Levels of representation in a recurrent neural model of visually grounded language learning

Lieke Gelderloos
ANR 1270964

Master's Thesis
Communication and Information Sciences
Specialization Human Aspects of Information Technology
Faculty of Humanities
Tilburg University, Tilburg

Supervisor: Dr. G. Chrupała
Second Reader: Dr. A. Alishahi

July 2016

TILBURG UNIVERSITY

Abstract

Computational models of language acquisition often simplify the problem of word learning by abstracting away from the input signal children receive. Typically, they learn to match symbolic representations of words and candidate referents. In reality, both the language input and the perceptual context information that contains the candidate referents are continuous. This study presents a model of visually grounded language learning: a deep recurrent neural model that is trained to predict the features of a visual scene, based on a description in the form of a sequence of phonemes. Like children, the model has to discover structure in the phoneme stream and learn meaning from noisy and ambiguous data. A series of experiments shows that the model indeed learns to predict the visual scene, encodes aspects of both word form and visual semantic information, and is able to exploit sentence order information as well. Encoding of form and meaning shows a hierarchical organization with regards to abstraction: lower layers are comparatively more sensitive to form, whereas higher layers are more sensitive to meaning.

Declaration

This thesis describes a collaborative project between Dr. Grzegorz Chrupała and me. The implementation and training of the neural models described in chapters 2 and 3 was done by Dr. Chrupała. I prepared the phoneme transcriptions for the input data, conducted the experiments described in chapter 4, and naturally, wrote this thesis.

A substantial part of the work in this thesis has been reported in a paper that is currently under review.

Acknowledgements

First and foremost, I would like to thank my supervisor, Grzegorz Chrupała, for giving me this very exciting thesis project, and for sharing the knowledge and teaching me the skills necessary to complete it. On a more personal level, you almost managed to make me see mistakes as teachable moments instead of simple proofs of inability. Thank you for the amount of time and energy you have put in this project, for sharing my enthusiasm for it, and for introducing me to the practice of research.

Ákos Kádár, thank you for letting me pick your brain at irregular intervals, for your invaluable technical help, and for simply being interested in the project.

Cyril Cleven, thank you for pretending to be a prescriptivist for the day and organizing a crash course in phonetic transcription and [ɑː piː]. I am also grateful to you and the other corner of our triangle, Nikki Heinsman, for your moral support during the writing process, and for regularly coming up with ideas for adventures and distracting me from my thesis duties.

My parents, Roel and Margreet; this thesis would not even have been started without your moral and financial support throughout the years I have spent in university. Thank you for continually encouraging and enabling me to pursue my dreams and make the most of my talents – even though I kept changing my mind about what those are, exactly.

Florean, when most people only saw my enthusiasm about this project, you also saw how difficult or even seemingly impossible it was at times. Thank you for also seeing how much I loved working on it, and for finding that enough reason to support me every step of the way.

Contents

Introduction

A language learning child has a daunting challenge ahead. A non-exhaustive to-do list includes learning to identify words, understand what they mean, how to pronounce the phonemes of which they are composed, and how to combine them into larger utterances. Learning the meaning of words involves more than matching objects or concepts in the world to the correct word forms. The information available to human language learners does not present itself in the form of neatly separated words and pre-defined concepts. The language input consists of an uninterrupted stream of phonemes, and it is up to the child to decide which subsequences form words. The possible referents are also not a given set of concepts; they are objects and abstract concepts that exist in the outside world, which is perceived through continuous, multisensory input.

It makes intuitive sense to decompose the language learning process in a hierarchical manner, starting at the phoneme level and growing towards the comprehension and production of full conversations. Indeed, many aspects of the language learning process have been studied in isolation. Learning to segment the speech signal into words is a key problem in language acquisition research, studied experimentally with language learning children and in artificial language learning experiments with adults, as well as using computational methods. Mapping words to objects and concepts is also studied experimentally as well as computationally, specifically in the cross-situational learning paradigm. Recently, however, computational and experimental studies have shown that visual cues may aid in word segmentation learning (e.g. Thiessen (2010), Glicksohn and Cohen (2013)), indicating that learning of form and meaning should be considered in interaction, rather than in sequence. Several joint models of word segmentation and cross-situational learning have been proposed, such as the Cross-channel Early Lexical Learning (CELL) model of Roy and Pentland (2002) and the statistical model of Räsänen and Rasilo (2015). These models work with continuous speech as language input, but the visual context is pre-processed in such a way that it no longer contains the noise and ambiguity that language learners encounter.

Chrupała et al. (2015) propose Imaginet, a model of visually grounded language learning, that learns from full utterances in noisy, ambiguous visual context. Imaginet makes realistic assumptions about the visual context, but bypasses the segmentation problem by taking words as the basic unit of language input.

Phoneme GRU, the model that forms the subject of this thesis, takes the work of Chrupała et al. (2015) one step further towards the learning task of a human language learner, by taking phoneme rather than word level language input. It consists of a recurrent neural network that receives language input in the form of phonetically transcribed utterances, and learns to predict the visual scene that the utterance describes. To successfully perform the learning task, Phoneme GRU will have to acquire structural language knowledge of form as well as function, solely through visual feedback - making it a fully grounded model of language learning.

Phoneme GRU is evaluated on the visual prediction task, on which it is outperformed by a word level recurrent architecture, but outperforms a bag-of-words model. Stepping down from word to phoneme level language input thus complicates the learning task, but Phoneme GRU is able to exploit sentence order information.

The architecture of Phoneme GRU comprises multiple hidden layers. A series of experiments on processing time scales and the encoding of several types of linguistic knowledge shows hierarchical organization in the stack of hidden layers. Lower layers operate on shorter time scales, and are comparatively more sensitive to form, whereas higher layers retain information over the full length of the sentence and are more sensitive to meaning.

Chapter 1 sketches the discussions in language acquisition research on word segmentation and reference learning, and explains how Phoneme GRU fits in the discussion. It also describes technically related work in natural language processing, specifically recurrent neural architectures that work with character level language input or multimodal data. In chapter 2, the architectures of Phoneme GRU and comparison models are specified, and the corpus that was used as input is described. The ability of Phoneme GRU and comparison word level models to predict the visual scene is reported in chapter 3. Chapter 4 reports several experiments that explore the encoding of linguistic knowledge in Phoneme GRU. Finally, chapter 5 summarizes and interprets the findings, relates them to discussions in lexical acquisition research, and poses considerations for future work.

Chapter 1

Background

This thesis presents Phoneme GRU, a model of visually grounded language learning with phoneme-level language input. It is based on the Imaginet model (Chrupała et al., 2015). In contrast to other models of word meaning acquisition, in which the perceptual context is highly simplified, Imaginet learns word meaning from visual input with realistic levels of noise and ambiguity. The language input, on the other hand, consists of neat, word-sized input units. A beginning human language learner does not have access to word forms, but has to learn how to segment the continuous speech signal into meaningful units, as well as connect these units to referents. Phoneme GRU is designed to take the approach of Chrupała et al. (2015) one step further towards the learning task of a natural language learner: the language input signal consists of sequences of phonemes, rather than sequences of words.

This endeavour is motivated by open problems in language acquisition research, and is technically related to work in natural language processing (NLP) and the recently emerged cross-over field of computer vision and NLP. Section 1.1 introduces the learning problems of word segmentation and reference in lexical acquisition research, as well as some computational models that address these questions. Joint models of lexical acquisition, in which the problems of word segmentation and reference are solved concurrently, are described in section 1.1.3. Section 1.2 describes recent advances in neural NLP with character level language input, and multimodal models in which visual and language data are processed together.

1.1 Computational models of lexical acquisition

1.1.1 Word segmentation

If we were to describe the language learning process step-by-step, one of the first problems to solve would likely be deciding which parts of the speech signal constitute words. There is no acoustic equivalent of the blank space in written text; speech contains no pauses between words. Adult language users may rely on their knowledge of words and syntax to determine which segmentation of the phoneme stream is most likely. Beginning language learners, however, have yet to acquire a lexicon, which makes separating the speech signal into word-sized units especially difficult.

Language acquisition research has proposed several mechanisms that may be involved in early segmentation. Perhaps the most intuitive proposition is that children learn words from occurrences in isolation. An initial vocabulary, learnt through isolated presentation of words, may then be recognized in multi-word utterances and serve as bootstrapping material for further segmentation. Around 9.0% of maternal child-directed utterances consists of single words, and the frequency with which words occur in isolation in maternal speech is predictive of later productive use of that word, whereas the overall frequency is not (Brent and Siskind, 2001). However, many words never occur in isolation, such as function words. Furthermore, when encountering a one-word utterance, the learner does not know that it contains only one word. Therefore, relying on isolated occurrences alone is likely too shallow a basis for solving the word segmentation problem.

Although there are no pauses in between words, the speech signal may contain acoustic cues that may help to determine where word boundaries are, such as metrical patterns and allophonic variation. Allophonic variation may be helpful, because some phonemes are pronounced differently when they appear at word boundaries as opposed to in the middel of a word. Adult language users are unaware of these variations, but newborn infants do notice them (Christophe et al., 1994). Some languages have fixed or predominant word stress patterns, and adults use these cues in segmentation (Cutler and Butterfield, 1992). Indeed, the ability of 7.5 month old infants to segment words from fluent speech depends on whether they follow the predominant stress pattern (Jusczyk et al., 1999). Acoustic cues are language-specific, and therefore, so is the effectiveness of segmentation strategies based on them.

Perhaps the most consistent and language independent source of information about word boundaries lies in the sequence of phonemes itself. The phonotactic rules of a language result in some sequences of phonemes commonly occurring within words, whereas others can only occur at word boundaries. The information in the phoneme sequence itself becomes richer if we take into account statistical regularities by looking at the transitional probabilities between phonemes: the next phoneme is more predictable within a word than between two words, as each word end may be followed by any number of starts of other words, starting with any phoneme. Intuitively, the more surprised you are about one phoneme following the other, the more likely it is that this phoneme pair contains a word boundary. Behavioural studies have shown that infants as young as 8 or 7 months show sensitivity to transitional probabilities in phoneme sequences (Saffran et al., 1996a; Thiessen and Saffran, 2003). The principle has shown to be an effective learning mechanism in computational studies as well as in applications in natural language processing (Saffran et al., 1996b; Cohen and Adams, 2001; Ando and Lee, 2000; Feng et al., 2004).

Experiments such as performed by Saffran et al. (1996a) use artificial language data in which word collocations are random, but in natural language, the presence of one word has predictive value over the presence of other words. Goldwater et al. (2009)

show that learning from transitional probabilities under the assumption of independent words leads to undersegmentation when faced with natural language data. When the predictive value of words over other words is exploited, however, learning results in far more accurate segmentation.

Interesting in relation to the work presented in this thesis, though perhaps not at the heart of the segmentation discussion, is the use of recurrent neural networks (RNNs, further covered in section 1.2) in word boundary prediction. RNNs are (in theory, at least) capable of taking into account meaningful sequential information over long timescales, and may be sensitive not only to the dependencies between phonemes, but also take into account dependencies between words. In fact, the now classic paper by Elman (1990) proposes word boundary identification as one of the use cases of RNNs, through predicting the next phoneme and subsequently checking how surprising the actual phoneme was.

More recently, work in natural language processing has shown that a word boundary identification implementation based on LSTMs (more in section 1.2) reaches state of the art performance in recognizing word boundaries in Chinese written text (which has no equivalent of the blank space in Latin script) (Chen et al., 2015b). Although the symbolic representation system is of course very different from phoneme symbols or Latin script, it does indicate that the location of word boundaries is learnable from distributional information in a symbol sequence using modern recurrent neural network techniques.

1.1.2 Cross-situational learning

Learning the mapping of words to meanings is an essential part of language acquisition, as it is what makes language functional in communication. A naive example of this process might go as follows: a child hears the word *bird*, sees a robin flying by, and assumes the word *bird* refers to that robin. Learning the meaning of words in this (overly simplified) example is simply remembering which words and objects appear together.

This basic principle has been implemented in associative models of word learning such as LEX (Regier, 2005). This model receives word forms paired with a representation of meaning, and learns to associate them. Regier (2005) shows that this type of model follows several patterns observed in human word learning, such as *fast mapping* (forming and retaining an association between novel words and new referents after a single encounter), and an initial reluctance to, but eventual learning of second words for referents.

The problem with this explanation, however, is that words and referents usually do not appear alone. Words typically appear together in larger utterances, and the visual scene also contains multiple objects. In our example, the child may very well also be seeing a tree and a house, and it is quite likely that she hears the sequence *Do you see the bird?*, rather than just the word *bird* in isolation. To arrive at the correct word-meaning pairing, she must figure out that *bird* does not refer to a tree or a house, or even the action of flying or the red color of the robins' belly; and she must also find out that it is the word *bird* that refers to the robin, rather than *do* or *the.* The challenge would be even greater if in fact she did *not* see the robin - now what could the word *bird* possibly relate to? The relationship between language and referents in the outside world is obscured by noise, alignment ambiguity (not knowing which part of the utterance relates to which referent in the scene), and referential uncertainty (uncertainty about which objects in the scene, or which parts of them, are being referred to). So how do learners eventually figure out the meaning of words?

The cross-situational word learning hypothesis proposes a mechanism to deal with referential uncertainty and alignment ambiguity. According to this hypothesis, language learning infants take into account co-occurrences of words and referent candidates over multiple situations. When looking at a single instance, it may not be clear what the correct word-referent pairings are, but the uncertainty may be reduced by looking at statistical patterns in co-occurence.

The first computational model of cross-situational learning was proposed by Siskind (1996). The model consists of a set of principles that guide inference from the available evidence and partial knowledge. Whenever a new combination of words and aspects of possible referents is encountered, the sets of *possible* and *necessary* aspects of the meaning of words are updated following the inference rules. Because of the discrete nature of the rules, this model has difficulty dealing with noise. Specific mechanisms are in place to handle exceptions, but they lead to unnecessary double entries in the lexicon. Later models implement cross-situational learning as a probabilistic process, which makes it more robust to noise (e.g. Frank et al. (2007); Yu (2005)). Most models of cross-situational learning operate in a batch-like manner, processing all input data at once. However, children receive input one piece at a time, and update their knowledge incrementally. This is implemented in the model of Fazly et al. (2010).

In the cross-situational models mentioned so far, learning the meaning of words is essentially a matching task between symbolic representations of words and referents. In reality, both the visual context and the language signal are continuous rather than symbolic. A visual scene is necessarily richer and more confusing than any symbolic transcription of it can capture. The IMAGINET model by Chrupała et al. (2015) learns from continuous visual input data paired with sequences of words that describe the scene. IMAGINET consists of two recurrent neural pathways, a language model and an image scene predictor, that share the word embedding layer. The visual pathway learns to predict the visual scene based on the words in the description. The authors show that the model is able to learn the meaning of individual words through the predictive value they have over the environment.

Although IMAGINET can be seen as a cross-situational model of word learning, in fact it covers a much broader scope of the language learning process. Its recurrent architecture allows it to encode sentential structure as well. Additional studies show that both pathways selectively pay attention to certain lexical categories, and that the model may even treat the same word differently depending on the grammatical role it fills in the sentence (Kádár et al., 2016).

What IMAGINET does not take into account, however, is the continuous nature of the speech signal. IMAGINET takes word-sized units as input, and has no segmentation problem to solve. PHONEME GRU, the model described here, takes phoneme-level language as input. Informally, it can be understood as a joint model of word discovery as well as cross-situational learning.

1.1.3 Joint learning of word form and meaning

The computational models described in the previous sections focus on one aspect of language learning alone, as if it were performed in isolation from other language learning processes. Recent experimental and computational studies have found that co-occurring visual information may help to learn word forms (Thiessen (2010); Cunillera et al. (2010); Glicksohn and Cohen (2013); Yurovsky et al. (2012)). This suggests that acquisition of word form and meaning should be seen as interactive, rather than separate processes.

Räsänen and Rasilo (2015) propose an integrated view of lexical learning: rather than linguistically correct segmentation or referent assignment, they argue 'segment meaningfulness' is the primary criterion in pre-lexical speech perception. Instead of trying to segment the speech signal into units first, and subsequently trying to attach these blocks to a referent in the outside world, the infant is looking for segments that have predictive power over the environment. In essence, this is the language learning strategy implemented in the Phoneme GRU model. Phoneme GRUs task is to predict the visual context based on the phonemic signal. Knowledge of word forms and reference is acquired only because it helps to perform this learning task.

Johnson et al. (2010) propose an argument for the integrated view of lexical learning. They present several Bayesian models that learn to segment phoneme-level utterance transcriptions into words, while also learning the referents of some of the words in these utterances. The input consists of phonetically transcribed child-directed speech, with manual annotations of potential referents. The authors show that learning dependencies between words contributes to word segmentation as well as reference learning, and that word segmentation is facilitated by concurrent reference learning. The models in Johnson et al. (2010) operate on symbolic representations of language and context input, with a very limited number of possible referents, significantly reducing the noise and ambiguity faced by beginning language learners.

Several models have been proposed that use speech signal and visual input instead of symbolic representations. The first model of this sort was the Cross-channel Early Lexical Learning (CELL) model of Roy and Pentland (2002). CELL learns to discover words from continuous speech, and learn their referent. The model has speech and visual processing modules that convert the speech to phone-like units, and strips the object in the visual context of all properties but shape. Rather than segmenting the whole speech signal, the model looks for the parts of the auditory signal that co-occur with certain shapes. While CELL manages to learn words from co-occurring spoken utterances and visual objects alone, it does so by removing all referential uncertainty, noise, and ambiguity on the visual side.

Yu and Ballard (2004) present a model that integrates visual and auditive data, as well as information about eye gaze and the position of head and hands, to learn to recognize words in a grounded and meaningful way. The input data consists of speech and sensory information that is recorded while participants were performing everyday tasks and commenting on their actions. Co-occurence of speech and sensory information is used both in segmentation and grounding. The model uses remarkably rich sensory data. However, the vocabulary is limited: it covers no more than seven action-verb pairs and four object-noun pairs.

Räsänen and Rasilo (2015) propose a probabilistic, cross-situational joint model of word segmentation and meaning acquisition from raw speech and visual context. As is customary in cross-situational models of lexical acquisition, the visual context in this model consists of a set of possible referents present in the environment.

The models described above all show the benefits of integrated lexical learning, especially as a bootstrapping mechanism for early acquisition of word forms. However, the perceptual context is significantly reduced in complexity, either by converting it to a symbolic set of possible referents, or by taking only one visible object and stripping it of any attributes other than form. The model of Yu and Ballard (2004) does use rich sensory context input, but only learns a very limited set of words and referents. The visual context available to Phoneme GRU consists of high-level visual feature vectors, extracted from pictures that contain multiple possible referents. It is noisy, and the relation between the language and visual data is ambiguous.

Phoneme GRU also differs with regards to the described models in that it does not assume the sequence of phonemes has to be divided into words. The training task of Phoneme GRU is not to segment speech into words, or to attach words to objects in the visual context; the task is to predict the visual context based on the phoneme sequence. Any structural language knowledge – either about word form or reference – is acquired not because the model is explicitly looking for it, but because the recurrent neural architecture allows to capture sequential knowledge that has predictive value over the context.

Most joint models of lexical acquisition take the acoustic speech signal as input. Phoneme GRU will operate on the phoneme level instead. The main aim in this study is to investigate the possibility of acquiring meaningful language knowledge through predicting the visual context from sub-word level, sequential language data. Using phoneme-level input data allows us to analyse the acquired lexical knowledge (see chapter 4) without the need for additional annotation. This work is certainly meant as a first step towards a joint model of language learning that operates on raw speech and visual data, but the additional technical challenges that speech input data pose are left for future work.

1.2 Recurrent Neural Networks

Language data is inherently sequential in nature. Even the smallest carrier of meaning, the morpheme, is sequential in nature; place its components, the phonemes, in a different order, and you might end up with a meaningless sequence or a completely different meaning. Meaning is further determined in the way morphemes are glued together to form words, which can be composed into complex expressions by placing one after the other in a meaningful sentential order. When we take the word as the basic unit, it is nevertheless possible to infer meaningful information without looking at their relative position. Approaching a text as a 'bag-of-words', simply looking at which words it contains, rather than at their constellation, may already give a good indication of what the text is about. When taking the phoneme (or its counterpart in written text: the character) as the basic unit of input, however, the *only* way to infer meaning from the signal is to look at the sequence; a phoneme by itself carries no meaning.

Recurrent neural networks (RNNs), introduced by Elman (1990), are specifically designed to work with sequential data. Rather than processing an instance in the input in one go, RNNs receive a sequence piece by piece. Each unit in the input sequence is processed at one time step. The power of RNNs is that it has 'working memory': the activation pattern of a hidden layer at time step *t-1* is recycled as input to that same hidden layer at time step t (in addition to the new input). This property enables RNNs to learn structure in sequential data, making them especially suitable to tasks in language processing.

Although RNNs are theoretically capable of learning dependencies over long distances, in practice they face the problem of vanishing gradients. During learning, the error signal is backpropagated through time. As it gets smaller at every timestep, weight updates at distant timesteps are small in size, making learning of long-term dependencies very slow. Simple Recurrent Neural Networks therefore tend to get stuck in suboptimal solutions, that capture short-term dependencies, but disregard long-term dependencies (see Bengio et al. (1994) for a full formal discussion of the problem). Special types of units have been proposed to make learning long dependencies more efficient and effective, most notably Long Short Term Memory units (LSTM) (Hochreiter and Schmidhuber, 1997) and, more recently, Gated Recurrent Units (GRU) (Cho et al., 2014). Both LSTMs and GRUs employ a mechanism that copies part of the activation from the last step directly. This enables architectures composed of these units to capture dependencies over long time scales (Chung et al., 2014). Both LSTM and GRU based architectures are now used in a wide variety of natural language processing tasks.

1.2.1 Character-level Language Learning

Recent neural approaches to problems in Natural Language Processing (NLP) often take the word as the basic input unit. An important reason for that is the popularity and effectiveness of word embeddings in various applications. Word embeddings are essentially distributed vector representations of the words in a lexicon, that may either be learned during model training, or may be pre-trained in some other task. The Imaginet model by Chrupała et al. (2015) is based on trainable word embeddings that are shared by the visual and language pathways. As we step down from the word level to the phoneme level, we cannot use word embeddings. This may be a handicap, as word embeddings typically capture meaningful semantic similarities. However, embedding vectors are independent in principle; two morphologically related words are as different or similar as any other word pair, at least until they are adjusted in training. As Phoneme GRU looks at phonemes in sequence, it has the advantage of accessing the surface form over

Imaginet. In NLP, the equivalent of phoneme-level input is character-level input, which has recently (re)gained attention.

A model for 'composing' word representation vectors from their surface form is presented by Ling et al. (2015a). The model comprises bidirectional LSTMs (Bi-LSTMs) that read every word character-by-character, where one starts at the end of the word and the other starts at the beginning. At the last reading time step, the activation patterns of both LSTMs are combined into a word vector, which is based on, and encodes information about, the character sequence that constitutes it. There are two major benefits of character-based word vectors. Firstly, model size can be greatly reduced by using character-level input, as the lexicon no longer needs to be stored. Secondly, access to word form as well as function allows to exploit meaningful similarities in form. This improves performance in language modelling and POS-tagging, especially in morphologically rich languages (Ling et al., 2015a; Plank et al., 2016). Word representations composed by Bi-LSTMs have also been used in machine translation, allowing for character-level machine translation that is on par with word based models (Ling et al., 2015b).

Composed word embeddings show the benefits of surface form information, but these models essentially still operate on the word level - there is no segmentation problem to solve. Character-level neural NLP *without* explicit word boundaries is studied in some specific cases, typically when fixed vocabularies are problematic due to the nature of the task. For example, in Chrupała (2013), input data includes natural as well as programming language. Xie et al. (2016) present a system that automatically provides writing feedback, specifically dealing with misspelled words. Chung et al. (2016) present machine translation with character level *output*, but the *input* consists of sub-word units as in Sennrich et al. (2015), which *may* correspond to phonemes, but certainly also to words or morphemes. Importantly, these units never cross word boundaries.

Hermans and Schrauwen (2013) and Karpathy et al. (2015) describe the representation of linguistic knowledge in character-level deep recurrent neural networks. Both studies show that character level recurrent language models are sensitive to long-range dependencies. For example, they show that certain units in the network keep track of opening and closing parentheses over long stretches of text. Hermans and Schrauwen (2013) describe the hierarchical organization that seems to emerge during training, with higher layers processing information over longer timescales. Phoneme GRU is a multilayer recurrent model too, and similar techniques will be used to investigate timescale processing differences. The differences in timescale processing will also be interpreted in terms of to linguistic levels of abstraction.

Note that even in character-based approaches that do not explicitly separate words, information about word boundaries is typically present in the input data, in the form of whitespaces and punctuation (exceptions can be found in studies on non-Latin script, such as the studies by Chen et al. (2015b), Ando and Lee (2000) and Feng et al. (2004) mentioned in section 1.1.1, but the character level in these scripts does not correspond to the phoneme-level). Since it is our goal to model a multi-modal language learning process that includes the segmentation problem, the input data will not contain word boundary markers.

1.2.2 Visually grounded language learning

Interestingly, the grounding problem is now not only studied with regards to language acquisition, but also with regards to natural language processing and computer vision. While NLP and computer vision traditionally are separate disciplines, machine learning research has recently seen many cross-over studies of multimodal semantic learning. The key task in this field is automatic image captioning (see Bernardi et al. (2016) for a recent overview). The aim is to automatically understand the content of an image, and then produce a suitable description in natural language. This requires both human-like image understanding as well as natural language generation, which makes it an equally challenging task in both computer vision and NLP. The Imaginet model, on which Phoneme GRU is based, was certainly technically inspired by image captioning research, but is decidedly different as the aim is to predict the visual context from the description, instead of the other way around. Image captioning research has made large datasets of captioned images available, such as MS COCO (Lin et al., 2014; Chen et al., 2015a), which is used in the work presented here.

Inspired both by image captioning research and cross-situational human language acquisition, two recent Automatic Speech Recognition models learn to recognize word forms from visual data. In Synnaeve et al. (2014), language input consists of single spoken words and visual data consists of image fragments, which the model learns to associate. Harwath and Glass (2015) employ a convolutional visual object recognition model and another convolutional word recognition model, and an embedding alignment model that learns to map recognized words and objects into the same high-dimensional space. Although the object recognition works on the raw visual input, the speech signal is segmented into words before presenting it to the word recognition model. Synnaeve et al. (2014) do not consider the larger utterance context at all, as words are presented in isolation. In this study, we are especially interested in a language learning task that operates on full utterances.

As Harwath and Glass (2015) note, burdening a multimodal model with word segmentation on raw speech significantly complicates the task at hand. We are primarily interested in learning structure from language data that does not contain word boundaries. Whereas Harwath and Glass (2015) choose to pre-segment the speech signal, we choose to take the symbolic phoneme transcription of the caption, without word boundaries. While the language input signal in this study is symbolic rather than continuous, the work presented is certainly meant as a first exploration in the direction of learning directly from full spoken utterances in visual context.

1.3 Learning Language from Phonemes and Pictures

This thesis studies Phoneme GRU, a model that learns to predict the visual context from phoneme-level utterances. It takes the work of Chrupała et al. (2015) one step further towards the task of a language learning task, by taking phoneme-level transcriptions rather than individual words as input. The learning task of the model is not to segment speech, nor is it to connect word-sized chunks to elements of the visual scene; it is to predict the visual context. One may wonder why this is a language learning task. The idea is that when a child hears an utterance, she will update her language knowledge such that the next time she hears the sentence, she is less surprised by the visual context it appears in. If she were to hear *Do you see the bird?* while spotting a robin, she would update her language knowledge so that the next time she hears *Do you see the bird?*, she would expect to see a robin again. This idea is implemented in Phoneme GRU as prediction of the visual scene. To perform the learning task, then, Phoneme GRU will need to learn to identify visually meaningful chunks of language, and connect them to referents in the visual scene in a cross-situational manner.

Phoneme GRU will be evaluated on its ability to perform the learning task, i.e. how well it is able to predict aspects of the visual scene given a phoneme sequence, in chapter 3. The learning task of Phoneme GRU is arguably more complicated than the learning task of Imaginet, because of the additional level of linguistic knowledge that needs to be acquired. Solving the problems of reference and segmentation in an integrated fashion may however also facilitate learning. Phoneme GRU has an advantage over Imaginet in that it has access to word form, which may help to determine reference when lexically related words are known.

The emergence of hierarchical linguistic knowledge is explored through various experiments in chapter 4. Rather than making assumptions about how linguistic knowledge at several levels of granularity is stored, the aim is to have a recurrent neural model figure the levels of representation out by itself. Recent advances in visually grounded natural language processing and character-level language modeling give reason to believe a neural implementation of a phoneme-level image prediction model can be successful, if it is composed of units that are designed to deal with the problem of vanishing gradients, and comprises multiple hidden layers.

Chapter 2

Methods

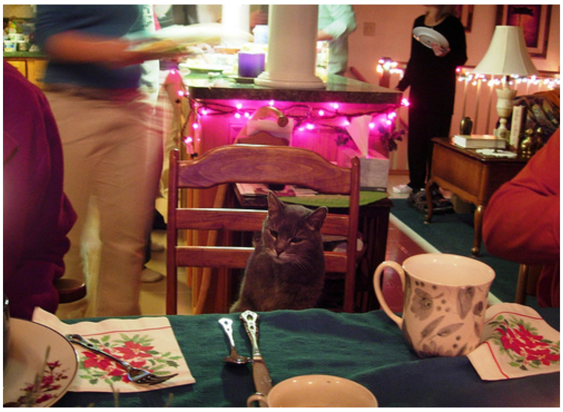
This chapter describes the architectures of the neural models analysed in chapter 4 and 5. It also describes the corpus that was used as input data, MS COCO (Lin et al., 2014; Chen et al., 2015a), and the procedures by which the pictures in this corpus were converted to high level visual feature vectors, and their orthographic captions to sequences of phonemes.

2.1 Input data

The Microsoft Common Objects in Context corpus (Lin et al., 2014) is an image dataset that is widely used in object recognition, image segmentation, and image captioning research. It contains over 328,000 images that are annotated for the presence of objects. A large portion of this dataset (over 163,000 images) is accompanied by at least five captions by human annotators describing the visual scene, collected through crowdsourcing (Chen et al., 2015a). This collection of captioned images (hencefort referred to as MS COCO) is suitable to our purposes, both because of the content of the images, and the nature of their descriptions. The selection of images in the Microsoft Common Objects in Context corpus is such that they contain (at least) one of 91 object types that would be easily recognizable by a four year old. Importantly, these objects are *in context*: the images depict naturally occurring scenes containing multiple objects (an average of 7.7 labelled object instances per image, to give an impression). The visual data thus contains multiple possible referents, making reference to the scene ambiguous. For the captions, annotators were instructed to use at least 8 words and describe all major parts of the scene, but not the unimportant details.

Figure 2.1 shows an example image with captions. The image contains more objects and persons than any of the captions mention. The captions also contain parts that can not directly be mapped to objects in the scene, but show additional word knowledge or interpretation. For example, the first caption given mentions *christmas dinner*, while the dinner itself is not visible in the scene, and even gives an interpretation of the thoughts of the depicted cat.

All models were trained on the full standard training set of MS COCO. For validation



Captions and transcriptions
the family cat wants to join in during christmas dinner ðəfamɪlikatwɒntstədʒɔɪnɪndjʊəɹɪŋkɹɪsməsdɪnə
a cat sitting on a chair at the table with a place setting ɐkatsɪtɪŋɒnɐtʃeəɹatðəteɪbəlwɪðɐpleɪssɛtɪŋ
a gray cat sits in a chair next to a table that has a blue table cloth on it and is set with silverware and cups. ɐɠđɹeɪkatsɪtsɪnɐtʃeənɛksttʊɐteɪbəlðathɐzɐbluːteɪbəlklɒθɒnɪtandɪzsɛtwɪðsɪlvəweəandkʌps
a cat sitting at a table filled with silverware ang plates əkatsɪtɪŋatəteɪbəlfɪldwɪðsɪlvəweəɹaŋpleɪts
a cat sits on a chair at the dinner table. ɐkatsɪtsɒnɐtʃeəɹatðədɪnəteɪbəl

Figure 2.1: A picture from MS COCO with captions and phoneme transcriptions. Phonemes are not separated in this example, but they are presented one-by-one to Phoneme GRU (even if they are represented by two characters in the transcription).

and testing, two random samples of 5,000 images (and the corresponding captions) were selected from the MS COCO validation set.

2.1.1 Phoneme sequences

The captions in MS COCO are in conventional English spelling (although they do contain the occasional typo or misspelling). They were converted to phonetic representations using eSpeak[1], using the IPA-output option with separated phonemes, and the default English voicing. Pauses and word and sentence stress markers were cleaned out. The extra elongated *iː:* was converted to *iː* (a transcription artefact caused by the apparently quite frequent presence of the 'Wii' console in MS COCO). An end-of-utterance symbol was added. Word boundaries were removed from the phoneme stream before presenting it to the model. They were, however, preserved for use in the word boundary prediction

[1] Available at `http://espeak.sourceforge.net/`

experiment described in section 4.3. The phoneme transcriptions are available at `https://github.com/lgelderloos/levels_representation/tree/master/data`. Figure 2.1 contains phoneme transcriptions as they are presented to the model.

2.1.2 Visual feature vectors

Rather than using raw images, the prediction objective of the multimodal models consisted of fixed image feature vectors that encode high-level features of the image. All images were fed to the 16-layer convolutional object recognition network by Simonyan and Zisserman (2014) which was pretained on the ImageNet dataset (Russakovsky et al., 2015). The 4096-dimensional activation vector of the pre-final layer is used as target visual feature vector. The visual input to any of the multimodal models thus did not consist of the raw image, but rather of an activation vector that contains information about high-level visual features of the object types that are present in the scene. This assumes object recognition processing on the visual information. We believe this is a reasonable assumption, given the capacities of very young children to form visual categories (Mareschal and Quinn, 2001).

2.2 Models

The main topic of interest in this thesis is the Phoneme GRU model, but three other models are described for comparison: the multimodal language learning models Word GRU and Word Sum and the purely linguistic Language Model. The architecture of all models is summarized in table 2.1.

Phoneme GRU is the model that learns to predict the visual scene from phoneme The architecture of Phoneme GRU is based on the visual pathway of the Imaginet architecture by Chrupała et al. (2015), but takes sequences of phonemes as input, rather than sequences of words. Phoneme GRU consists of a phoneme encoding layer, one or more hidden layers, and mapping from the final state of the top hidden layer to the image feature vector. As prior research has shown that a deep recurrent architect allows for hierarchical processing of character-level data (Hermans and Schrauwen, 2013; Karpathy et al., 2015), several versions of Phoneme GRU were explored, with varying numbers of hidden layers. Comparison and selection is described in section 3.1.

The Language Model is similar in architecture to Phoneme GRU, but is trained to predict the next phoneme in sequence rather than the visual context at the end-of-sentence symbol.

Word GRU and Word Sum are defined and trained in order to compare the performance of Phoneme GRU to word-level models. Performance of all multimodal models on the training task is reported in chapter 3. Word GRU is a word-level, single GRU-layer version of Phoneme GRU, taking words rather than character sequences as input, and thus is similar to the visual pathway of Imaginet (Chrupała et al., 2015). It consists of a word embedding layer, a single hidden layer, and a mapping from the final state of the hidden layer to the image feature vector. Word Sum is simply a sum of the embeddings of all words in a sentence, which is then mapped to the image feature vector. It contains no word order information, and can thus be considered a continuous bag-of-words model. WORD SUM and WORD GRU should *not* be considered baseline models for the visual feature prediction task. PHONEME GRU may have the advantage of access to surface form, but WORD GRU and WORD SUM certainly also exploit information that is unavailable to PHONEME GRU: the location of word boundaries and the identity of words is unambiguous to them.

As their names indicate, the hidden layers on PHONEME GRU, WORD GRU and the Phoneme GRU LANGUAGE MODEL are built from Gated Recurrent Units (GRU), that were proposed by Cho et al. (2014) as a solution to the vanishing gradient problem. GRU is a simpler architecture than the earlier proposed LSTMs Hochreiter and Schmidhuber (1997), but performs similarly on sequential tasks involving long-range dependencies (Chung et al., 2014).

A GRU computes the hidden state at the current time step, $\mathbf{h}_t$, as the linear combination of previous activation $\mathbf{h_{t-1}}$, and a new *candidate* activation $\tilde{\mathbf{h}}_t$:

$$\text{gru}(\mathbf{x}_t, \mathbf{h}_{t-1}) = (1 - \mathbf{z}_t) \odot \mathbf{h}_{t-1} + \mathbf{z}_t \odot \tilde{\mathbf{h}}_t \tag{2.1}$$

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

$$\mathbf{z}_t = \sigma_s(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1}) \tag{2.2}$$

The candidate activation is computed as:

$$\tilde{\mathbf{h}}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1})) \tag{2.3}$$

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:

$$\mathbf{r}_t = \sigma_s(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1}) \tag{2.4}$$

By applying the gru function repeatedly, a GRU layer maps a sequence of inputs to a sequence of states.

$$\text{GRU}(\mathbf{X}, \mathbf{h}_0) = \text{gru}(\mathbf{x}_n, \ldots, \text{gru}(\mathbf{x}_2, \text{gru}(\mathbf{x}_1, \mathbf{h}_0))) \tag{2.5}$$

where $\mathbf{X}$ stands for the matrix composed of input column vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Two or more GRU layers can be composed into a stack:

$$\text{GRU}_2(\text{GRU}_1(\mathbf{X}, \mathbf{h_1}_0), \mathbf{h_2}_0). \tag{2.6}$$

The multi-layer PHONEME GRU and LANGUAGE MODEL architectures are composed of *residualized* layers. Residualization means adding the input vectors to the GRU function to obtain the activation. Residual connections were introduced for use with convolutional networks by (He et al., 2015) and applied in recurrent architectures by (Oord et al., 2016). In exploratory stages of this project, it was observed that residual connections speed up learning in stacks of several GRU layers, which is why they are adopted here.

$$\mathrm{GRU}_{\mathrm{res}}(\mathbf{X}, \mathbf{h}_0) = \mathrm{GRU}(\mathbf{X}, \mathbf{h}_0) + \mathbf{X} \tag{2.7}$$

The gated recurrent units use steep sigmoids for the gate activations:

$$\sigma_s(z) = \frac{1}{1 + \exp(-3.75z)}$$

and rectified linear units clipped between 0 and 5 for unit activations:

$$\sigma(z) = \mathrm{clip}(0.5(\mathrm{z} + \mathrm{abs(z)}), 0, 5)$$

The phoneme encoding layer of Phoneme GRU is a lookup table **E** whose columns correspond to one-hot-encoded phoneme vectors. The input phoneme p_t of utterance p at each step t indexes into the encoding matrix and produces the input column vector:

$$\mathbf{x}_t = \mathbf{E}[:, p_t]. \tag{2.8}$$

Mapping of the final state of the (top) GRU layer $\mathbf{h}_{\mathbf{K}n}$ for the recurrent models, and the summed embedding vector for Word Sum, is mapped to the vector of image features using a fully connected layer:

$$\hat{\mathbf{i}} = \mathbf{I}\mathbf{h}_{\mathbf{K}n} \tag{2.9}$$

All models were implemented in Theano (Bergstra et al., 2010) and optimized with Adam (Kingma and Ba, 2014). Training was done in minibatches of 64 captions. The initial learning rate was set to 0.0002.

Model	**Input**	**Hidden**	**Prediction made at**	**Prediction**
Phoneme GRU	one-hot encoding of phoneme	3 1024-dimensional GRU layers*	end of sentence	4096-dimensional image feature vector
Word GRU	1024-dimensional word embedding	1 1024-dimensional GRU layer	end of sentence	4096-dimensional image feature vector
Word Sum	1024-dimensional word embedding	None (word embeddings are summed)	end of sentence	4096-dimensional image feature vector
Language model	one-hot encoding of phoneme	3 1024-dimensional GRU layers	every time step	phoneme at $t + 1$

* This is the final architecture of Phoneme GRU, but several numbers of GRU layers have been explored. See section 3.1 for details.

Table 2.1: Overview of models

Chapter 3

Prediction of visual features

During training of all multimodal models, the loss function was the cosine distance between the predicted and actual visual feature vector. Evaluation of the models was done using an image retrieval task: all images in the evaluation set were ranked on cosine distance to the predicted feature vector. The evaluation score is *accuracy @ 5*: the proportion of times the correct image was amongst the five images that most closely corresponded to the predicted feature vector. Note that *correct* here refers to the image that originally elicited the caption. In reality, a caption may of course also be applicable to images other than the one it originally described.

3.1 Model selection for Phoneme GRU

Five different versions of PHONEME GRU were implemented, with one to five GRU layers in the stack. To determine the optimal number of GRU layers, all variants were evaluated on the image retrieval task. Figure 3.1 shows accuracy @ 5 on the validation data, for three initializations of all versions of PHONEME GRU. Clearly, the optimal number of GRU-layers for PHONEME GRU was three. The training and evaluation reported the rest of this chapter, as well as the experiments reported in chapter 4, were performed on the three hidden layer-version of PHONEME GRU. Subsequently, PHONEME GRU refers to this version.

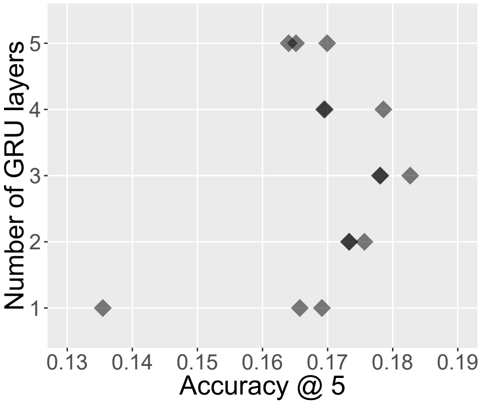


Figure 3.1: Validation accuracy @ 5 for different versions of PHONEME GRU. Reported scores are for best performing epochs of random three initializations.

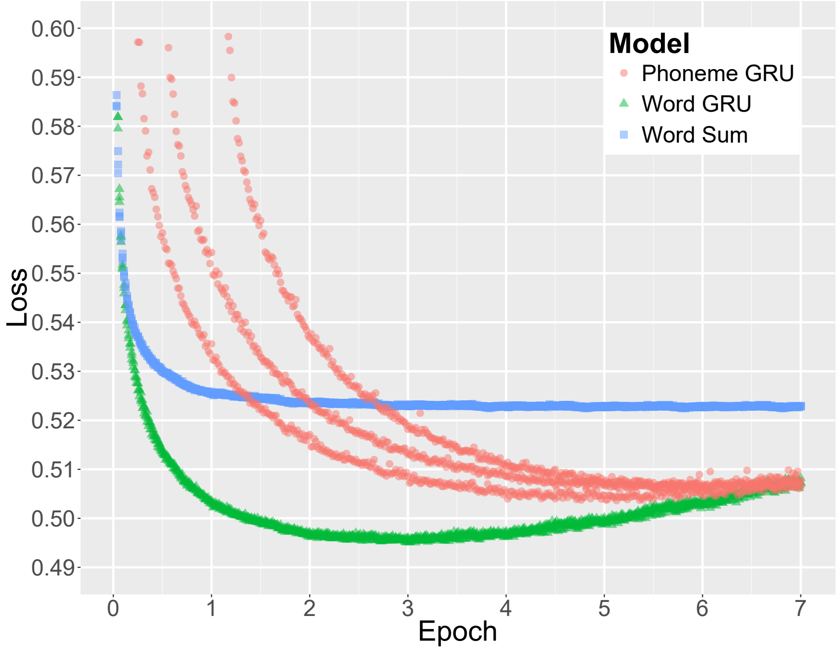


Figure 3.2: Value of the loss function on validation data during training. Three random initialization of each model are shown.

3.2 Training

Figure 3.2 shows the value of the validation average cosine distance between the predicted visual vector and the target vector, for three random initializations of each of the model types. For Word GRU and Word Sum, the random initialization does not make much of a difference: the loss function trajectories can not be discerned. For Phoneme GRU, however, the initializations lead to clearly separate trajectories. For the word level models, the loss value diminishes much quicker than for Phoneme GRU. After enough training, all initializations of Phoneme GRU outperform Word Sum, but not Word GRU.

3.3 Evaluation

After each epoch in training, the models were evaluated on the image retrieval task on the validation data. The plotted scores in figure 3.3 correspond to the optimal epoch for each of the models. On the validation data, accuracy @ 5 for Phoneme GRU falls in between the scores for Word Sum and Word GRU. This is consistent with the value of the loss function reported in figure 3.2. The best performing versions of Word GRU, Word Sum and the three-hidden-layer Phoneme GRU were selected for use in the experiments reported in chapter 4, and evaluated on the unseen test data.

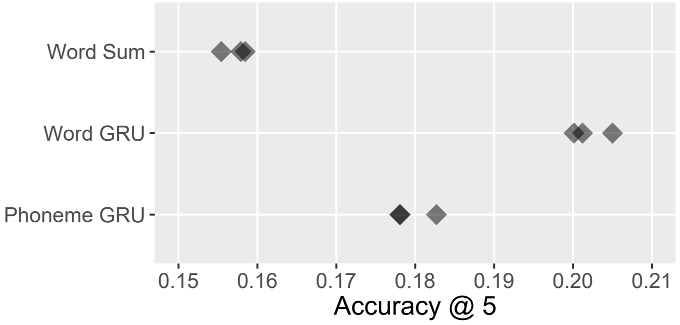


Figure 3.3: Validation accuracy @ 5 on the image retrieval task. Three different initializations of each model; reported scores are for best performing epochs.

The image retrieval test scores for the optimal models are reported in table 3.1. The accuracy @ 5 for WORD GRU is comparable to what Chrupała et al. (2015) report for IMAGINET, whose visual pathway has the same structure. Evaluation on the test data shows the same pattern as observed during validation: performance of PHONEME GRU is better than that of WORD SUM, but less than the performance of WORD GRU.

Model	**Accuracy @ 5**
WORD SUM	0.158
WORD GRU	0.205
PHONEME GRU	0.180

Table 3.1: Image retrieval accuracy @ 5 on test data for the versions of WORD SUM, WORD GRU and PHONEME GRU chosen by validation.

Chapter 4

Representation of linguistic knowledge

This chapter reports several experiments that were performed to investigate the levels of representation of linguistic information in PHONEME GRU. The time scale at which its hidden layers operate is explored in section 4.1 by investigating for how long a permutation at the phoneme level is detectable in the activation vectors, and in 4.2 by showing the position of shared sub-sequences in sentences with similar distributed representations. The encoding of sub-word linguistic information is studied by means of a word boundary detection experiment in 4.3. Finally, section 4.4 describes a word activation vector similarity experiment that studies the encoding of phonetic form and semantic features. The code for all experiments described in this chapter is available at `https://github.com/lgelderloos/levels_representation/tree/master/experiments`.

4.1 Phoneme replacement

This experiment explores the time scale at which the three hidden layers in PHONEME GRU operate. It is based on an exploratory analysis of the differential processing time scales of layers in deep recurrent character-level language models by Hermans and Schrauwen (2013). One item in a sequence is replaced by another item that is present in the data, and the difference this makes to the activation vector at each layer is tracked through time. Hermans and Schrauwen (2013) find substantial differences between low layers, in which the change in activation disappears over several tens of time steps, and high layers, in which it remains clearly detectable after a hundred time steps.

In our adaptation of this experiment, the first phoneme in a sentence is replaced by another phoneme. Both the original phoneme sequence and the perturbed version are fed to PHONEME GRU. For each hidden layer, the cosine distance between the activation vectors for the original and the perturbed sequence is recorded at all time steps. The development of this distance through time, and how this differs for the three hidden layers, can give us a first indication of differential processing time scales in PHONEME GRU.

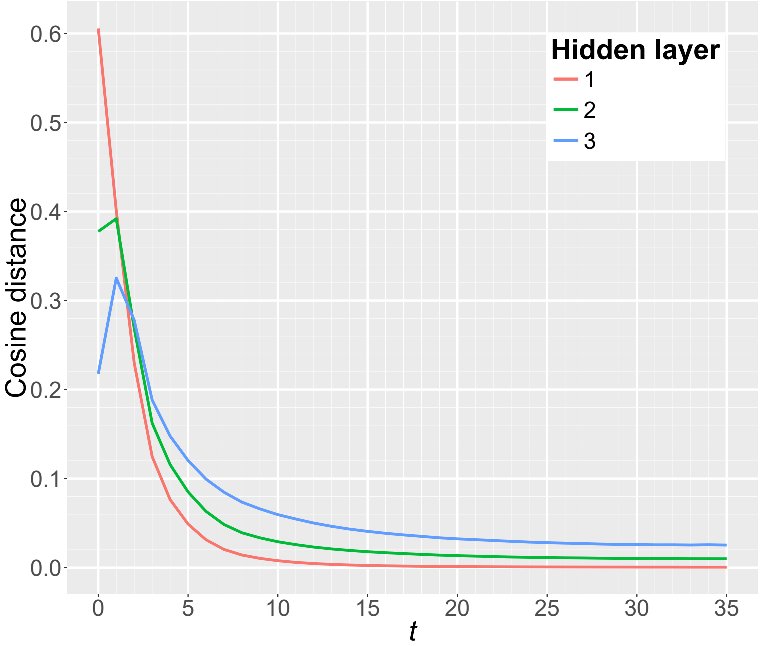


Figure 4.1: Average cosine distance between activation vector at t time steps after replaced phoneme. Graph is cut off at $t = 35$; average sentence length in validation set is 34.4 phonemes, mode is 31.

Figure 4.1 shows how replacing the first phoneme in a sequence affects the activation vectors through time. The plotted cosine distances are averages over the captions from 5,000 images in the validation set of MS COCO that were also used as validation data in chapter 3. In the first layer, the effect of changing one phoneme quickly diminishes, and is close to zero around $t = 15$. In the second and third layer, some effect of the perturbation remains present for the duration of the sentence (the plot is cut off at $t = 35$, just over the average sentence length), but this difference is considerably larger in layer 3 than it is in layer 2. In layer 2 and 3, the activation vector distance at $t = 1$ is larger than at the time step of the replaced phoneme itself. This first exploration indicates that the layers in Phoneme GRU differ in the way they retain information over time. In the following experiments, we explore where these temporal differences come from.

4.2 Shared subsequences in nearest-neighbour sentences

In the previous experiment we made a single alteration to the linguistic input, and explored for how long the change in the activity lingered at the different layers of Phoneme GRU. The experiment in this paragraph also investigates time scale processing differences between the levels in the stack of GRUs, but it takes the opposite approach: it considers sentences that show similar activation patterns in Phoneme GRU, and investigate what parts of their phoneme sequences match. The outcome measure is the average position of matching substrings in the original sentence. This gives an indication of the time scale the layer operates on: if the shared properties of nearest neighbours are all at the end of the sentence, that the activation pattern mostly depends on information from close time steps. However, if similar end-of-sentence representations are generated by phoneme sequences that only share substrings in the beginning of the sentence, information must be retained over the full length of the sentence. While the position of shared substrings is arguably a less direct time scale indicator than the change in activation that was analysed in the last experiment, this experiment reflects the behaviour of PHONEME GRU with natural utterances, which typically differ on the word level, rather than the level of the phoneme.

The captions of the 5,000 validation images for visual feature prediction were included in this experiment. For each sentence in this set, the nearest neighbour was determined for each hidden layer in PHONEME GRU. The nearest neighbour is the sentence in the validation set for which the activation vector at the end of sentence symbol has the smallest cosine distance to the activation vector of the original sentence. Average substring position is quantified as the average position in the original sentence of all phonemes in substrings that it shares with its nearest neighbour, counted from the end of the sentence. A high mean average substring position thus means that the shared substrings appear early in the sentence, and that information is retained for longer periods of time. See figure 4.2 for an illustration of this idea.

Layer	**Mean position**
1	12.1
2	14.9
3	16.8

Table 4.1: Average position of symbols in shared substrings between nearest neighbour sentences according to PHONEME GRU activation patterns. Positions are indexed from end of sentence, i.e. index 0 is the last symbol.

As can be seen in Table 4.1, the average position of shared substrings in neighbour sentences is closest to the end for the first hidden layer and moves towards the beginning of the sentence for the second and third hidden layer. These findings corroborate the findings in the last experiment: information from distant time steps is more contributive to the activation pattern of the highest layer than to that of the lowest layer.

4.3 Word boundary prediction

To explore the sensitivity of PHONEME GRU to linguistic structure at the sub-word level, we investigated the encoding of information about word ends in the hidden layers. A logistic regression model was trained on activation patterns of the hidden layers at all timesteps, with the objective of identifying phonemes that preceded a word boundary. For comparison, several logistic regression models were trained to perform the same task on activation vectors from the PHONEME GRU LANGUAGE MODEL, and on positional n-gram data.

The location of the word boundaries was taken from the eSpeak transcriptions. Mostly, these match the location of word boundaries in conventional English spelling.

Layer 1	A horse and rider jumping over a bar **on a track**.
	A train carrying carts coming around a curve **on a track**.
	ɐhɔːsandɹaɪdədʒʌmpɪŋəʊvəɹɐbɑːɹ**ɒnɐtɹak**
	ɐtɹeɪnkaɹɪɪŋkɑːtskʌmɪŋɐɹaʊndɐkɜːv**ɒnɐtɹak**
Layer 2	A **horse** and rider jumping over a bar **on a track**.
	Two **horse-**drawn chariots racing **on a track**.
	ɐ**hɔːs**andɹaɪdədʒʌmpɪŋəʊvəɹɐbɑːɹ**ɒnɐtɹak**
	tuː**hɔːs**dɹɔːntʃaɹiətsɹeɪsɪŋ**ɒnɐtɹak**
Layer 3	A **horse and rider jumping over** a bar on a track.
	The **horse and rider** are **jumping over** the yellow structure.
	ɐ**hɔːsandɹaɪdədʒʌmpɪŋəʊvə**ɹɐbɑːɹɒnɐtɹak
	ð**əhɔːsandɹaɪdə**ɹɑː**dʒʌmpɪŋəʊvə**ðəjɛləʊstɹʌktʃə

Table 4.2: Example of a phoneme sequence with nearest neighbours for each GRU-layer in Phoneme GRU. Conventional spelling given for convenience. Shared subsequences of 3 or more phonemes are printed in bold. In reality, shared subsequences of all lengths were taken into account.

However, eSpeak models some coarticulation effects, which sometimes leads to word boundaries disappearing from the transcription. For example, *bank of a river* is transcribed as [baŋk əvə ɹɪvə]. The features for the positional n-grams are all 1- to n-grams between t-n and t (where t is the current phoneme), marked for the position relative to t.

All models were implemented using the `LogisticRegression` implementation from Scikit-learn (Pedregosa et al., 2011) with L2-normalization. The captions of the visual feature prediction validation data were used as training data, and those of the test set as test data. For all models, the optimal value of regularization parameter C was determined using `GridSearchCV` with 5-fold cross validation on the training set. Models with C at the optimal value were then trained on the full training set.

Table 4.3 reports recall, precision, accuracy and F_1 score on the test set. The proportion of phonemes preceding a word boundary is 0.29, meaning that predicting *no word boundary* by default would be correct in 0.71 of cases. At the highest hidden layer of Phoneme GRU, enough information about the word form is available for correct prediction in 0.82 of cases - substantially above the majority baseline. The lower levels allow for more accurate prediction of word boundaries: 0.86 at the middle hidden layer, and 0.88 at the bottom level. Prediction accuracy of the logistic regression model based on the activation patterns of the lowest hidden layer is comparable to that of a bigram logistic regression model.

Figure 4.4 illustrates the wourd boundary predictions (and errors) that are made by each Phoneme GRU-layer based regression model.

Activation vectors of the Phoneme GRU Language model at any level provide enough information for accurate prediction of word boundaries in over 0.94 of cases. Although Phoneme GRU shares the architecture of this model up to the prediction

Model		**Accuracy**	**Precision**	**Recall**	$\mathbf{F}_1$
Majority baseline		0.71			
Phoneme GRU	Layer 1	0.88	0.82	0.78	0.80
	Layer 2	0.86	0.79	0.71	0.75
	Layer 3	0.82	0.74	0.60	0.66
Language Model	Layer 1	0.94	0.90	0.92	0.91
	Layer 2	0.96	0.92	0.94	0.93
	Layer 3	0.96	0.92	0.94	0.93
n-gram	$n = 1$	0.80	0.79	0.41	0.54
	$n = 2$	0.87	0.79	0.78	0.79
	$n = 3$	0.93	0.86	0.90	0.88
	$n = 4$	0.95	0.90	0.93	0.92

Table 4.3: Word boundary prediction scores of linear regression models based on activation vectors of Phoneme GRU, the Language Model, and positional n-grams

layer, its activation vectors contain considerably less information that allow for accurate and precise word boundary prediction. There is a distinct difference with regards to recall in particular, which in the Phoneme GRU-based classifiers is always smaller than precision, whereas it is in fact larger than precision for the Language Model-based classifiers.

These results indicate that information on sub-word structure is only partially encoded by Phoneme GRU. It is mostly absent by the time the signal from the input propagates to the top layer, at which point only 0.60 of word ends are correctly identified as such. Although the bottom layer does learn to encode a fair amount of word boundary information, the recall of 0.78 indicates that it is rather selective.

Original	ɐ smɔːl blak dɒg standɪŋ əʊvəɹ ɐ pleɪt ɒv fuːd	A small black dog standing over a plate of food
Layer 1	ɐ smɔːl blak dɒg stand ɪŋ əʊvə ɹɐ pleɪt ɒv fuːd	A small black dog stand ing ove ra plate of food
Layer 2	ɐ smɔːl blak d ɒg s tand ɪŋ əʊvə ɹɐ pleɪt ɒv fuːd	A small black d og s tand ing ove ra plate of food
Layer 3	ɐ smɔːlblak d ɒg st and ɪŋ əʊvəɹɐpleɪtɒv fuːd	A smallblack d og st and ing overaplateof food
Original	tuː zɛbɹə havɪŋ ɐ faɪt ɒn tɒp əvə dɹaɪ gɹas fiːld	Two zebra having a fight on top ofa dry grass field
Layer 1	tuː z ɛbɹəh avɪŋ ɐ faɪtɒn t ɒp əvə dɹaɪgɹasfiːld	Two z ebrah aving a fighton t op ofa drygrassfield
Layer 2	tuː z ɛbɹəhavɪŋ ɐ faɪtɒn tɒp əvə dɹaɪgɹasfiːld	Two z ebrahaving a fighton top ofa drygrassfield
Layer 3	tuː z ɛbɹəhavɪŋ ɐfaɪtɒntɒpəvədɹaɪgɹasfiːld*	Two z ebrahaving afightontopofadrygrassfield*

* in this case, the word boundary at the end of the sentence was missed, too.

Table 4.4: Two sentences segmented according to word boundary predictions by classifiers based on activation patterns in Phoneme GRU. Conventional spelling equivalents provided for convenience.

		All words	Frequent words
PHONEME GRU	Layer 1	0.09	0.12
	Layer 2	0.21	0.33
	Layer 3	0.28	0.45
WORD GRU		0.48	0.60
WORD SUM		0.42	0.56

Table 4.5: Spearman's correlation coefficient between word-word cosine similarity and human similarity judgements. Frequent words appear at least 100 times in the training data. All correlations significant at $p < 1\text{e}{-4}$.

4.4 Word similarity

The previous experiments have indicated that the layers in the stack of GRUs in PHONEME GRU operate on different time scales, and that information about word form is mostly encoded in the lowest layer, and dissipates as the input signal propagates through the network. This experiment studies where information about word form and meaning is represented in PHONEME GRU. The correlation of cosine similarity between word pairs with human similarity ratings is investigated, as well as their correlation with surface form similarity.

To understand the encoding of semantic information in PHONEME GRU, the cosine similarity of activation vectors for word pairs from the MEN dataset Bruni et al. (2014) were compared to human similarity judgements. The MEN dataset contains 3,000 word pairs that are rated for semantic relatedness. For each word pair in the MEN dataset, the words were transcribed phonetically using eSpeak and then fed to PHONEME GRU individually. For comparison, the words were also fed to WORD GRU and WORD SUM. Word pair similarity was quantified as the cosine similarity between the activation patterns of the hidden layers at the end-of-sentence symbol.

Table 4.5 shows Spearman's rank correlation coefficient between human similarity ratings from the MEN dataset and cosine similarity at the last timestep for all hidden layers. In all layers, the cosine similarities between the activation vectors for two words are significantly correlated with human similarity judgements. The strength of the correlation differs considerably between the layers, ranging from 0.09 in the first layer to 0.28 in the highest hidden layer. Correlations for both WORD GRU and WORD SUM are considerably higher than for PHONEME GRU. This is expected given that these are word level models with explicit word embeddings, while PHONEME GRU builds word representations by forwarding phoneme-level input through several layers of processing.

The second column in Table 4.5 shows the correlations when only taking into account the 1,283 word pairs of which both words appear at least 100 times in the training data. For all cosine similarities, but most distinctively for those of the third layer of PHONEME GRU, the correlations with human similarity ratings are considerably higher when only taking into account well known words.

The relative contribution of surface form information to activation pattern in PHONEME

GRU is analysed through the correlation of cosine similarity with a measure of phonetic difference: the Levenshtein distance between the phonetic transcriptions of the two words, normalized by the length of the longer transcription.

Table 4.6 shows Spearman's rank correlation coefficient between the normalized edit distance of word pairs and the cosine similarity of activation vectors at the hidden layers of PHONEME GRU. As expected, edit distance and cosine similarity of the activation vectors are negatively correlated, which means that words which are more similar in form also have more similar representations in PHONEME GRU. (Note that in the MEN dataset, meaning and word form are also (weakly) correlated: human similarity judgements and edit distance are correlated at -0.08 ($p < 1\text{e}{-5}$).) The negative correlation between edit distances and cosine similarities is strongest at the lowest hidden layer and weakest, though still present and stronger than for human judgements, at the third hidden layer.

Layer	ρ
1	-0.30
2	-0.24
3	-0.15

Table 4.6: Spearman's rank correlation coefficient between PHONEME GRU cosine similarity and phoneme-level edit distance. All correlations significant at $p < 1\text{e}{-15}$.

Table 4.7 shows the 20 most similar pairs of well-known words in the MEN dataset for each layer in PHONEME GRU, as determined by calculating the cosine distance at the end of sentence symbol. Only pairs of words that occur at least 100 times in the training data are included in this table. In the first layer, all most-similar word pairs are similar in phonetic form, and most word pairs are related in meaning. A couple of word pairs, however, are similar in form, but not in meaning, such as "licking - sitting". In the second layer, although almost all word pairs are still similar in form, all word pairs are *also* similar in meaning, which is reflected in the high semantic similarity ratings. At layer 3, several word pairs are similar in meaning, yet do not share substrings (larger than one phoneme), such as "sandwich - burger" and "city - town".

The correlations of cosine similarities with edit distance on the one hand, and human similarity rating on the other hand, indicate that the different hidden layers in PHONEME GRU reflect increasingly abstract levels of representation. Qualitative analysis of the most similar word pairs shows the same pattern: whereas information about phonetic form is gradually lost as the input signal propagates up through PHONEME GRU, semantic information becomes more prominent.

Layer 1					**Layer 2**					**Layer 3**				
word pair		IPA		MEN	word pair		IPA		MEN	word pair		IPA		MEN
airplane	plane	eəpleɪn	pleɪn	42	airplane	plane	eəpleɪn	pleɪn	42	airplane	plane	eəpleɪn	pleɪn	42
light	sunlight	laɪt	sʌnlaɪt	44	dirt	dirty	dɜːt	dɜːti	45	painted	painting	peɪntɪd	peɪntɪŋ	42
bedroom	room	bɛdɹuːm	ɹuːm	41	sun	sunny	sʌn	sʌni	44	sun	sunny	sʌn	sʌni	44
licking	*sitting*	*lɪkɪŋ*	*sɪtɪŋ*	*14*	painted	painting	peɪntɪd	peɪntɪŋ	42	dirt	dirty	dɜːt	dɜːti	45
eat	meat	iːt	miːt	35	skate	skating	skeɪt	skeɪtɪŋ	45	weather	wet	wɛðə	wɛt	33
dirt	dirty	dɜːt	dɜːti	45	weather	wet	wɛðə	wɛt	33	bicycle	bike	baɪsɪkəl	baɪk	45
flight	flying	flaɪt	flaɪɪŋ	42	bicycle	bike	baɪsɪkəl	baɪk	45	aircraft	airplane	eəkɹaft	eəpleɪn	46
weather	winter	wɛðə	wɪntə	36	**car**	**vehicle**	**kɑː**	**viəkəl**	**46**	shop	shopping	ʃɒp	ʃɒpɪŋ	44
weather	wet	wɛðə	wɛt	33	smoke	smoking	sməʊk	sməʊkɪŋ	45	weather	winter	wɛðə	wɪntə	36
bathroom	bedroom	baθɹuːm	bɛdɹuːm	38	food	fruit	fuːd	fɹuːt	31	**fruit**	**tomato**	**fɹuːt**	**təmɑːtəʊ**	**39**
painted	painting	peɪntɪd	peɪntɪŋ	42	shop	shopping	ʃɒp	ʃɒpɪŋ	44	**food**	**meat**	**fuːd**	**miːt**	**42**
bathroom	room	baθɹuːm	ɹuːm	37	bed	bedroom	bɛd	bɛdɹuːm	42	**truck**	**vehicle**	**tɹʌk**	**viəkəl**	**45**
skate	skating	skeɪt	skeɪtɪŋ	45	weather	winter	wɛðə	wɪntə	36	bedroom	room	bɛdɹuːm	ɹuːm	41
fish	fishing	fɪʃ	fɪʃɪŋ	46	light	sunlight	laɪt	sʌnlaɪt	44	smoke	smoking	sməʊk	sməʊkɪŋ	45
hanging	*sitting*	*haŋɪŋ*	*sɪtɪŋ*	*21*	light	lighting	laɪt	laɪtɪŋ	46	**city**	**town**	**sɪti**	**taʊn**	**39**
bacon	chicken	beɪkən	tʃɪkɪn	36	bedroom	room	bɛdɹuːm	ɹuːm	41	skate	skating	skeɪt	skeɪtɪŋ	45
hand	stand	hand	stand	31	bright	light	bɹaɪt	laɪt	39	bed	bedroom	bɛd	bɛdɹuːm	42
railway	subway	ɹeɪlweɪ	sʌbweɪ	37	surf	surfers	sɜːf	sɜːfəz	39	**burger**	**sandwich**	**bɜːgə**	**sandwɪtʃ**	**41**
children	fun	tʃɪldɹən	fʌn	27	race	racing	ɹeɪs	ɹeɪsɪŋ	42	river	water	ɹɪvə	wɔːtə	49
lunch	sandwich	lʌntʃ	sandwɪtʃ	42	sunny	sunset	sʌni	sʌnsɛt	40	**children**	**kids**	**tʃɪldɹən**	**kɪdz**	**46**
	Average similarity rating:			36.7		Average similarity rating:			41.6		Average similarity rating:			42.6

Table 4.7: Word pairs from the MEN dataset with highest cosine similarity in PHONEME GRU. Word pairs similar in form, but not meaning are in italics; Word pairs similar in meaning, but not in form are printed in bold. Human similarity judgements from the MEN dataset are included; possible values for similarity are integers between 0 (least similar) and 50 (most similar).

Chapter 5

Discussion

This study analysed the behaviour of PHONEME GRU, a model of visually grounded language learning from phoneme-level data. PHONEME GRU processes the input phoneme-by-phoneme, and predicts the visual context after arriving encountering the end of the sentence. The language knowledge encoded in PHONEME GRU was examined in a series of experiments. Evaluation on the learning task and exploratory analysis of the encoded knowledge showed that PHONEME GRU learns to encode word forms as well as visually grounded meaning, supporting the idea that lexical acquisition can be understood as interactive, rather than separate processes of word segmentation and cross-situational meaning acquisition.

5.1 Findings

The ability of PHONEME GRU to predict high level image features from phonetically transcribed caption was evaluated in an image retrieval task. PHONEME GRU was outperformed by the word-level recurrent WORD GRU model. Interestingly, however, it was more accurate than WORD SUM, the analog of a bag of words model. The word-based models have a definite advantage over PHONEME GRU in this task: they know exactly what the words in the sentence are. PHONEME GRU outperforms the summed word embedding-model despite this handicap.

The role of the three layers in the final architecture of PHONEME GRU was analysed in a series of experiments. Exploratory analysis showed differential processing time scales: replacing one phoneme in the sequence with another phoneme lead to a change in the activation vectors that became untraceable roughly halfway the sentence in the first layer, but in the second and third layer, it remained noticeable for the duration of the sentence. This is consistent with the findings of Hermans and Schrauwen (2013) of differential time scales in character-level language modelling. Differential processing time scales were also observed when comparing nearest neighbour sentences, as judged by cosine similarity. The average position of shared subsequences in nearest neighbours was closest to the end of the sentence for the first layer, and around the middle of the sentence for the third layer, with the second layer falling in between.

The rationale behind this analysis was that sentences that lead to similar activation vectors share parts of their form, but that the place of these parts may differ due to processing timescale differences. This interpretation is corroborated by the findings of the replacement experiment, but it may not be the whole story. It may also be that the third layer is simply less sensitive to form than the lower levels, and operates on more abstract semantic representations. If we were to pick two sentences at random, it is most likely that the average position of subsequences is around the middle of the sentence. Of course we are not picking sentences at random here - hopefully, we are picking sentence pairs that may appear in similar visual contexts. Could it be that the third layer cares more about (visual aspects of) meaning, whereas the lower layers care more about form?

A word boundary prediction experiment indicated that the lowest layer is most involved in encoding information about word form, as its activation pattern provided the most accurate predictor of word boundaries. Indeed, the higher levels were less reliable sources of information to make this prediction. A word similarity experiment on the MEN dataset showed that cosine similarity at any hidden layer and edit distance of two word pairs were negatively correlated, most strongly so in the lowest hidden layer, again indicating that information about word form is represented low in the stack of GRUs. Human judgements of semantic relatedness of word pairs, on the other hand, were correlated most strongly with cosine similarities between activation vectors of the third hidden layer, and decreasingly so for the lower layers.

These results indicate that there is indeed not just a difference in the timescale at which the layers operate, but also in the type of information they encode: the lowest layer mostly attends to form, and as the signal propagates through the stack of GRUs, representations become more semantic.

Similarity of activation patterns in the hidden layers of the word-based models were a much closer match to human similarity judgements than those in Phoneme GRU. We proposed that access to word form may help to determine the meaning of unfamiliar words, which might give Phoneme GRU an advantage in this respect, an effect that is present in NLP-applications that incorporate character information (Ling et al., 2015a; Plank et al., 2016). We did not observe a beneficial effect of the access to word form, but also did not look at unfamiliar forms specifically. It would be interesting to do additional analysis of the interpretation of novel forms by Phoneme GRU, taking into account well-known words that are similar in form.

If the beneficial effect of access to word form exists, it may be obscured by the advantage that the word level models have in our study: they know where the word boundaries are. There is no ambiguity between a sequence form and a word, as there is for Phoneme GRU. The correlations between human word similarity ratings and cosine similarities of activation vectors in Phoneme GRU are stronger for frequent words, suggesting that the disambiguation is learnable.

Taking the poorer encoding of semantic information into consideration, the fact that Phoneme GRU outperformed the Word Sum model on the retrieval task is especially impressive. Clearly, Phoneme GRU is well able to use sentence order information, as this is the advantage it has over Word Sum. The better performance of Word GRU may be due to the better encoding of semantic information, but it could also have been an indication of better capturing of long-range dependencies. The fact that Phoneme GRU outperforms Word Sum makes the latter explanation unlikely.

The word boundary prediction experiment showed that Phoneme GRU is quite selective in the information about word form it holds on to. The Language Model showed that the architecture itself in principle is well capable of tracking word form, yet even at the lowest level, Phoneme GRU only encodes about as much information about word form as a bigram model. Because Phoneme GRU was never explicitly instructed to identify word units, these results cannot really be interpreted in terms of under- or oversegmentation, as is often done for segmentation models. However, the fact that recall is particularly low in all levels (which is certainly not the case for the Language Model) does give the impression that some parts of the sequence form are simply forgotten about, or not attended to, by Phoneme GRU. A possible explanation is that not all sequence parts are of equal importance for the objective function: the prediction of the visual context. Kádár et al. (2016) show that the language and visual pathways of Imaginet attend to different parts of the language input. The visual pathway is much more attentive to content than to function words. It is very possible that this is also the case in Phoneme GRU.

The hierarchical representation of linguistic structure is not absolutely separated between the layers. Although there is a clear pattern of short-timescale information in the lower layers and larger dependencies in the higher layers, the third layer still encodes information about the phonetic form: its activation patterns were predictive of word boundaries, and similarities between word pairs at this level were more strongly correlated with edit distance than human similarity judgments are. It would be interesting to investigate exactly what information that is. In humans, both word phonological form and word meaning can act as primes, which hints that human representations encode both aspects. This may be the case in Phoneme GRU. It may also be that the model holds on only to semantically relevant aspects of form; for example, suffixes that indicate plurality.

5.2 Limitations

MS COCO does not consist of child-directed speech, and the input data misses many of the cues available to children. Word- and sentence stress markers from the input signal have been cleared out from the input signal, but given the potential role of these cues in word segmentation (Cutler and Butterfield, 1992), it may have been better to include them. However, the good encoding of form information in the Language Model suggests that Phoneme GRU selectively ignores certain information about form, rather than simply not being able to grasp it.

The MS COCO dataset does not contain visual social cues such as eye-gaze of the speaker, or the object that the speaker is holding, which have been proposed to play a role in language development in general and word learning in particular (e.g. Law et al. (2012), Yu and Smith (2013)). Exploitation of these cues has been implemented successfully in computational models of cross-situational learning with word-level language input (Frank et al., 2007; Lazaridou et al., 2016). Phoneme GRU may learn to use these cues, when given input data in which they are present.

Perhaps the most serious shortcoming of Phoneme GRU is that there is no mechanism in place for segmentation of speech without visual information. Infants do display the ability to segment speech without non-linguistic cues (Saffran et al., 1996a; Thiessen and Saffran, 2003). This limitation becomes all the more serious when we consider that the only study that investigated the use of visual cues for word segmentation in infants, did not find evidence that these cues were in fact exploited, whereas statistical cues in the language signal were (Thiessen, 2010). Phoneme GRU cannot account for these findings.

5.3 Future work

Phoneme GRU learns to encode information of word form as well as meaning, but it is rather selective in what word form knowledge it learns to encode, and is unable to learn from language unaccompanied by visual data. These shortcomings could be addressed by adding an explicit language learning task. A similar implementation as the multi-purpose Imaginet (Chrupała et al., 2015) model is probably not suitable to phoneme-level input. Being able to predict the next word, as in Imaginet, requires much more meaningful knowledge than predicting the next phoneme does, and a shared phoneme embedding layer probably does not provide the same benefits as a shared word embedding. However, some other implementation may be possible. As an example, we might add a phoneme prediction output layer which gets input from the first hidden layer and is trained on predicting the next phoneme in sequence. The first hidden layer would then be trained both on a language modelling task and the visual output prediction task, urging it to encode more information about form.

In chapter 1, it was mentioned that the work in this thesis is meant as an intermediate step towards grounded language learning from raw perceptual data: continuous visual input and acoustic speech. Phoneme GRU assumes knowledge of phonetic categories, but in reality, the developmental periods for word segmentation and phonetic category acquisition overlap. Perhaps counter-intuitively, it is possible that stepping down to the speech signal would also result in acquiring more knowledge about form, as lexical knowledge facilitates learning of phonetic categories (Feldman et al., 2013). Raw speech also incorporates non-phonemic acoustic cues to word boundaries, which Phoneme GRU does not have access to now.

Stepping down from phoneme transcriptions to acoustic speech will be a challenging endeavour, but worthwhile; it would result in a grounded model of language learning that covers the full scope of language learning, incorporating learning at and interaction between all intermediate levels of granularity.

5.4 Conclusion

Phoneme GRU acquires knowledge of word form, meaning, and sentential ordering, through visually grounded language learning from phoneme-level input in highly noisy and ambiguous visual context. The hidden layers in the deep recurrent architecture show a hierarchical organization, both in processing timescales, and in abstraction: lower levels encode more aspects of form, and higher levels encode more semantic knowledge. The results support an integrated view of language learning, and bode well for future endeavours in recurrent neural language learning from raw perceptual data.

Bibliography

Ando, R. K. and Lee, L. (2000). Mostly-unsupervised statistical segmentation of Japanese: Applications to Kanji. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 241–248. Association for Computational Linguistics.

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

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation.

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

Brent, M. R. and Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81(2):B33–B44.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49:1–47.

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

Chen, X., Qiu, X., Zhu, C., Liu, P., and Huang, X. (2015b). Long short-term memory neural networks for Chinese word segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1385–1394.

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

Christophe, A., Dupoux, E., Bertoncini, J., and Mehler, J. (1994). Do infants perceive word boundaries? an empirical study of the bootstrapping of lexical acquisition. *The Journal of the Acoustical Society of America*, 95(3):1570–1580.

Chrupała, G. (2013). Text segmentation with character-level text embeddings. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*.

Chrupała, G., Kádár, A., and Alishahi, A. (2015). Learning language through pictures. In *ACL*.

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

Cohen, P. and Adams, N. (2001). An algorithm for segmenting categorical time series into meaningful episodes. In *International Symposium on Intelligent Data Analysis*, pages 198–207. Springer.

Cunillera, T., Laine, M., Càmara, E., and Rodríguez-Fornells, A. (2010). Bridging the gap between speech segmentation and word-to-world mappings: Evidence from an audiovisual statistical learning task. *Journal of Memory and Language*, 63(3):295 – 305.

Cutler, A. and Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, 31(2):218–236.

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

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

Feldman, N. H., Griffiths, T. L., Goldwater, S., and Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, 120(4):751–778.

Feng, H., Chen, K., Deng, X., and Zheng, W. (2004). Accessor variety criteria for Chinese word extraction. *Computational Linguistics*, 30(1):75–93.

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

Glicksohn, A. and Cohen, A. (2013). The role of cross-modal associations in statistical learning. *Psychonomic Bulletin & Review*, 20(6):1161–1169.

Goldwater, S., Griffiths, T. L., and Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Harwath, D. and Glass, J. (2015). Deep multimodal semantic embeddings for speech and images. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 237–244. IEEE.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*.

Hermans, M. and Schrauwen, B. (2013). Training and analysing deep recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 190–198.

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

Johnson, M., Demuth, K., Jones, B., and Black, M. J. (2010). Synergies in learning words and their referents. In *Advances in neural information processing systems*, pages 1018–1026.

Jusczyk, P. W., Houston, D. M., and Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology*, 39(3):159–207.

Kádár, A., Chrupała, G., and Alishahi, A. (2016). Representation of linguistic form and function in recurrent neural networks. *CoRR*, abs/1602.08952.

Karpathy, A., Johnson, J., and Li, F.-F. (2015). Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Law, B., Houston-Price, C., and Loucas, T. (2012). Using gaze direction to learn words at 18 months: Relationships with later vocabulary. *LANGUAGE*, 4:3–14.

Lazaridou, A., Chrupała, G., Fernández, R., and Baroni, M. (2016). Multimodal semantic learning from child-directed input. In *The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, pages 740–755. Springer.

Ling, W., Luís, T., Marujo, L., Astudillo, R. F., Amir, S., Dyer, C., Black, A. W., and Trancoso, I. (2015a). Finding function in form: Compositional character models for open vocabulary word representation. *arXiv preprint arXiv:1508.02096*.

Ling, W., Trancoso, I., Dyer, C., and Black, A. W. (2015b). Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Mareschal, D. and Quinn, P. C. (2001). Categorization in infancy. *Trends in cognitive sciences*, 5(10):443–450.

Oord, A. v. d., Kalchbrenner, N., and Kavukcuoglu, K. (2016). Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and douard Duchesnay (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Plank, B., Søgaard, A., and Goldberg, Y. (2016). Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *arXiv preprint arXiv:1604.05529*.

Räsänen, O. and Rasilo, H. (2015). A joint model of word segmentation and meaning acquisition through cross-situational learning. *Psychological review*, 122(4):792.

Regier, T. (2005). The emergence of words: Attentional learning in form and meaning. *Cognitive science*, 29(6):819–865.

Roy, D. K. and Pentland, A. P. (2002). Learning words from sights and sounds: a computational model. *Cognitive Science*, 26(1):113 – 146.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Saffran, J., Aslin, R., and Newport, E. (1996a). Statistical learning by 8-month-old infants. *Science (New York, NY)*, 274(5294):1926–1928.

Saffran, J. R., Newport, E. L., and Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of memory and language*, 35(4):606–621.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1):39–91.

Synnaeve, G., Versteegh, M., and Dupoux, E. (2014). Learning words from images and speech. In *NIPS Workshop on Learning Semantics, Montreal, Canada*.

Thiessen, E. D. (2010). Effects of visual information on adults and infants auditory statistical learning. *Cognitive Science*, 34:1093–1106.

Thiessen, E. D. and Saffran, J. R. (2003). When cues collide: use of stress and statistical cues to word boundaries by 7-to 9-month-old infants. *Developmental psychology*, 39(4):706.

Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., and Ng, A. Y. (2016). Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*.

Yu, C. (2005). The emergence of links between lexical acquisition and object categorization: A computational study. *Connection science*, 17(3-4):381–397.

Yu, C. and Ballard, D. H. (2004). A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception (TAP)*, 1(1):57–80.

Yu, C. and Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS one*, 8(11):e79659.

Yurovsky, D., Yu, C., and Smith, L. B. (2012). Statistical speech segmentation and word learning in parallel: scaffolding from child-directed speech. *Frontiers in Psychology*, 3(374).