focus on jointly learning representations for multiple languages, grounding in visual modality and the interaction of these two views. Grounding in visual modality is largely motivated by evidence of perceptual grounding in human concept acquisition and representation Barsalou et al. (2003) and as such it brings computational language learning systems closer to human-like learning. Harnessing the visual modality to learn language representations that link linguistic knowledge to the external world Kiela et al. (2014); Baroni (2016); Elliott & Kádár (2017); Kiela et al. (2017); Yoo et al. (2017) has been empirically shown to improve performance on several semantic tasks

such as paraphrase identification, semantic entailment Dolan et al. (2004); Marelli et al. (2014). Traditionally, given a task for each language separate corpora are created and separate sets of parameters are learned. Learning shared cross-lingual representations, however, allows researchers and practitioners to train a single model allowing to transfer knowledge between languages leading to practical consequences such as mitigating the low-resource problem, cross-lingual applications and processing data with code-mixing. Multilingual systems also offer a fertile ground computational typology as shared models are forced to exploit cross-lingual regularities. The first part of the thesis Chapter 1 and 2. focuses on learning grounded representations for a single language. In Chapter 1. we start by learning visual representations for words using a novel computational cognitive model of cross-situational word learning that takes words and high-level continuous image representations as input. Chapter 2. introduces a modern recurrent and convolutional neural network based model that learns from both visual-grounding signals and word-word co-occurrences. Furthermore, we introduce a technique to interpret the learned representations of such architecture and investigate if certain linguistic phenomena is encoded in the learned model. The second part of the thesis focuses on multimodal multilingual models. Chapter 3. provides evidence that grounded learning can potentially improve machine translation and in Chapter 4. we show under what conditions multilinguality an help improve grounded representations.

Continuous representations of language

My work focuses on learning grounded continuous language embeddings. The first half of the thesis is about learning the representa-

tions of words and followed-up by learning sentence embeddings for a single language. Later we move on to learning multilingual visually grounded representations. In other works of mine I contributed to generating referring expressions grounded in Wikipedia entities and learning language through interacting with text-games grounded in a knowledge-base.

Continuous 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 the earlier days of American structuralism Nevin & Johnson (2002). In the seminal paper "Distributional structure" Harris (1954) Harris already claims back in 1954 that distribution should be taken as an explanation for word meaning and that similarity classes can be constructed based on co-occurrence statistics. Cruse & Cruse (1986) writes 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'. One of the first computational verification attempts of this distributional hypothesis was put forward in Miller & Charles (1991). Here we discuss computational models of distributional word representations using the 'count-based' vs. 'prediction-based' distinction borrowed from Baroni et al. (2014). Early computational linguistics models of distributional semantics fall in the count-based approaches: they count the number of times target words appear occur in different contexts resulting in a co-occurence matrix. To these matrices various re-weighting schemes are applied usually followed by some dimension-

ality reduction technique. One notable approach is 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 pointwise mutual information and local mutual information Evert (2005) or different dimensionality reduction techniques such as non-negative matrix factorization Baroni et al. (2014). For a comprehensive empirical experiments on count-based approaches please consult Bullinaria & Levy (2007) and Bullinaria & Levy (2012). In more recent years – and more related to the present thesis – various deep learning methods have been applied to learn continuous word-representations usually referred to as ´word-embeddings´ in the literature. Contrary to count-based methods prediction-based approaches fit into the standard supervised 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. The first modern approach to neural language models on realistic data sets was introduced in Bengio et al. (2003). It is a feed-forward multilayer perceptron with continuous word-embeddings, a single hidden layer and a softmax output layer. The model is trained to maximize the probability of the target word given the previous two words as context – 2-gram language model – trained by stochastic gradient descent Cauchy (1847) through the backpropagation algorithm Rumelhart et al. (1985). The superior performance of the feed-forward neural language model on language modeling has soon been shown to improve performance in speech recognition Schwenk & Gauvain (2005). The convolutional architecture of Collobert & Weston (2008) based

on the time-delay neural network model Waibel et al. (1990) takes several steps towards the by now standard practices in neural NLP. Contrary to the simple feed-forward network of Bengio et al. (2003) the convolution over-time structure can handle sequences of variable length, which is essential in NLP applications as sentences contain varying number of words. The paper also introduces the idea jointly learning many tasks at the same time such as part-of-speech tagging, chunking, named entity recognition and semantic role labeling through multi-task learning. Finally Collobert & Weston (2008) were the first to show the utility of pre-trained word-embeddings in other tasks. This architecture was later refined in Collobert et al. (2011) and the pre-trained full model was made available alongside the standalone word-embeddings in the SENNA toolkit. The architecture, however, that arguably became most popular in NLP and is used in all my papers except for Chapter 1. is the recurrent neural network. It was argued that that the simple recurrent network architecture introduced by Elman is difficult to train for practical applications on longer sequences Bengio et al. (1994), never the less the RNNLM implementation Mikolov et al. (2010) established a new state-of-the-art. The word-embeddings learned by the recurrent language model were shown to represent syntactic and semantic regularities Mikolov et al. (2013d). In earlier work pre-trained word-embeddings were also shown through a larger-scale empirical investigation by Turian et al. (2010) to transfer to sequence prediction tasks when used as features in state-of-the-art systems. However, it wasn't until the introduction of the much simpler and faster CBOW and skip-thought algorithms packaged into the easy to use word2vec toolkit that word-embeddings became ubiquitous in computational linguistics and NLP research.

Mention count-based methods for completeness, but dont worry about it cite baroni guys. Do the SENNA, Turian, Skip-gram and GloVe I guess.

From words to sentences

Earlier work developed under the framework of compositional distributional semantics produce continuous representations for phrases up to sentences using additive and multiplicative interactions of 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 unified theory of distributional semantics and type logical grammars based on pregroups Coecke et al. (2010). From the more practical point of view of learning transferable distributed sentence representations – akin to pre-trained word-embeddings – the first notable approach is the skip-thought vectors model Kiros et al. (2015). This method was confirmed to be a successful unsupervised method for transfer learning in a number of sentence classifications 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). Later approaches have also focused on identifying supervised tasks such as natural language inference Conneau et al. (2017) that lead to representations that generalize well to other tasks or to combine a number of supervised tasks with unsupervised training through multi-task learning Subramanian et al. (2018). The state-of-the-art in learning universal sentence representations at the time of writing the present thesis is represented by ELMo approach Peters et al. (2018), which trains a stack convolution and recurrent layers large-scale bidi-

rectional language modeling. For each task the forward and backward representations are extracted from the model for each word position and then fed to a task specific recurrent network. Such contextualized word-representations have been explored before ELMo using the LSTM states of machine translation encoders McCann et al. (2017) pose a powerful alternative to universal sentence encoders.

Visually grounded representations of words and sentences

The approaches to learn distributed representations discussed so far focus on 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 the representations. Interestingly the articles introducing Latent Semantic Analysis Landauer & Dumais (1997) and Hyperspace Analogue to Language Lund & Burgess (1996) already mention that a possible limitation of 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." On the defense of textual models Louwerse (2011) argue that the corpora used to train distributional semantic models are generated by humans and as such reflect the perceptual world. In practice, however, much work on multi-modal distributional semantics have found that text-only spaces tend to represent more encyclopedic knowledge, whereas mutli-modal representations capture more concrete aspects Andrews et al. (2009); Baroni & Lenci (2008). A famous example is that few pieces of text are going to state the obvious fact that "bananas are yellow" Bruni et al. (2014). However, one does not necessarily need to reach a conclusion

on whether grounded or distributional models are superior, rather combining their merits in a pragmatic way is an attractive alternative Riordan & Jones (2011). x Learning representation from both linguistic and visual input is a step towards realistic models of language learners, however, it is as close as assuming children learn language by sitting still and watching TV. Probably the first approach to visual word representations Feng & Lapata (2010). Perceptually grounded word representations Bruni et al. (2012). Using bag-of-visual-words Bruni et al. (2011). For a big review check Bruni et al. (2014). Modern skip-gram + VGG word embeddings Kiela & Bottou (2014). Multi-modal skip-gram Lazaridou et al. (2015b). MMFEAT toolkit for internet search style multimodal word embeddings Kiela (2016). Good results with image search style on STS Glavaš et al. (2017). Cross-modal ranking models: Deep belief network to learn a joint probabilistic generative model of images and tags Srivastava & Salakhutdinov (2012), cross-modal correspondence auto-encoder Feng et al. (2014), image annotation with joint embeddings Weston et al. (2010). Grounded sentence representations for ranking with dependency tree recursive networks Socher et al. (2014).Unified RNN-CNN architecture for captioning and ranking Kiros et al. (2014a). Learning to hash image-caption pairs Jiang & Li (2016); Cao et al. (2016).

Concatenating grounded sentence embeddings with the skip-thought embeddings lead to improvements on a large number of semantic sentence classifications tasks Kiela et al. (2017). We show that in the domain of image-captions grounded learning improves translation quality and that learning multi-modal representations provides gains on top of learning from larger bilingual corpora Elliott & Kádár

(2017). This lead us to the hypothesis that much of the observed improvements of multi-modal translation models over text-only baselines is due to grounded learning and not to the effective use of visual context. LATER DESMOND SHOWS THAT MODELS ARE NOT SENSITIVE TO CONTEXT CITE THE EMNLP PAPER. SOMEOW MENTION THE FOLLOWUP HUMAN STUDY TOO.

Multilingual representations of words and sentences

Multilingual or cross-lingual word embedding approaches map words of different languages in a shared space. These representations allow to transfer knowledge between languages and perform cross-lingual tasks such as cross-lingual document retrieval. The bulk of such approaches apply the techniques developed for monolingual word-embedding induction as a component of their full pipeline. One of the early approaches trains two separate word spaces with using the negative-sampling implementation of the skip-gram algorithm and then maps these spaces onto eachother by minimizing the squared loss between the 5000 top most frequent word-vectors Mikolov et al. (2013b). This main approach went through improvents such as by switching mean-squared error to ranking loss Lazaridou et al. (2015a) or enforcing orthogonality constraint on the mapping matrix Xing et al. (2015) (WE DID THE SAME OVER TIME). Another prominent technique to learn a mapping between word (and other types of) spaces is Canonical Correlation Analysis, first used by Faruqui & Dyer (2014), which was later improved by the application of Deep CCA Lu et al. (2015). Another approach is to create a pseudo-bilingual corpora by replacing words in a source language by their translation and train an embedding model on the resulting corpus

The uber multilingual word-embedding Ammar et al. (2016b).

Visually grounded multilingual sentences representations

Now do the multi-lingual multimodal stuff and put the IJCNLP and ConLL papers

Grounding and other modalities

Grounding words in auditory signals Kiela & Clark (2015); Lopopolo & Miltenburg (2015) and olfactory perception Kiela et al. (2015)r. Growing body of literature in learning joint representations of speech and images Harwath et al. (2016); Chrupała et al. (2017); Harwath et al. (2018b)

Mention semantic parsing and the WebNLG paper i did with Thiago. Mention the TextWorld thing. Grzegorz and Afra´s work speech, plus the fail paper that im on too.

Interpreting continuous representations

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). Another angle of model interpretability is to train a complex opaque model and use them to discover patterns in the input data that are crucial in solving the task. Deep neural networks learn to solve tasks from close to raw input representations, however, the regularities they uncover from the input signal during training is opaque. As such there is a growing

interest in deriving methods to *explanain* the decision of such architectures. These methods assign a real-valued "relevance" score to each unit in the input signal, which signifies how much impact it had on the final prediction of the model. One of the most well researched paradigms in generating such relevance scores is gradient based methods: they take the gradient of the ouput of the network with respect to the input Simonyan et al. (2013). Deep neural models of language tasks learn distributed representations of input symbols and as such further operations have to applied to reduce the resulting gradient vectors to single scalars e.g. using ℓ_2 norm Bansal et al. (2016). Another prominent approach is layerwise relevance propagation Bach et al. (2015)

Here put the CL paper, but also refer to the other stuff. I will also mention the interpretable models and the caveat of the complexity of these things with the Bowman negative result and my HMLSTM result.

Integrating Vision and Language

Mention the bunch of work thats here. Do image-captioning and image-sentence ranking and then metion that we do the latter always. Say the importance of attention in this field and mention the DIDEC corpus. Mention VQA and fine-grained reasoning and the FigureQA paper.

Methods

Data sets

For image-feature extraction all approaches presented in the thesis use some CNN architecture pre-trained on a version of the ImageNet data set Deng et al. (2009) prepared for a large-scale image classification challenge Russakovsky et al. (2015). ImageNet annotates the noun synsets from WordNet Miller (1995) with images collected from the internet. At the time of writing this thesis on average the data set consists of 500 images per node. The subsection used in the pre-trained networks is the ILSVRC2012 containing 1.2 million images annotated with 1000 classes. For learning grounded language representation we use data sets introduced for image-captioning. These data sets annotate images found in online resources with descriptions through crowd-sourcing. The smallest data set we use is Flickr8K Hodosh et al. (2013a) with 8000 images, which was later extended to contain 30K images resulting in the Flickr30K benchmark Young et al. (2014) and finally the largest collection we use is the COCO image-caption benchmark Chen et al. (2015b) with 123K images. All data sets contain 5 captions per image. The descriptions collected for these data sets 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 (frizbee, dog, boy) their attributes (yellow, fluffy, young) and their relations between

each other. The images depict common real-life scenes such as bus turning left or people playing soccer in the park. As such annotation collected independently from different 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). For our multi-lingual experiments we use the Multi30K data set Elliott et al. (2016a, 2017a); a multilingual extension of Flickr30K. It consists of a *translation* and a *comparable* portions both containing all images from the original Flickr30K. The *translation* portion annotates a single English description with a German, French and Czech caption, while the *comparable* the original 5 English and additional 5 German descriptions collected independently using the same crowdsourcing process.

Architectures

Just do the usual GRU + CNN thing. cite Jamie (heart).

Recurrent network

Do the GRU mention that the LSTM is strictly more powerful according to Yoav, but it doesnt seem to matter at all according to Chung.

Convolutional network

Just super quick run-down on VGG and ResNet.

Optimization

Just mention that all this shit is optimized with some version of SGD and we use Adam all the time. Mention that Adam sucks but its just some common thing to do. Say this thing about the MSE or cosine loss that we start with that but i results in hubness actually as Angeliki points it out. So we move to the contrastive kinda losses.

Transfer learning

Don't really worry about it too much just say something about how awesome is that the CNNs trained on image net actually end up being so useful for us.

Multi-task learning

Multi-task learning has been an important paradigm since the earliest works on training neural architectures for natural-language processing Collobert & Weston (2008). Multi-task learning involves optimizing multiple loss function jointly usually with the goal of exploiting commonalities between tasks. As described in the seminal work of Caruana (1997a):

> Multitask Learning is an approach to inductive transfer that improves generalization by using the domain information contained in the training signals of related tasks as an inductive bias. It does this by learning tasks in parallel while using a shared representation; what is learned for each task can help other tasks be learned better.

Introducing inductive bias using different tasks means that the learning algorithm prefers a specific set of hypotheses over others just as in the case of other forms of regularization such as ℓ_1 or ℓ_2 penalties. Here I only discuss the specific instantiation of the mutli-task learning paradigm that are applicable to the chapters presented in this thesis and for a more comprehensive overview please consult Ruder (2017). The technique used in the thesis is hard-parameter sharing Caruana (1997a)

2

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 Castro Ferreira, T., Krahmer, E., & Wubben, S. (2016). Individual variation in the choice of referential form. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT'2016 (pp. 423-427). San Diego, California: Association for Computational Linguistics.

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.

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. (2010). 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. (2010).

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).

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 & Fei-Fei, 2014), (Mao et al., 2014b), (Kiros et al., 2014b), (Donahue et al., 2014), (Vinyals et al., 2014), (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., 2014) 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., 2015b) 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.* Thus, strictly speaking their model does not implement cross-situational learning because a subset of words is unambiguously associated with certain images.

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 2.5).

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

*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.

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. (2010). 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 im-

age 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.

Visual input

The features of the images are extracted by training a 16-layer convolutional neural network (CNN) Simonyan & Zisserman (2014) on an object recognition task.[†] 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 2.1 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.

[†]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 2.1: *Dimensions with three most closely aligned images from F8k.*

Learning algorithm

We adapt the IBM model 1 estimation algorithm in the following ways[‡]: (i) like (Fazly et al., 2010) 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

[‡]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.

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[§].

- 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,c). WORD2VEC learns a vector representation (embedding) of a word which maximizes performance on predicting words in a small window around it.

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

Experiments

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 & Fei-Fei, 2014), leaving out 1000 images for validation and 1000 for testing from each set. Table 2.1 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 2.1: *Flickr image caption datasets.*

For the Single-concept image descriptions experiments reported in section 2.3.4, we also use the ILSVRC2012 subset of ImageNet Russakovsky et al. (2014), 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.

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.¶

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

¶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.

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 model-

ing 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 2.2 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.

	#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 2.2: *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.*

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.

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[‖] 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

$$P(w_i|i_j) \propto P(i_j|w_i) \times P(w_i), \tag{2.1}$$

where w_i is a word from the vocabulary of the captions and i_j

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

is an image from the collections of images I. The probability of an image given a word is defined as

$$P(i_j|w_i) = \frac{\text{cosine}(i_j, w_i)}{\sum_{k=1}^{|I|} \text{cosine}(i_k, w_i),} \tag{2.2}$$

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 2.3, where N is the number of words in the vocabulary.

$$P(w_i) = \frac{\log(count(w_i))}{\sum_{j=1}^{N} \log(count(w_j))} \tag{2.3}$$

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 2.2 shows an example of an image and its multi-word captions in the validation portion of the F30k dataset.

Single-concept image descriptions

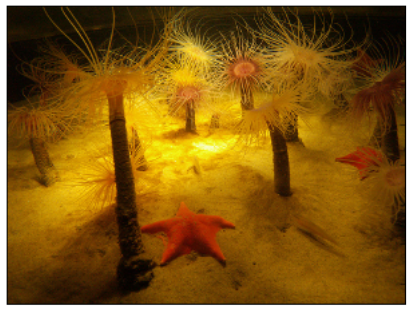
Even though we use separate portions of F8k and F30k for training and testing, these subsets are still very similar. To test how general the VISUAL word representations are, we use images from the ILSVRC2012 subset of ImageNet Russakovsky et al. (2014) as benchmark. The major difference between these images and the ones from F8k and F30k datasets is that labels of the images in ImageNet are

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 2.2: *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.*



Label: sea anemone anemone
Hypernym: animal

Figure 2.3: *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.*

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 2.3 illustrates these issues.

Results

We evaluate our model on two main tasks: simulating human judgments of word similarity[**] 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[††] Efron (1982).

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[‡‡] 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 2.4 illustrates the correlation of the similarity judgments by the three models with those of humans on four datasets. Table 2.3

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

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

[‡‡]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.

Concreteness

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

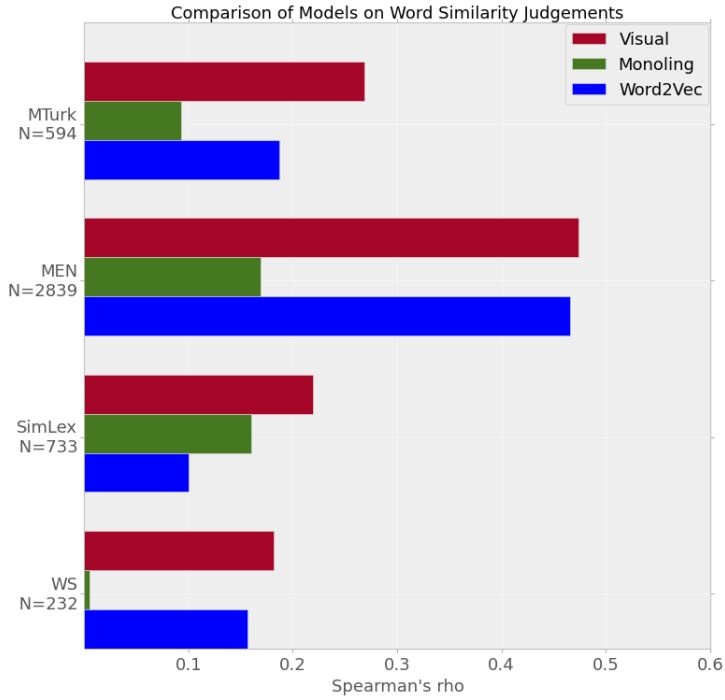


Figure 2.4: *Comparison of models on approximating word similarity judgments. The length of the bars indicate the size of the correlation measured by Spearman's ρ, longer bars indicate better similarity between the models' predictions and the human data. The labels on the y-axis contain the names of 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 2.3: *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 2.3.3, 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 2.4 only reports the results on MEN and Simlex999 as these were the only benchmarks that had at least 200 word-pairs after partitioning. Table 2.2 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

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

Table 2.4: *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* $p < 0.05$.

there is no significant difference in its performance given the different concreteness levels of the word pairs. Splitting the word pairs in two 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.

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 2.3.4. These experiments can be viewed as computational simulations of a language task where human subjects associate words to given images. Words were ranked accord-

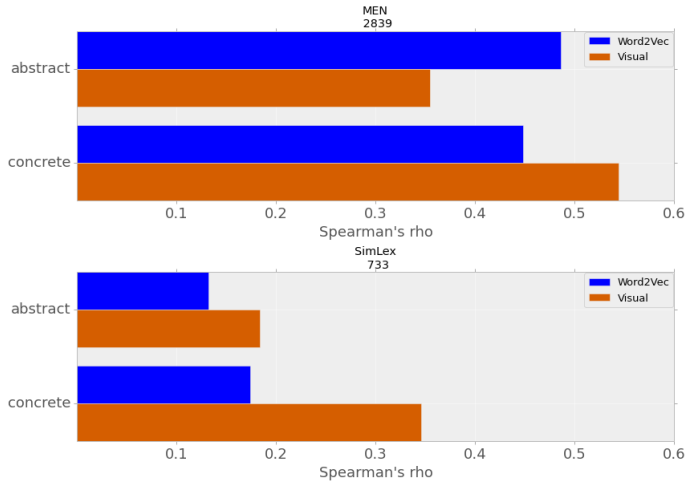


Figure 2.5: *Models' performance on word similarity judgments as a function of the concreteness of the word pairs.*

ing 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.

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 2.3.4. Table 2.5 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*, 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 2.5). 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.

	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 2.5: *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.*

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 ILSVRC2012 portion of ImageNet Russakovsky et al. (2014). 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 2.6, 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 Figureẽffig: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 be seen in the examples in Figure 2.6, 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. (2014) would be largely beneficial.

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 mul-

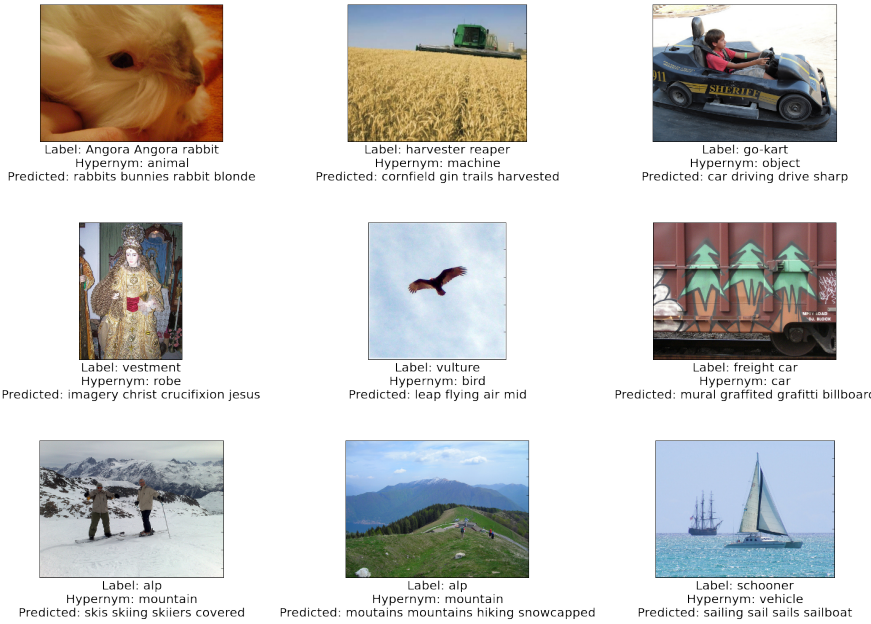


Figure 2.6: *The caption above the images show the target labels, the hypernyms that were considered as a new target if the original was not in the vocabulary and the top N predicted words. In a large number of cases the guesses of the model are conceptually similar to the images, although, do not actually overlap with the labels or the hypernyms.*

tiword 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 2.1.2). We are currently developing this style of integrated model to investigate the impact of structure on word learning from a cognitive point of view.

3

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.

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. (2015a), 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., 2015a). 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.

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., 2015a) 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 con-

clude 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. (2013); 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. (2013); 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.

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.

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 $\tilde{\mathbf{h}}_t$:

$$\text{GRU}(\mathbf{h}_{t-1}, \mathbf{x}_t) = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \tag{3.1}$$

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

$$\mathbf{z}_t = \sigma_s(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1}) \tag{3.2}$$

The candidate activation is computed as:

$$\tilde{\mathbf{h}}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1})) \tag{3.3}$$

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:

$$\mathbf{r}_t = \sigma_s(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1}) \tag{3.4}$$

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

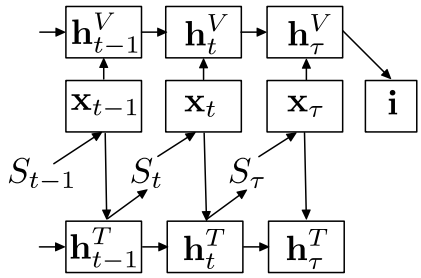


Figure 3.1: Structure of IMAGINET, adapted from Chrupała et al. (2015a).

Imaginet

IMAGINET introduced in (Chrupała et al., 2015a) 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 3.1 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 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:

$$\mathbf{h}_t^V = \text{GRU}^V(\mathbf{h}_{t-1}^V, \mathbf{x}_t) \tag{3.5}$$

$$\mathbf{h}_t^T = \text{GRU}^T(\mathbf{h}_{t-1}^T, \mathbf{x}_t) \tag{3.6}$$

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* $\hat{\mathbf{i}}$ from its last hidden representation $\mathbf{h}_t^V$.

$$\hat{\mathbf{i}} = \mathbf{V}\mathbf{h}_\tau^V \tag{3.7}$$

$$p(S_{t+1}|S_{1:t}) = \text{softmax}(\mathbf{L}\mathbf{h}_t^T) \tag{3.8}$$

The loss function is a multi-task objective which penalizes error on the visual and the textual targets simultaneously. The objective combines cross-entropy loss L^T for the word predictions and cosine distance L^V for the image predictions†, weighting them with the parameter α (set to 0.1).

$$L^T(\theta) = -\frac{1}{\tau}\sum_{t=1}^{\tau} \log p(S_t|S_{1:t}) \tag{3.9}$$

$$L^V(\theta) = 1 - \frac{\hat{\mathbf{i}} \cdot \mathbf{i}}{\|\hat{\mathbf{i}}\|\|\mathbf{i}\|} \tag{3.10}$$

$$L = \alpha L^T + (1-\alpha)L^V \tag{3.11}$$

For more details about the IMAGINET model and its performance

*Representing the full image, extracted from the pre-trained Convolutional Neural Network of (Simonyan & Zisserman, 2014).

†Note that the original formulation in (Chrupała et al., 2015a) 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.

see (Chrupała et al., 2015a). 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.

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.

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.

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 3.4.1 we introduce *omission score*, a metric to measure the contribution of each token to the prediction of the networks, and in Section 3.4.2 we analyze how omission scores are dis-

tributed over dependency relations and part-of-speech categories. In Section 3.4.3 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 3.4.4 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 target image representations were extracted from the pre-softmax layer of the 16-layer CNN of (Simonyan & Zisserman, 2014).

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 activation 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

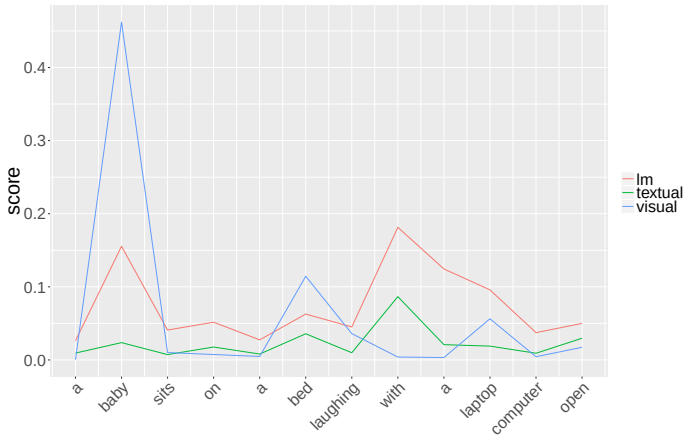


Figure 3.2: Omission scores for the example sentence *a baby sits on a bed laughing with a laptop computer open* for LM and the two pathways, TEXTUAL and VISUAL, of IMAGINET.

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:

$$\text{omission}(i, S) = 1 - \text{cosine}(\mathbf{h}_{\text{end}}(S), \mathbf{h}_{\text{end}}(S_{\setminus i})) \tag{3.12}$$

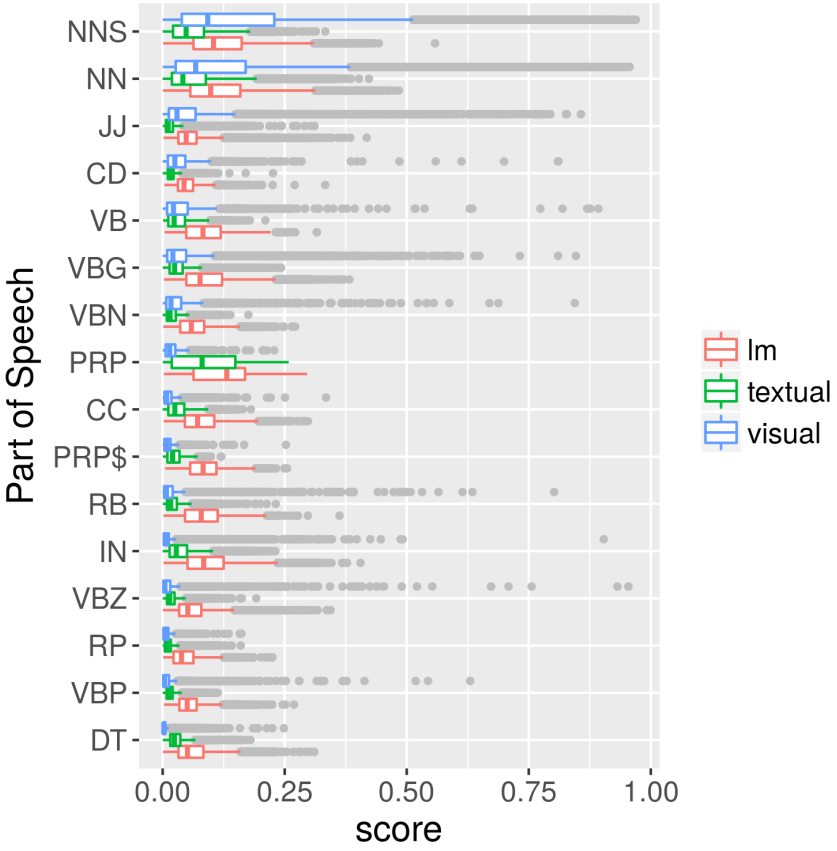
Figure 3.2 demonstrates the omission scores for the LM, VISUAL and TEXTUAL models for an example caption. Figure 3.3 shows the images retrieved by VISUAL for the full caption and for the one with the word *baby* omitted. The images are retrieved from the validation set of MS-COCO by: 1) computing the image representation of the

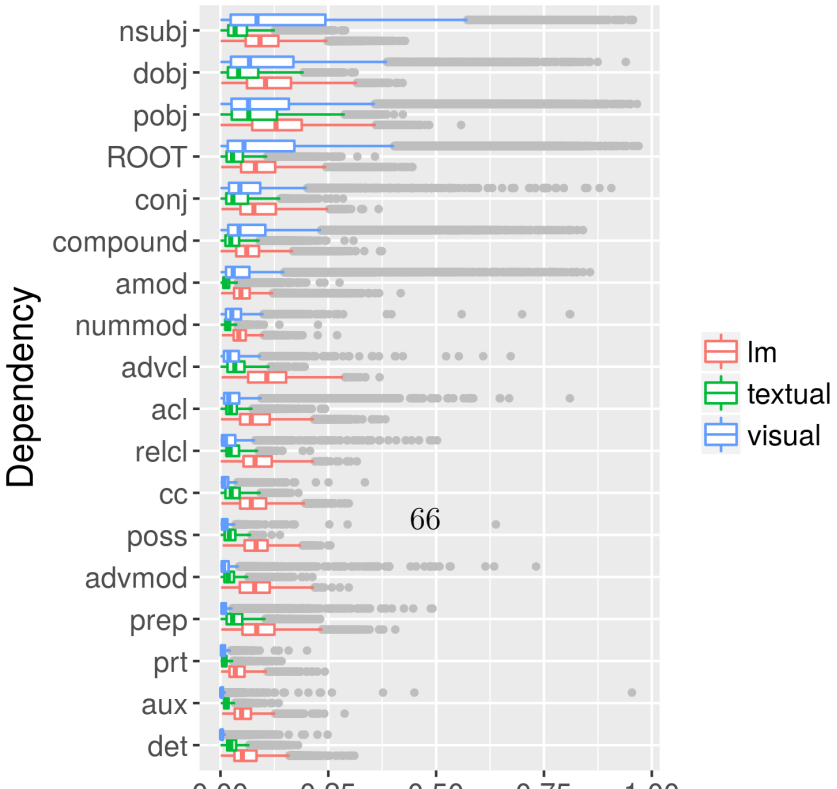

Figure 3.3: Images retrieved for the example sentence *a baby sits on a bed laughing with a laptop computer open* (left) and the same sentence with the second word omitted (right).

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 3.2 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.

Omission score distributions

The omission scores can be used not only to estimate the importance of individual words, but also of syntactic categories. We estimate


Part of Speech
NNS
NN
JJ
CD
VB
VBG
VBN
PRP
CC
PRP$
RB
IN
VBZ
RP
VBP
DT
0.00
0.25
0.50
0.75
1.00
score
lm
textual
visual


Dependency
nsubj
dobj
pobj
ROOT
conj
compound
amod
nummod
advcl
acl
relcl
cc
poss
advmod
prep
prt
aux
det
lm
textual
visual

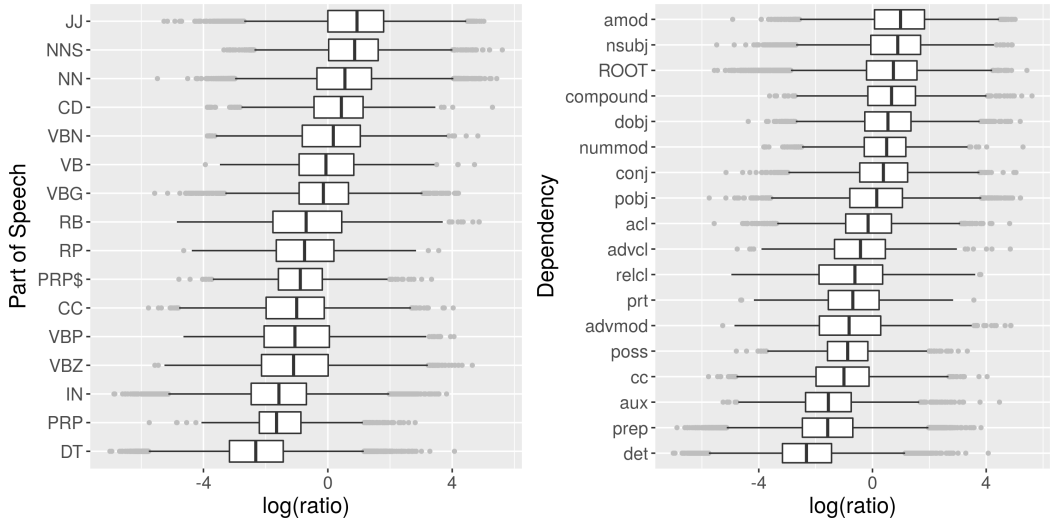


Figure 3.5: Distributions of log ratios of omission scores of TEXTUAL to VISUAL per POS (left) and dependency labels (right). Only labels which occur at least 1250 times are included.

the 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.[‡]

Figure 3.4 shows the distribution of omission scores per POS and dependency label for the two pathways of IMAGINET and for LM.[§]

[‡]Available at `https://spacy.io/`.

[§]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 extend past the lower and upper quartile to 1.5× the interquartile range (i.e.

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 3.5 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

75th percentile - 25th percentile); the points are outliers.

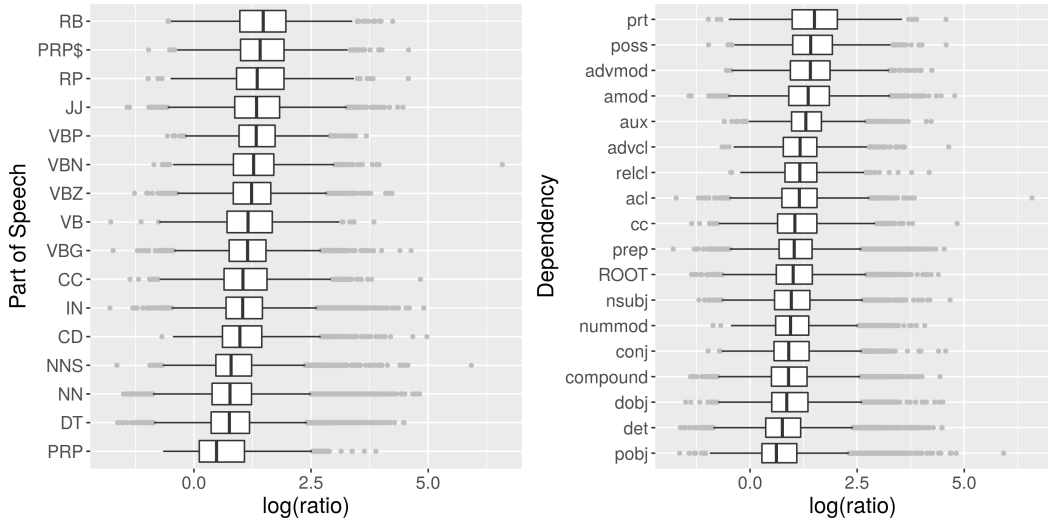


Figure 3.6: Distributions of log ratios of omission scores of LM to TEXTUAL per POS (left) and dependency labels (right). Only labels which occur at least 1250 times are included.

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 3.6 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.

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 3.4.2 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

Table 3.1: 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

4. LR FULL: word type, dependency label, position, word:dependency interaction, word:position interaction

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 3.1 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 3.7 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.

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

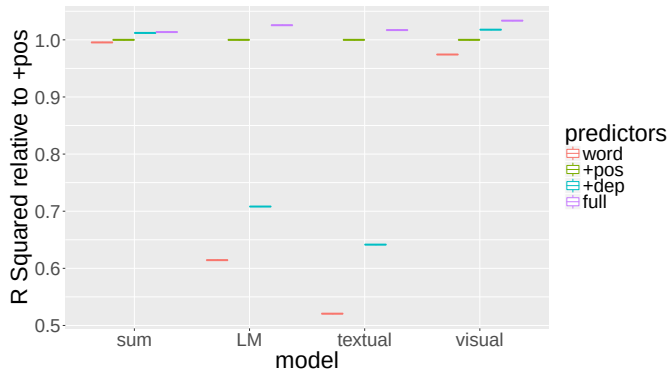


Figure 3.7: Proportion of variance in omission scores explained by the linear regression models for SUM, LM, VISUAL and TEXTUAL, relative to regressing on word identity and position only.

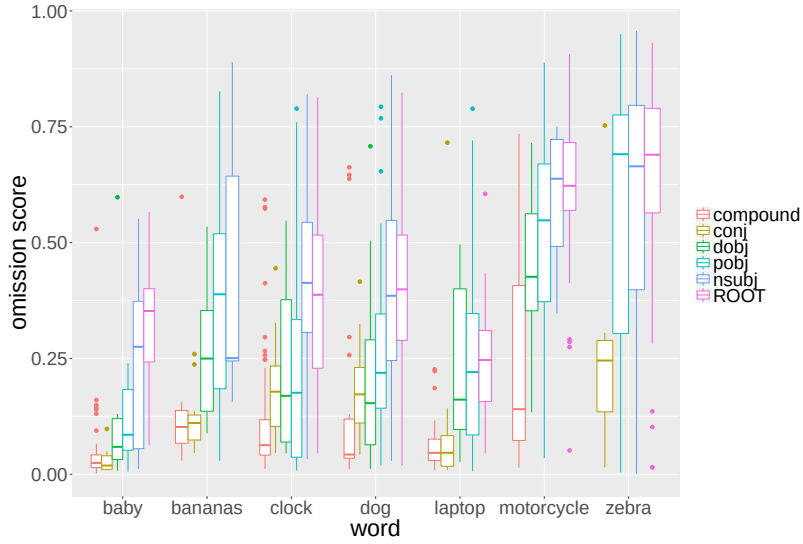


Figure 3.8: Distribution of omission scores per dependency label for the selected word types.

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

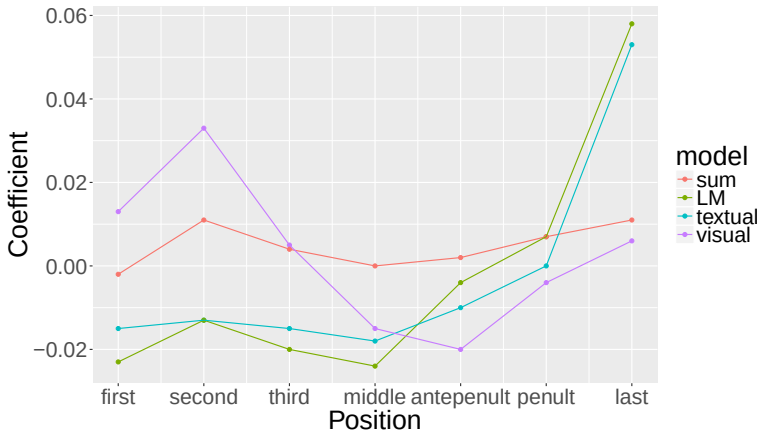


Figure 3.9: Coefficients on the y-axis of LR FULL corresponding to the position variables on the x-axis.

Sensitivity to linear structure

As observed in Section 3.4.3, adding extra information about the position of words explains more of the variance in the case of VISUAL and especially TEXTUAL and LM. Figure 3.9 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 3.9 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.

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:

$$\mathrm{I}(A_i; C) = \sum_{a \in A_i} \sum_{c \in C} p(a, c) \log \left(\frac{p(a, c)}{p(a)p(c)} \right) \tag{3.13}$$

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 3.10 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,

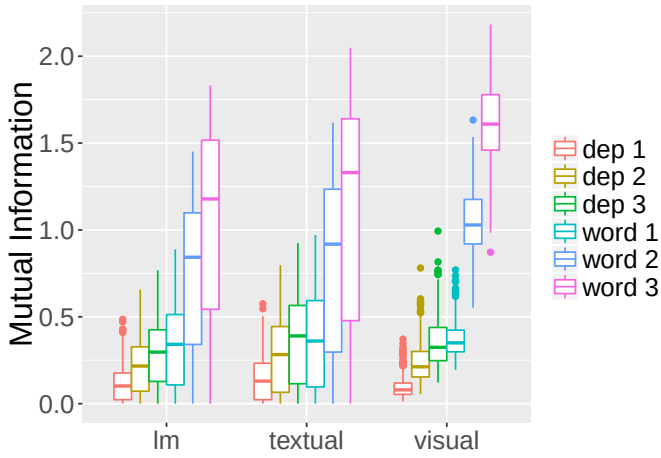


Figure 3.10: Distributions of the mutual information scores for the three networks and the six context types.

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 $\log(\text{MI}_C^T/\text{MI}_C^V)$ and $\log(\text{MI}_C^{LM}/\text{MI}_C^T)$ for all six context types C. In order to quantify variability we bootstrap this statistic with 5000 replicates. Figure 3.11 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

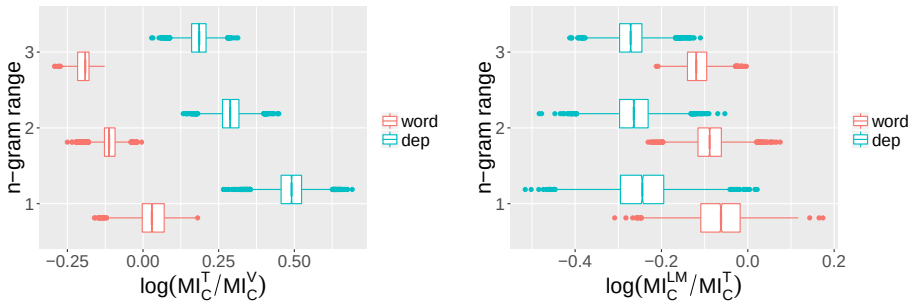


Figure 3.11: Bootstrap distributions of log ratios of median mutual information scores for word and dependency contexts. Left: TEXTUAL vs VISUAL; right: LM vs TEXTUAL

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 3.2 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.*

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 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.

Table 3.2: 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

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.

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., 2015a) 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.

4

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.

Introduction

Multimodal machine translation is the task of translating sentences in context, such as images paired with a parallel text Specia et al. (2016a). 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. (2016a). 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 4.1 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. (2015b), 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.,

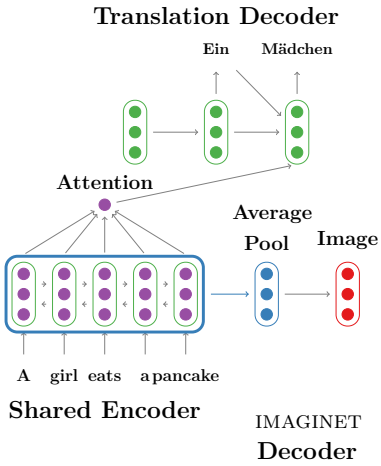


Figure 4.1: The Imagination model learns visually-grounded representations by sharing the encoder network between the Translation Decoder with image prediction in the IMAGINET Decoder.

2015b, 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. (2016) and correlate better with human similarity judgements Chrupała et al. (2015b). 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. (2013b). 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. (2016b) 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. (2015a). 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.

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. (2016a). 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:

$$J(\theta, \phi) = wJ_T(\theta, \phi^t) + (1 - w)J_G(\theta, \phi^g) \tag{4.1}$$

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.

Imagination Model

Shared Encoder

The encoder network of our model learns a representation of a sequence of N tokens $x_{1...n}$ 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 $\mathbf{E}$:

$$\mathbf{e_i} = \mathbf{x_i} \cdot \mathbf{E} \tag{4.2}$$

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:

$$\overrightarrow{\mathbf{h_i}} = \overrightarrow{\text{RNN}}(\overrightarrow{\mathbf{h_{i-1}}}, \mathbf{e_i}) \tag{4.3}$$

$$\overleftarrow{\mathbf{h_i}} = \overleftarrow{\text{RNN}}(\overleftarrow{\mathbf{h_{i-1}}}, \mathbf{e_i}) \tag{4.4}$$

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

$$\mathbf{h_i} = [\overrightarrow{\mathbf{h_i}}; \overleftarrow{\mathbf{h_i}}] \tag{4.5}$$

Neural Machine Translation Decoder

The translation model decoder is an attention-based recurrent neural network Bahdanau et al. (2015b). 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 $\mathbf{E_y}$:

$$\mathbf{e_j} = \mathbf{y_j} \cdot \mathbf{E_y} \tag{4.6}$$

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:

$$\mathbf{d_j} = \text{RNN}(\mathbf{d_{j-1}}, \mathbf{y_{j-1}}, \mathbf{e_j}) \tag{4.7}$$

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

$$\mathbf{d}_{-1} = \tanh(\mathbf{W}_{init} \cdot \frac{1}{N} \sum_{i}^{N} \mathbf{h_i}) \tag{4.8}$$

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

$$\mathbf{c_j} = \sum_{i=1}^{N} \alpha_{ji} \mathbf{h_i} \tag{4.9}$$

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:

$$\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_{l=1}^{N} \exp(e_{li})} \tag{4.10}$$

$$e_{ji} = \mathbf{v_a} \cdot \tanh(\mathbf{W_a} \cdot \mathbf{d_{j-1}} + \mathbf{U_a} \cdot \mathbf{h_i}) \tag{4.11}$$

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:

$$p(y_j | y_{<j}, c) = \text{softmax}(\tanh(\mathbf{e_{j-1}} + \mathbf{d_j} + \mathbf{c_j})) \tag{4.12}$$

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

$$J_{NLL}(\theta, \phi^t) = -\sum_{j} \log \, \mathrm{p}(y_j | y_{<j}, x) \tag{4.13}$$

Imaginet Decoder

The image prediction decoder is trained to predict the visual feature vector of the image associated with a sentence Chrupała et al. (2015b). 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 4.3.1. 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. 4.8). 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 $\mathbf{W_{vis}}$:

$$\hat{\mathbf{v}} = \tanh(\mathbf{W_{vis}} \cdot \frac{1}{N} \sum_{i}^{N} \mathbf{h_i}) \tag{4.14}$$

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); **?**. The contrastive examples $\mathbf{v}'$ are drawn from the other instances in a minibatch:

$$\begin{aligned} J_{MAR}(\theta, \phi^t) = \sum_{\mathbf{v}' \neq \mathbf{v}} \max\{0, \alpha - d(\hat{\mathbf{v}}, \mathbf{v}) \\ + d(\hat{\mathbf{v}}, \mathbf{v}')\} \end{aligned} \tag{4.15}$$

	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 4.1: The datasets used in our experiments.

Data

We evaluate our model using the benchmark Multi30K dataset Elliott et al. (2016b), 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.*

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

*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.

10,214 types for English and 16,022 for German.

We also use two external datasets to evaluate our model: the MS COCO dataset of English described images Chen et al. (2015a), 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).

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.

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 auxil-

iary 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 $\mathbf{W_{vis}} \in \mathbb{R}^{2048x2000}$ to predict the true image vector with $\alpha = 0.1$ for the Imaginet objective (Equation 4.15). The models are trained using the Adam optimiser with the default hyperparameters Kingma & Ba (2015) 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.

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. (2016a). 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, 2017) 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

	Meteor	BLEU
NMT	54.0 ± 0.6	35.5 ± 0.8
(Calixto et al., 2017a)	55.0	36.5
(Calixto & Liu, 2017)	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 4.2: 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.

source language semantics from the image and linguistic data.

Table 4.2 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.

External described image data

Recall from Section 4.2 that we are interested in scenarios where x, y, and v are drawn from different sources. We now experiment

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

Table 4.3: 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 4.4: 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.

with separating the translation data from the described image data using $\mathcal{D}_{image}$: MS COCO dataset of 83K described images[†] and $\mathcal{D}_{text}$: Multi30K parallel text.

Table 4.3 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

[†]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 4.5: Ensemble decoding results. Zmorge denotes models trained with decompounded German words; Sub-word denotes joint SentencePiece word splitting (see Section 5.4 for more details).

images.

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 4.4 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 4.5.2). 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.

Ensemble results

Table 4.5 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 pi
	NMT:	zwei kinder auf ihren gesichtern liegen unter dem boden auf d boden
	Ours:	zwei kinder liegen bäuchlings auf dem boden unter ei schaukel
	Source:	small dog in costume stands on hind legs to reach dangling fl ers
	NMT:	ein kleiner hund steht auf dem hinterbeinen und läuft , na links von blumen zu sehen
	Ours:	ein kleiner hund in einem kostüm steht auf den hinterbeine 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 4.6: 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.

Multi30K 2017 results

We also evaluate our approach against 16 submissions to the WMT Shared Task on Multimodal Translation and Multilingual Image De-

scription Elliott et al. (2017b). This shared task features a new evaluation dataset: Multi30K Test 2017 Elliott et al. (2017b), 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.

Qualitative examples

Table 4.6 shows examples of where the multitasking model improves or worsens translation performance compared to the baseline model[‡]. 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.

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

Discussion

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 4.2 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.

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. (2016a); 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 Simonyan & Zisserman (2015), 2048D features extracted from the 'pool5/7x7_s1' layer of InceptionNet V3 Szegedy et al. (2015), and

(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 4.2: We can interpret the IMAGINET Decoder by visualising the predictions made by our model.

2048D features extracted from 'avg_pool' layer of ResNet-50 He et al. (2016a). Table 4.7 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[§].

[§]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 4.7: The type of visual features predicted by the IMAGINET Decoder has a strong impact on the Multitask model performance.

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, 2017b) 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 (1997b). 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 bene-

ficial 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., 2015b) 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 improve-

ments we find from using out-of-domain parallel text and described images.

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[¶]. 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).

[¶]Code: http://github.com/elliottd/imagination

Acknowledgments

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

5

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.

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. (2014); Specia et al. (2016b). 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. (2016a).

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.

[*] 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 5.1: 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).

In real-life applications, many tasks and domains can involve code switching Barman et al. (2014), which is easier to deal with using 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 5.5) as well as adding French and Czech captioned images (Section 5.6).

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 5.1. 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 (2017a). We compare translated vs. independently collected captions in Sections 5.5.2 and 5.6.1, and overlapping vs. disjoint images in Section 5.5.3.

High-to-low resource transfer: In Section 5.6.2 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.

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 repre-

[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.

sentations have been shown to lead to higher correlation with human judgements on word-similarity benchmarks such as WordSim353 Finkelstein et al. (2001b) or SimLex999 Hill et al. (2015) compared to uni-modal representations Kádár et al. (2015); 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. (2014b); 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. (2016b), 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. (2016a) 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. (2017) 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. (2018a) 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.

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. (2014b); Karpathy & Fei-Fei (2015). The training procedure is illustrated by the pseudo-code in Figure 5.2. 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):

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

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 $< c^d, i >$ or a $< c^k_a, c^m_b >$ 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

Require: p: task switching probability.
$\mathcal{D}_{c2i}$: datasets $D_1 \dots D_k$ 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**
$T \sim \text{Bern}(p)$
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
$[\theta_\phi; \theta_\psi] \leftarrow \text{SGD}(\nabla_{[\theta_\phi;\theta_\psi]} \mathcal{J}(\mathbf{a}, \mathbf{b}))$
end while

Figure 5.2: Pseudo-code of the training procedure used to train our multilingual multi-task model.

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. (2016b) 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

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

train the **c2c** ranking model for ℓ languages is generated as follows. For a given image i, a set of languages $1 \cdots \ell$, and a set of captions $C_1^i, \ldots, C_\ell^i$ 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 $C_m^i \times C_n^i$ in $\mathcal{C}^i$ to the training set.

Experimental setup

Datasets. We train and evaluate our models on the *translation* and *comparable* portions of the Multi30K dataset Elliott et al. (2016a, 2017a). 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 5.1 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 occurring fewer than four times with a special "UNK" symbol. Table 5.1 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 5.6 we only report R@10 due to space limitations and because it has less variance than R@1 or R@5.

Bilingual Experiments

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`

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

		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 5.2: 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).

(Bilingual) and `Parallel-Sym` (Bilingual + c2c). The `OE`, `Pivot-Asym` and `Parallel-Asym` models are trained using the asymmetric similarity measure introduced for the order-embeddings Vendrov et al. (2015). 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 5.2 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 in-

		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 5.3: 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).

stead of the VGG image features). Nevertheless, our results show a similar trend to the symmetric cosine similarity models from (Gella 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 5.3 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

	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 5.4: 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.

German, whereas (Gella et al., 2017) reports the overall best results for the monolingual baseline `VSE`. Their models that use the asymmetric 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.

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 5.4 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.

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

	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 5.5: 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.

and captions in a new language. We compare these two settings here.

In the Bi-overlap condition, we collect captions for the existing 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 5.5 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 sam-

pling 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 5.5.1. 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.

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 5.5.3 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.

Translation vs. independent captions

Table 5.6 shows the results of repeating the translations vs. comparable captions experiment from Section 5.5.2 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.

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 5.7 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 points if trained on the extra 155K English and German *comparable*

	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 5.6: 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.

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.

Bilingual vs. multilingual

Finally, we investigate how useful it is to train on four languages instead of two. Figure 5.3 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 5.5.2 and 5.6.1. The Multilingual models are trained with the additional French and Czech *translation* data.

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

Table 5.7: Multilingual is trained on all *translation pairs*, + Comparable adds the *comparable* data set.

As can be seen in Figure 5.3, the performance on both tasks and for both languages improves as we move from using data from one to two to four languages.

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 improved performance for image-sentence ranking. Additionally, we

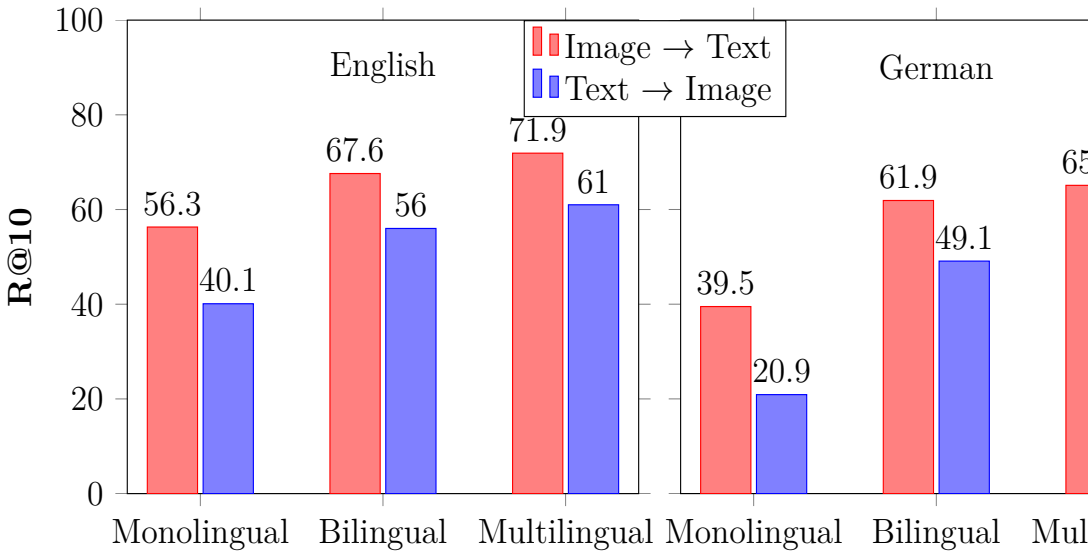


Figure 5.3: Comparing models from the Monolingual, Bilingual and Multilingual settings. The Monolingual and Bilingual models are trained on the downsampled English and German *comparable* sets with additional c2c objective. The Multilingual model uses the French and Czech *translation pairs* as additional data. The results are reported on the full 2016 test set of the *comparable* portion of Multi30K.

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.

6

General discussion and conclusion

Summary

Publication list

Journal papers

Papers in conference proceedings (peer reviewed)

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).

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.

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

Ammar, W., Mulcaire, G., Tsvetkov, Y., Lample, G., Dyer, C., & Smith, N. A. (2016b). Massively multilingual word embeddings. *arXiv preprint arXiv:1602.01925*.

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).

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. (2015a). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representation (ICLR)*.

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

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., Dinu, G., & Kruszewski, G. (2014). 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.

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.

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., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of machine learning research*, 3(Feb), 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.

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.

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).

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. *Journal of Artificial Intelligence Research*, 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., Barrault, L., & Bougares, F. (2016). Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

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. (2017). 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., & 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*.

Cao, Y., Long, M., Wang, J., Yang, Q., & Yu, P. S. (2016). Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1445–1454).: ACM.

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

Caruana, R. (1997b). 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.

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

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., & Zitnick, C. L. (2015b). Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

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

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

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).

Chrupała, G., Gelderloos, L., & Alishahi, A. (2017). Representations of language in a model of visually grounded speech signal. *arXiv preprint arXiv:1702.01991*.

Chrupała, G., Kádár, Á., & Alishahi, A. (2015a). 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).

Chrupała, G., Kádár, A., & Alishahi, A. (2015b). 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* (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. *arXiv preprint arXiv:1412.3555*.

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

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.

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

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*.

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. (2014). Long-term recurrent convolutional networks for visual recognition and description. *arXiv preprint arXiv:1411.4389*.

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.

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)*.

Elliott, D., Frank, S., Barrault, L., Bougares, F., & Specia, L. (2017a). 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., Barrault, L., Bougares, F., & Specia, L. (2017b). 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* (pp. 215–233). Copenhagen, Denmark: Association for Computational Linguistics.

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. (2016a). Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

Elliott, D., Frank, S., Sima'an, K., & Specia, L. (2016b). Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*.

Elliott, D. & Kádár, A. (2017). Imagination improves multimodal translation. *arXiv preprint arXiv:1705.04350*.

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.

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*.

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). Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 462–471).

Fazly, A., Alishahi, A., & Stevenson, S. (2010). 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, F., Wang, X., & Li, R. (2014). Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the 22nd ACM international conference on Multimedia* (pp. 7–16).: ACM.

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.

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

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2001b). 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*.

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.

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).

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. *arXiv preprint arXiv:1707.07601*.

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).

Glavaš, G., Vulić, I., & Ponzetto, S. P. (2017). If sentences could see: Investigating visual information for semantic textual similarity. In *IWCS 2017-12th International Conference on Computational Semantics-Long papers*.

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)*.

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

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

Harwath, D., Recasens, A., Surís, D., Chuang, G., Torralba, A., & Glass, J. (2018b). Jointly discovering visual objects and spoken words from raw sensory input. *arXiv preprint arXiv:1804.01452*.

Harwath, D., Torralba, A., & Glass, J. (2016). Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems* (pp. 1858–1866).

He, K., Zhang, X., Ren, S., & Sun, J. (2016a). Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778).

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

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. (2014). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *arXiv preprint arXiv:1408.3456*.

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. (2013a). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47, 853–899.

Hodosh, M., Young, P., & Hockenmaier, J. (2013b). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47, 853–899.

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).

Jabri, A., Joulin, A., & van der Maaten, L. (2016a). Revisiting visual question answering baselines. In *European conference on computer vision* (pp. 727–739).

Jabri, A., Joulin, A., & van der Maaten, L. (2016b). Revisiting visual question answering baselines. In *European conference on computer vision* (pp. 727–739).: Springer.

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*.

Jiang, Q.-Y. & Li, W.-J. (2016). Deep cross-modal hashing. *CoRR*.

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*.

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

Kádár, A., Chrupała, G., & Alishahi, A. (2016). Representation of linguistic form and function in recurrent neural networks. *arXiv preprint arXiv:1602.08952*.

Karpathy, A. & Fei-Fei, L. (2014). Deep visual-semantic alignments for generating image descriptions. *arXiv preprint arXiv:1412.2306*.

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*.

Kiela, D. (2016). Mmfeat: A toolkit for extracting multi-modal features. *Proceedings of ACL-2016 System Demonstrations*, (pp. 55–60).

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., Bulat, L., & Clark, S. (2015). Grounding semantics in olfactory perception. 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. 231–236).

Kiela, D. & Clark, S. (2015). Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 2461–2470).

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).

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. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. & Ba, J. (2015). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kiros, R., Salakhutdinov, R., & Zemel, R. (2014a). Multimodal neural language models. In *International Conference on Machine Learning* (pp. 595–603).

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014b). 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. 3294–3302).

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).

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).

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., Dinu, G., & Baroni, M. (2015a). Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1 (pp. 270–280).

Lazaridou, A., Pham, N. T., & Baroni, M. (2015b). Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Lee, J., Cho, K., & Hofmann, T. (2017). Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5, 365–378.

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., 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.

Lopopolo, A. & Miltenburg, E. (2015). Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 70–75).

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, 3(2), 273–302.

Lu, A., Wang, W., Bansal, M., Gimpel, K., & Livescu, K. (2015). Deep multilingual correlation for improved word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 250–256).

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*.

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*.

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*.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., Bernardi, R., Zamparelli, R., et al. (2014). A sick cure for the evaluation of compositional distributional semantic models. In *LREC* (pp. 216–223).

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).

McCann, B., Bradbury, J., Xiong, C., & Socher, R. (2017). Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems* (pp. 6294–6305).

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., Le, Q. V., & Sutskever, I. (2013b). Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013c). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Mikolov, T., Yih, W.-t., & Zweig, G. (2013d). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 746–751).

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11), 39–41.

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.

Nakayama, H. & Nishida, N. (2017a). Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2), 49–64.

Nakayama, H. & Nishida, N. (2017b). Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot. *Machine Translation*, 31(1-2), 49–64.

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.

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)*.

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.

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).

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. 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).

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.

Rajendran, J., Khapra, M. M., Chandar, S., & Ravindran, B. (2015). Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519*.

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.

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.

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)*.

Rubenstein, H. & Goodenough, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10), 627–633.

Ruder, S. (2017). An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

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., Berg, A. C., & Fei-Fei, L. (2014). Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*.

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).

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).

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

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.

Specia, L., Frank, S., Sima'an, K., & Elliott, D. (2016a). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation* (pp. 543–553).

Specia, L., Frank, S., Sima'an, K., & Elliott, D. (2016b). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation* (pp. 543–553). Berlin, Germany: Association for Computational Linguistics.

Srivastava, N., Mansimov, E., & Salakhudinov, R. (2015). Unsupervised learning of video representations using LSTMs. In *International Conference on Machine Learning* (pp. 843–852).

Srivastava, N. & Salakhutdinov, R. (2012). Learning representations for multimodal data with deep belief nets. In *International conference on machine learning workshop*, volume 79.

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)*.

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In *Eight International Conference on Language Resources and Evaluation (LREC'12)*.

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.

Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384–394).: Association for Computational Linguistics.

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).

Vendrov, I., Kiros, R., Fidler, S., & Urtasun, R. (2015). Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*.

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*.

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. (2014). Show and tell: A neural image caption generator. *arXiv preprint arXiv:1411.4555*.

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.

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.

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.

Xing, C., Wang, D., Liu, C., & Lin, Y. (2015). Normalized word embedding and orthogonal transform for bilingual word translation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1006–1011).

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.

Yoo, K. M., Shin, Y., & Lee, S.-g. (2017). Improving visually grounded sentence representations with self-attention. *arXiv preprint arXiv:1712.00609*.

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. (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., 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)*.

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)*.

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.