From phonemes to images: levels of representation in a recurrent neural model of visually-grounded language learning

Author 1 **Author 2**

Abstract

We present a model of visually-grounded language learning based on stacked gated recurrent neural networks which learns to predict visual features given an image description in the form of a sequence of phonemes. The learning task resembles that faced by human language learners who need to discover both structure and meaning from noisy and ambiguous data across modalities. We show that our model indeed learns to predict features of the visual context given phonetically transcribed image descriptions, and show that it represents linguistic information in a hierarchy of levels: lower layers in the stack are comparatively more sensitive to form, whereas higher layers are more sensitive to meaning.

1 Introduction

Children acquire their native language with little and weak supervision, exploiting noisy correlations between speech, visual, and other sensory signals, as well as via feedback from interaction with their peers and parents. Understanding this process is an important scientific challenge in its own right, but it also has potential to generate insights useful in engineering efforts to design conversational agents or robots. Computationally modeling the ability to learn linguistic form–meaning pairings has been the focus of much research, under scenarios simplified in a variety of ways, for example:

- distributional learning from pure word-word co-occurrences with no perceptual grounding (Landauer et al., 1998; Kiros et al., 2015);
- cross-situational learning with word sequences and sets of symbols representing sensory input (Siskind, 1996; Fazly et al., 2010);
- cross-situational learning using sensory audio and visual input, but with extremely limited sets of words and objects (Roy and Pentland, 2002; Iwahashi, 2003).

Some recent models have used more naturalistic, larger-scale inputs, for example in cross-modal distributional semantics (Lazaridou et al., 2015) or in implementations of the acquisition process trained on images paired with their descriptions (Chrupała et al., 2015). While in these works the representation of the *visual scene* consists of pixel-level perceptual data, the *linguistic* input consists of sentences segmented into discrete word symbols. In this paper we take a step towards addressing this major limitation, by using the phonetic transcription of input utterances. While this type of input is symbolic rather than perceptual, it goes a long way toward making the setting more naturalistic, and the acquisition problem more challenging: the learner may need to discover structure corresponding to morphemes, words and phrases in an unsegmented string of phonemes, and the length of the dependencies that need to be detected grows substantially when compared to word-level models.

Our contributions We design and implement a simple architecture based on stacked recurrent neural networks with Gated Recurrent Units (Chung et al., 2014): our model processes the utterance phoneme by phoneme while building a distributed low dimensional semantic representation through a series of recurrent layers. The model learns by projecting its sentence representation to image space and comparing it to the features of the corresponding visual scene.

We train this model on a phonetically transcribed version of MS-COCO (Lin et al., 2014) and show that it is able to successfully learn to understand aspects of sentence meaning from the visual signal, and exploits temporal structure in the input. In a number of experiments we show that different levels in the stack of recurrent layers represent different aspects of linguistic structure. Low levels focus on local, short time-scale spans of the input sequence, and are comparatively more sensitive to form. The top level encodes global aspects of the input sequence and is sensitive to visually salient elements of its meaning.

2 Related work

A major part of learning language consists in learning its structure, but in order to be able to communicate it is also of crucial to learn the relation of words to entities in the world. Grounded lexical acquisition is often modeled as cross-situational learning, a process of rule-based (Siskind, 1996) or statistical (Fazly et al., 2010; Frank et al., 2007) inference of word-to-referent mappings. Cross-situational models typically work on word-level language input and symbolic representations of the context. However, infants have to learn from continuous perceptual input. Lazaridou et al. (2016) partially remedy this shortcoming and propose a model of learning word meanings from text paired with continuous image representations; the limitation of their work is the toy evaluation dataset.

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 are interactive, rather than separate.

The Cross-channel Early Lexical Learning (CELL) model of Roy and Pentland (2002) and the more recent work of Räsänen and Rasilo (2015) take into account the continuous nature of the speech signal, and incorporate visual information as well. The CELL model learns to discover words in continuous speech through co-occurence with their visual referent, but the visual input only consists of the shape of single objects, effectively bypassing referential uncertainty. Räsänen and Rasilo (2015) propose a probabilistic joint model of word segmentation and meaning acquisition from raw speech and a set of possible referents that appear in the context. In both Roy and Pentland (2002) and Räsänen and Rasilo (2015) the visual context is considerably less noisy and ambiguous than that available to children.

There is an extensive line of research on image captioning (see Bernardi et al. (2016) for a recent overview). Typically, captioning models learn to recognize high-level image features and associate them with words. Inspired by both 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 two convolutional neural networks, a visual object recognition model and a 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. Both Harwath and Glass (2015) and Synnaeve et al. (2014) recognize words from pre-segmented speech.

Character-level input representations have recently gained attention in NLP. Ling et al. (2015) and Plank et al. (2016) use bidirectional LSTMs to compose characters into word embeddings, while Chung et al. (2016) propose machine translation model with character level output. These approaches exploit character-level information but crucially they assume that word boundaries are available in the input. Character-level neural NLP *without* explicit word boundaries in the input is studied in cases where fixed vocabularies are inherently problematic, e.g. with combined natural and programming language input (Chrupała, 2013) or when specifically dealing with misspelled words in automatic writing feedback (Xie et al., 2016).

Character-level language models are analyzed in Hermans and Schrauwen (2013) and Karpathy et al. (2015). Both studies show that character-level recurrent neural networks are sensitive to long-range dependencies: for example by keeping track of opening and closing parentheses over stretches of text. Hermans and Schrauwen (2013) describe the hierarchical organization that emerges during training, with higher layers processing information over longer timescales. In our work we show related effects in a model of visually-grounded language learning from unsegmented phonetic strings without word boundaries or strong cues such as whitespace and punctuation.

We use phonetic transcription of full sentences as a first step towards large-scale multimodal language learning from speech co-occurring with visual scenes. Our use of phonetic transcription rather than raw speech signal simplifies learning and allows us to perform experiments on the encoding of linguistic knowledge as reported in section 4 without additional annotation. In contrast to Roy and Pentland (2002) and Räsänen and Rasilo (2015), the visual input to our model consists of high-level visual features, which means it contains ambiguity and noise. In contrast to Synnaeve et al. (2014) and Harwath and Glass (2015), we consider full utterances rather than separate words. As Harwath and Glass (2015) note, the learning task of a multimodal model becomes significantly more complicated when the language input consists of unsegmented speech. The absence of word boundaries in the input data is essential to our aim.

To our knowledge, there is no work yet on multimodal phoneme or character-level language modeling with visual input. Although the language input in this study is low-level-symbolic rather than perceptual, the learning problem is similar to that of a human language learner: discover language structure as well as meaning, based on ambiguous and noisy data from another modality.

Chrupała et al. (2015) simulate visually grounded human language learning in face of noise and ambiguity in the visual domain. Their model predicts visual context given a sequence of words. While the visual input consists of a continuous representation, the language input consists of a sequence of words. The aim of this study is to take their approach one step further towards multimodal language learning from raw perceptual input. Kádár et al. (2016) develop techniques for understanding and interpretation of the representations of linguistic form and meaning in recurrent neural networks, and apply these to word-level models. In our work we share the goal of revealing the nature of emerging representations, but we do not assume words as their basic unit. Also, we are especially concerned with the emergence of a hierarchy of levels of representations in stacked recurrent networks.

3 Models

Consider a learner who sees a person pointing at a scene and uttering the unfamiliar phrase *Look, the monkeys're playing*. We may suppose that the learner will update her language understanding model such that the subsequent utterance of this phrase will evoke in her mind something close to impression of this visual scene. Our model is a particular instantiation of this simple idea.

3.1 Phon GRU

The architecture of our main model of interest, PHON GRU is schematically depicted in Figure 1 and consists of a phoneme encoding layer, followed by a stack of K Gated Recurrent Neural nets, followed by densely connected layer which maps the last hidden state of the top recurrent layer to a vector of visual features.

Gated Recurrent Units (GRU) were introduced in Cho et al. (2014) and Chung et al. (2014) as an attempt to alleviate the problem of vanishing gradient in standard simple recurrent nets as known since the work of Elman (1990). GRUs have a linear shortcut through timesteps which bypasses the nonlinearity and thus promotes gradient flow. Specifically, a GRU computes the hidden state at current time step, $\mathbf{h}_t$, as the linear combination of previous activation $\mathbf{h_{t-1}}$, and a new *candidate* activation $\tilde{\mathbf{h}}_t$:

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

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

$$\mathbf{z}_t = \sigma_s(\mathbf{W}_z\mathbf{x}_t + \mathbf{U}_z\mathbf{h}_{t-1}) \tag{2}$$

The candidate activation is computed as:

$$\tilde{\mathbf{h}}_t = \sigma(\mathbf{W}\mathbf{x}_t + \mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{t-1})) \tag{3}$$

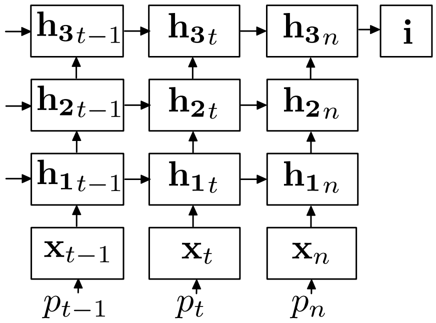

Figure 1: A three-timestep slice of the stacked recurrent architecture with three hidden layers.

ejʌŋwʊmənɹaɪdɪŋehɔːshəʊldɪŋeflag

A young woman riding a horse holding a flag

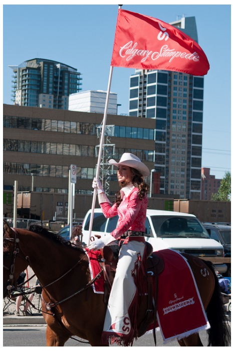

Figure 2: (Top) Example of a postprocessed phonetic transcription from eSpeak used as input to the PHON GRU model. (Bottom) Corresponding image (bottom)

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

$$\mathbf{r}_t = \sigma_s(\mathbf{W}_r\mathbf{x}_t + \mathbf{U}_r\mathbf{h}_{t-1}) \tag{4}$$

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

$$\mathrm{GRU}(\mathbf{X}, \mathbf{h}_0) = \mathrm{gru}(\mathbf{x}_n, \ldots, \mathrm{gru}(\mathbf{x}_2, \mathrm{gru}(\mathbf{x}_1, \mathbf{h}_0))) \tag{5}$$

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:

$$\mathrm{GRU}_2(\mathrm{GRU}_1(\mathbf{X}, \mathbf{h_1}_0), \mathbf{h_2}_0). \tag{6}$$

In our version of the Stacked GRU architecture we use *residualized* layers:

$$\mathrm{GRU}_{\mathrm{res}}(\mathbf{X}, \mathbf{h}_0) = \mathrm{GRU}(\mathbf{X}, \mathbf{h}_0) + \mathbf{X} \tag{7}$$

Residual convolutional networks were introduced by He et al. (2015), while Oord et al. (2016) showed their applicability to recurrent architectures. We adopt residualized layers here as we observed they speed up learning in stacks of several GRU layers.

Our gated recurrent units use steep sigmoids for gate activations:

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

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

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

There are two more components of our PHON GRU model: the phoneme encoding layer, and mapping from the final state of the top GRU layer to the image feature vector. The phoneme encoding layer is a simply a lookup table $\mathbf{E}$ 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:

$$\mathbf{x}_t = \mathbf{E}[:, p_t]. \tag{8}$$

Finally, we map the final state of the top GRU layer $\mathbf{h_K}_n$ to the vector of image features using a fully connected layer:

$$\hat{\mathbf{i}} = \mathbf{I}\mathbf{h_K}_n \tag{9}$$

Our main interest lies in recurrent phoneme-level modeling. However, in order to put the performance of the phoneme-level PHON GRU into perspective, we compare it to two word-level models. Importantly, the word models should **not** be seen as baselines, as they have access to word boundary and word identity information not available to PHON GRU.

3.2 Word GRU

The architecture of this model is the same as PHON GRU with the difference that we use words instead of phonemes as input symbols, use learnable word embeddings instead of fixed one-hot phoneme encodings, and reduce the number of layers in the GRU stack. See Section 4 for details.

3.3 Word Sum

The second model we use for comparison is a word-based non-sequential model, consisting of a word embedding matrix, a vector sum operator, and a mapping to the image feature vector:

$$\hat{\mathbf{i}} = \mathbf{I} \sum_{t=1}^{n} \mathbf{E}[:, w_t] \tag{10}$$

where w_t is the word at position t in the input utterance. This model simply learns word embeddings which are then summed into a single vector and projected to the target image vector. Thus this model does not have access to word sequence information, and is a distributed analog of a bag-of-words model.

4 Experiments

For all experiments, the models were trained on the training set of MS COCO. Textual input for the PHON GRU models was transcribed automatically using the grapheme-to-phoneme functionality with the default English voice of the eSpeak speech synthesis toolkit.[1] Stress and pause markers were removed, as well as word boundaries (after storing their position for use in experiments), leaving only phoneme symbols. See Figure 2 for an example transcription.

Visual input for all models was obtained by forwarding images through the 16-layer convolutional neural network described in Simonyan and Zisserman (2014) pre-trained on Imagenet (Russakovsky et al., 2014), and recording the activation vectors of the pre-softmax layer. The z-score transformation was applied to these features to ease optimization.

Most of the details of the three model types and training hyperparameters were adopted from related work, and adapted via informal exploration. Full exploration of the search space was not feasible due to the large number of adjustable settings in these models and their long running time. Given the importance of depth for our purposes, we did systematically explore the number of layers in the PHON GRU and WORD GRU models. A single layer is optimal for WORD GRU. For PHON GRU, see Section 4.1 below. Other important settings were as follows:

- All models: Implemented in Theano (Bastien et al., 2012), optimized with Adam (Kingma and Ba, 2014), initial learning rate of 0.0002, minibatch size of 64, gradient norm clipped to 5.0.
- WORD SUM: 1024-dimensional word embeddings, words with frequencies below 10 replaced by UNK token.
- WORD GRU: 1024-dimensional word embeddings, a single 1024 dimensional hidden layer, words with frequencies below 10 replaced by UNK token.
- PHON GRU: 1024-dimensional hidden layers.

4.1 Prediction of visual features

The models are trained and evaluated on the prediction of visual feature vectors from captions. We are not devising an image retrieval method, but this task reflects the ability to extract visually salient semantic information from language. For the experiments on the prediction of visual features all models were trained on the training set of MS COCO. As validation and test data we used a random sample of 5000 images each from the MS COCO validation set.

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

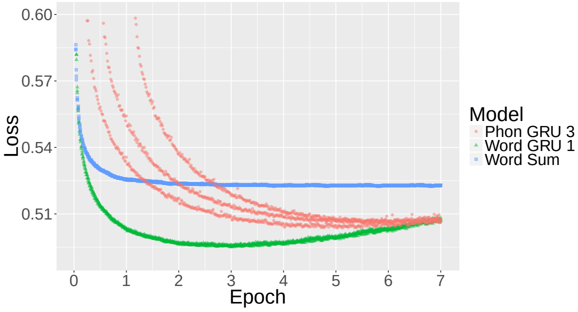

Figure 3: Value of the loss function on validation data during training. Three random initialization of each model are shown.

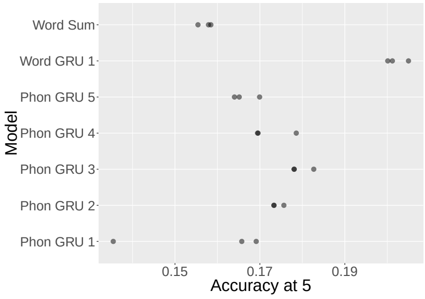

Figure 4: Validation accuracy at 5 on the image retrieval task.

Figure 3 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.

The Phonetic GRU model is more sensitive to the initialization: one can clearly distinguish three separate trajectories. The word-level models are much less affected by random initialization. In terms of the overall performance, the PHON GRU model falls between the WORD SUM model and the WORD GRU model.

We also evaluated the models on how well they perform when used to search images: for each validation sentence the model was used to predict the visual vector. The image vectors in the validation data were then ranked by cosine similarity to the predicted vector, and the proportion of times the correct image was among the top 5 was reported. By *correct* image we mean the one which the sentence was used to describe (even though often many other images are also good matches to the sentence).

In Figure 4 we report the validation accuracies on this task for the two word-level models, as well as for the Phon GRU model with different number of hidden layers. We trained each model version with three random initializations for each model setting, and evaluate after each epoch. We report the score of the best epoch for each initialization. The overall ranking of the models matches the direct evaluation of the loss function above: the phoneme-level models are in between the two word-level models. PHON GRU with three hidden layers is the best of the phoneme-level models.

In Table 1 we show the accuracies of the best version of each of the models types on the test images; these are also the model versions used in all subsequent experiments. The accuracy @ 5 for the WORD GRU is comparable to what Chrupała et al. (2015) report for their multitask IMAGINET model, whose visual pathway has the same structure.

4.2 Word boundary prediction

To explore the sensitivity of the PHON GRU model to linguistic structure at the sub-word level, we investigated the encoding of information about word-ends in the hidden layers. Logistic regression models were trained on activation patterns of the hidden layers at all timesteps, with the objective of identifying phonemes that preceded a word boundary. For comparison, we also trained logistic regression models on *n*-gram data to perform the same tasks, with positional phoneme *n*-grams in the range 1-*n*. The location of the word boundaries was taken from the eSpeak transcriptions, which mostly matches the location of word boundaries according to conventional English spelling. 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]. 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. The optimal value of regularization parameter *C* was determined using `GridSearchCV` with 5-fold cross validation on the training set, after which the model with the optimal settings was trained on all training data.

Table 2 reports the scores 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, 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 scores of the logistic regression model based on the activation patterns of the lowest hidden layer are comparable to those of a bigram logistic regression model.

These results indicate that information on sub-word structure is only partially encoded by PHON GRU, and is mostly absent by the time the signal from the input propagates to the top layer. The bottom layer does learn to encode a fair amount of word boundary information, but the prediction score substantially below 100% indicates that it is rather selective.

Model	Acc @ 5
WORD SUM	0.158
WORD GRU	0.205
PHON GRU	0.180

Table 1: Image retrieval accuracy at 5 on test data for the versions of WORD SUM, WORD GRU and PHON GRU chosen by validation.

Model		**Acc**	**Prec**	**Rec**
Majority		0.71		
Phon GRU	Layer 1	0.88	0.82	0.78
	Layer 2	0.86	0.79	0.71
	Layer 3	0.82	0.74	0.60
n-gram	$n = 1$	0.80	0.79	0.41
	$n = 2$	0.87	0.79	0.78
	$n = 3$	0.93	0.86	0.90
	$n = 4$	0.95	0.90	0.93

Table 2: Prediction scores of linear regression models based on activation vectors of PHON GRU and on positional n-grams

4.3 Word similarity

To understand the encoding of semantic information in PHON GRU, we analyzed the cosine similarity of activation vectors for word pairs from the MEN dataset (Bruni et al., 2014), and compared them to human similarity judgements. For each word pair in the MEN dataset, the words were transcribed phonetically using eSpeak and then fed to PHON 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. In contrast to WORD GRU and WORD SUM, PHON GRU has access to the sub-word structure. To explore the role of phonemic form in word similarity, a measure of phonemic difference was included: the Levenshtein distance between the phonetic transcriptions of the two words, normalized by the length of the longer transcription.

Table 3 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. The second column in Table 3 shows the correlations when only taking into account the 1283 word pairs of which both words appear at least 100 times in the training data. Correlations for both WORD GRU and WORD SUM are considerably higher than for PHON GRU. This is expected given that these are word level models with explicit word-embeddings, while PHON GRU builds word representations by forwarding phoneme-level input through several layers of processing.

	All words	Frequent words
PHON 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 3: Spearman's correlation coefficient between word-word cosine similarity and human similarity judgements. All correlations significant at $p < 1\text{e}{-4}$. Frequent words appear at least 100 times in the training data.

Layer	ρ
1	-0.30
2	-0.24
3	-0.15

Table 4: Spearman's rank correlation coefficient between PHON GRU cosine similarity and phoneme-level edit distance. All correlations significant at $p < 1\text{e}{-15}$.

Table 4 shows Spearman's rank correlation coefficient between the edit distance and the cosine similarity of activation vectors at the hidden layers of PHON GRU. As expected, edit distance and cosine similarity of the activation vectors are negatively correlated: words which are more similar in form are also more similar according to the model.[2]

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.

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 reflect increasing levels of representation: whereas at the lowest level mostly encodes information about form, the highest layer mostly encodes semantic information.

4.4 Position of shared substrings

Here we quantify the time-scale at which information is retained in the different layers of PHON GRU. We looked at the location of phoneme strings shared by sentences and their nearest neighbours on validation data. We determined each sentences' nearest neighbour for each hidden layer in PHON GRU. The nearest neighbour is the sentence for which the activation vector at the end of sentence symbol has the smallest cosine distance to the activation vector of the original sentence. The position of matching substrings is the average position in the original sentence of symbols in substrings that are shared by the neighbour sentences, counted from the end of the sentence. A high mean average substring position thus means that the shared substring(s) appear early in the sentence. This gives an indirect measure of the timescale at which the different layers operate. Table 5 shows an example.

As can be seen in Table 6, 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. This indicates a difference between the layers with regards to the timescale they represent. Whereas in the lowest layer only information from the latest timesteps is present, the higher layers retain the input signal over longer timescales.

Layer 1
A metallic bench **on a path in** the **park**
A man riding a bicycle **on a path** in a **park**
Layer 3
A metallic **bench** on a path in the **park**
A stone park **bench** sitting in an empty green **park**

Table 5: An illustrative sentence with its nearest neighbour at layer 1 and layer 3. For readability, sentences are displayed in conventional spelling, and only highlight matching substrings of length ≥ 3. In reality we used phonetic transcriptions to compute shared substring positions, and substrings of all lengths.

Layer	Mean position
1	12.1
2	14.9
3	16.8

Table 6: Average position of symbols in shared substrings between nearest neighbour sentences according to PHON GRU representations at the different layers. Positions are indexed from end of string.

5 Future work

Although our analyses show 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, and to what extent it is analogous to language representation in the mind of human speakers. In humans both word phonological form and word meaning can act as primes, which is somewhat reminiscent of the behavior of our model.

Finally, we would like to take the next step towards grounded learning of language from raw perceptual input, and apply models similar to the one described here to acoustic speech signal coupled with visual

[2] 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 ($p < 1\mathrm{e}{-}5$).

input. We expect this to be a challenging but essential endeavor.

References

Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Bergeron, Nicolas Bouchard, and Yoshua Bengio. 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *arXiv preprint arXiv:1601.03896*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49:1–47.

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

Grzegorz Chrupała, Akos Kádár, and Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Grzegorz Chrupała. 2013. Text segmentation with character-level text embeddings. In *Proceedings of the ICML workshop on Deep Learning for Audio, Speech and Language Processing*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Deep Learning and Representation Learning Workshop*.

Junyoung Chung, Kyunghyun Cho, and Yoshua Bengio. 2016. A character-level decoder without explicit segmentation for neural machine translation. *arXiv preprint arXiv:1603.06147*.

Toni Cunillera, Matti Laine, Estela Càmara, and Antoni Rodríguez-Fornells. 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.

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

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

Michael C. Frank, Noah D. Goodman, and Joshua B. Tenenbaum. 2007. A Bayesian framework for cross-situational word-learning. In *Advances in Neural Information Processing Systems*, volume 20.

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

David Harwath and James Glass. 2015. Deep multimodal semantic embeddings for speech and images. *arXiv:1511.03690*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *arXiv:1512.03385*.

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

Naoto Iwahashi. 2003. Language acquisition through a human–robot interface by combining speech, visual, and behavioral information. *Information Sciences*, 156(1):109–121.

Ákos Kádár, Grzegorz Chrupała, and Afra Alishahi. 2016. Representation of linguistic form and function in recurrent neural networks. *CoRR*, abs/1602.08952.

Andrej Karpathy, Justin Johnson, and Fei-Fei Li. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*, pages 3276–3284.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL HLT 2015 (2015 Conference of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies)*.

Angeliki Lazaridou, Grzegorz Chrupała, Raquel Fernández, and Marco Baroni. 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*.

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

Wang Ling, Tiago Luís, Luís Marujo, Ramón Fernandez Astudillo, Silvio Amir, Chris Dyer, Alan W Black, and Isabel Trancoso. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of EMNLP*.

Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. 2016. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In *Proceedings of ACL 2016*. Association for Computational Linguistics (ACL).

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

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Jeffrey M. Siskind. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1-2):39–91.

Gabriel Synnaeve, Maarten Versteegh, and Emmanuel Dupoux. 2014. Learning words from images and speech. In *NIPS Workshop on Learning Semantics, Montreal, Canada*.

Erik D Thiessen. 2010. Effects of visual information on adults' and infants' auditory statistical learning. *Cognitive Science*, 34(6):1093–1106.

Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y Ng. 2016. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*.

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