Learning language through pictures
Chrupala, Grzegorz; Kadar, Akos; Alishahi, Afra

Published in:
Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)

Document version:
Publisher's PDF, also known as Version of record

Publication date:
2015

Link to publication

Citation for published version (APA):

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright, please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Learning language through pictures
Grzegorz Chrupała g.chrupala@uvt.nl
Ákos Kádár a.kadar@uvt.nl
Afra Alishahi a.alishahi@uvt.nl
Tilburg Center for Cognition and Communication
Tilburg University

Abstract

We propose IMAGINET, a model of learning visually grounded representations of language from coupled textual and visual input. The model consists of two Gated Recurrent Unit networks with shared word embeddings, and uses a multi-task objective by receiving a textual description of a scene and trying to concurrently predict its visual representation and the next word in the sentence. Mimicking an important aspect of human language learning, it acquires meaning representations for individual words from descriptions of visual scenes. Moreover, it learns to effectively use sequential structure in semantic interpretation of multi-word phrases.

1 Introduction

Vision is the most important sense for humans and visual sensory input plays an important role in language acquisition by grounding meanings of words and phrases in perception. Similarly, in practical applications processing multimodal data where text is accompanied by images or videos is increasingly important. In this paper we propose a novel model of learning visually-grounded representations of language from paired textual and visual input. The model learns language through comprehension and production, by receiving a textual description of a scene and trying to “imagine” a visual representation of it, while predicting the next word at the same time.

The full model, which we dub IMAGINET, consists of two Gated Recurrent Unit (GRU) networks coupled via shared word embeddings. IMAGINET uses a multi-task Caruana (1997) objective: both networks read the sentence word-by-word in parallel; one of them predicts the feature representation of the image depicting the described scene after reading the whole sentence, while the other one predicts the next word at each position in the word sequence. The importance of the visual and textual objectives can be traded off, and either of them can be switched off entirely, enabling us to investigate the impact of visual vs textual information on the learned language representations.

Our approach to modeling human language learning has connections to recent models of image captioning (see Section 2). Unlike in many of these models, in IMAGINET the image is the target to predict rather than the input, and the model can build a visually-grounded representation of a sentence independently of an image. We can directly compare the performance of IMAGINET against a simple multivariate linear regression model with bag-of-words features and thus quantify the contribution of the added expressive power of a recurrent neural network.

We evaluate our model’s knowledge of word meaning and sentence structure through simulating human judgments of word similarity, retrieving images corresponding to single words as well as full sentences, and retrieving paraphrases of image captions. In all these tasks the model outperforms the baseline; the model significantly correlates with human ratings of word similarity, and predicts appropriate visual interpretations of single and multi-word phrases. The acquired knowledge of sentence structure boosts the model’s performance in both image and caption retrieval.

2 Related work

Several computational models have been proposed to study early language acquisition. The acquisition of word meaning has been mainly modeled using connectionist networks that learn to associate word forms with semantic or perceptual features (e.g., Li et al., 2004; Coventry et al., 2005; Regier, 2005), and rule-based or probabilistic implementations which use statistical reg-
ularities observed in the input to detect associations between linguistic labels and visual features or concepts (e.g., Siskind, 1996; Yu, 2008; Fazly et al., 2010). These models either use toy languages as input (e.g., Siskind, 1996), or child-directed utterances from the CHILDES database (MacWhinney, 2014) paired with artificially generated semantic information. Some models have investigated the acquisition of terminology for visual concepts from simple videos (Fleischman and Roy, 2005; Skocaj et al., 2011). Lazaridou et al. (2015) adapt the skip-gram word-embedding model (Mikolov et al., 2013) for learning word representations via a multi-task objective similar to ours, learning from a dataset where some words are individually aligned with corresponding images. All these models ignore sentence structure and treat inputs as bags of words.

A few models have looked at the concurrent acquisition of words and some aspect of sentence structure, such as lexical categories (Alishahi and Chrupala, 2012) or syntactic properties (Howell et al., 2005; Kwiatkowski et al., 2012), from utterances paired with an artificially generated representation of their meaning. To our knowledge, no existing model has been proposed for concurrent learning of grounded word meanings and sentence structure from large scale data and realistic visual input.

Recently, the engineering task of generating captions for images has received a lot of attention (Karpathy and Fei-Fei, 2014; Mao et al., 2014; Kiros et al., 2014; Donahue et al., 2014; Venugopalan et al., 2014; Vinyals et al., 2014; Zweig, 2014; Chen and Zitnick, 2014; Fang et al., 2014). From the point of view of modeling, the research most relevant to our interests is that of Chen and Zitnick (2014). They develop a model based on a context-dependent recurrent neural network (Mikolov and Zweig, 2012) which simultaneously processes textual and visual input and updates two parallel hidden states. Unlike theirs, our model receives the visual target only at the end of the sentence and is thus encouraged to store in the final hidden state of the visual pathway all aspects of the sentence needed to predict the image features successfully. Our setup is more suitable for the goal of learning representations of complete sentences.

3 Models

IMAGINET consists of two parallel recurrent pathways coupled via shared word embeddings. Both pathways are composed of Gated Recurrent Units (GRU) first introduced by Cho et al. (2014) and Chung et al. (2014). GRUs are related to the Long Short-Term Memory units (Hochreiter and Schmidhuber, 1997), but do not employ a separate memory cell. In a GRU, activation at time $t$ is the linear combination of previous activation, and candidate activation:

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t$$

(1)

where $\odot$ is elementwise multiplication. The update gate determines how much the activation is updated:

$$z_t = \sigma(s(W_z x_t + U_z h_{t-1}))$$

(2)

The candidate activation is computed as:

$$\tilde{h}_t = \sigma(W x_t + U(r_t \odot h_{t-1}))$$

(3)

The reset gate is defined as:

$$r_t = \sigma(s(W_r x_t + U_r h_{t-1}))$$

(4)

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:

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

Figure 1 illustrates the structure of the network. The word embeddings is a matrix of learned parameters $W_e$ with each column corresponding to a vector for a particular word. The input word symbol $S_t$ of sentence $S$ at each step $t$ indexes into the embeddings matrix and the vector $x_t$ forms input to both GRU networks:

$$x_t = W_e[:, S_t]$$

(5)
This input is mapped into two parallel hidden states, $h_V^t$ along the visual pathway, and $h_T^t$ along the textual pathway:

$$h_V^t = \text{GRU}^V(h_{V,t-1}, x_t) \quad (6)$$
$$h_T^t = \text{GRU}^T(h_{T,t-1}, x_t) \quad (7)$$

The final hidden state along the visual pathway $h_V^t$ is then mapped to the predicted target image representation $i$ by the fully connected layer with parameters $V$ and the clipped rectifier activation:

$$\hat{i} = \sigma(Vh_V^t) \quad (8)$$

Each hidden state along the textual pathway $h_T^t$ is used to predict the next symbol in the sentence $S$ via a softmax layer with parameters $\Theta$:

$$p(S_{t+1}|S_{1:t}) = \text{softmax}(Lh_T^t) \quad (9)$$

The loss function whose gradient is backpropagated through time to the GRUs and the embeddings is a composite objective with terms penalizing error on the visual and the textual targets simultaneously:

$$L(\theta) = \alpha L_T^T(\theta) + (1 - \alpha) L_V^V(\theta) \quad (10)$$

where $\theta$ is the set of all IMAGINET parameters. $L_T^T$ is the cross entropy function:

$$L_T^T(\theta) = -\frac{1}{\tau} \sum_{t=1}^{\tau} \log p(S_t|S_{1:t}) \quad (11)$$

while $L_V^V$ is the mean squared error:

$$L_V^V(\theta) = \frac{1}{K} \sum_{k=1}^{K} (\hat{i}_k - i_k)^2 \quad (12)$$

By setting $\alpha$ to 0 we can switch off the whole textual pathway and obtain the VISUAL model variant. Analogously, setting $\alpha$ to 1 gives the TEXTUAL model. Intermediate values of $\alpha$ (in the experiments below we use 0.1) give the full MULTITASK version. Finally, as baseline for some of the tasks we use a simple linear regression model LINREG with a bag-of-words representation of the sentence:

$$\hat{i} = Ax + b \quad (13)$$

where $\hat{i}$ is the vector of predicted image features, $x$ is the vector of word counts for the input sentence and $(A, b)$ the parameters of the linear model estimated via $L_2$-penalized sum-of-squared-errors loss.

4 Experiments

Settings The model was implemented in Theano (Bastien et al., 2012; Bergstra et al., 2010) and optimized by Adam (Kingma and Ba, 2014).\(^1\) The fixed 4096-dimensional target image representation come from the pre-softmax layer of the 16-layer CNN (Simonyan and Zisserman, 2014). We used 1024 dimensions for the embeddings and for the hidden states of each of the GRU networks. We ran 8 iterations of training, and we report either full learning curves, or the results for each model after iteration 7 (where they performed best for the image retrieval task). For training we use the standard MS-COCO training data. For validation and test, we take a sample of 5000 images each from the validation data.

4.1 Word representations

We assess the quality of the learned embeddings for single words via two tasks: (i) we measure similarity between embeddings of word pairs and compare them to elicited human ratings; (ii) we examine how well the model learns visual representations of words by projecting word embeddings into the visual space, and retrieving images of single concepts from ImageNet.

Word similarity judgment For similarity judgment correlations, we selected two existing benchmarks that have the largest vocabulary overlap with our data: MEN 3K (Bruni et al., 2014) and SimLex-999 (Hill et al., 2014). We measure the similarity between word pairs by computing the cosine similarity between their embeddings from three versions of our model, VISUAL, MULTITASK and TEXTUAL, and the baseline LINREG.

Table 1 summarizes the results. All IMAGINET models significantly correlate with human similarity judgments, and outperform LINREG. Examples of word pairs for which MULTITASK cap-

<table>
<thead>
<tr>
<th></th>
<th>SimLex</th>
<th>MEN 3K</th>
</tr>
</thead>
<tbody>
<tr>
<td>VISUAL</td>
<td>0.32</td>
<td>0.57</td>
</tr>
<tr>
<td>MULTITASK</td>
<td>0.39</td>
<td>0.63</td>
</tr>
<tr>
<td>TEXTUAL</td>
<td>0.31</td>
<td>0.53</td>
</tr>
<tr>
<td>LINREG</td>
<td>0.18</td>
<td>0.23</td>
</tr>
</tbody>
</table>

\(^1\)Code available at github.com/gchrupala/imaginet.
Table 2: Accuracy@5 of retrieving images with compatible labels from ImageNet.

<table>
<thead>
<tr>
<th>VISUAL</th>
<th>MultiTask</th>
<th>LinReg</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.38</td>
<td>0.38</td>
<td>0.33</td>
</tr>
</tbody>
</table>

fures human similarity judgments better than VISUAL include antonyms (dusk, dawn), collocations (sexy, smile), or related but not visually similar words (college, exhibition).

**Single-word image retrieval**  In order to visualize the acquired meaning for individual words, we use images from the ILSVRC2012 subset of ImageNet (Russakovsky et al., 2014) as benchmark. Labels of the images in ImageNet are synsets from WordNet, which identify a single concept in the image rather than providing descriptions of its full content. Since the synset labels in ImageNet are much more precise than the descriptions provided in the captions in our training data (e.g., elkhound), we use synset hypernyms from WordNet as substitute labels when the original labels are not in our vocabulary.

We extracted the features from the 50,000 images of the ImageNet validation set. The labels in this set result in 393 distinct (original or hypernym) words from our vocabulary. Each word was projected to the visual space by feeding it through the model as a one-word sentence. We ranked the vectors corresponding to all 50,000 images based on their similarity to the predicted vector, and measured the accuracy of retrieving an image with the correct label among the top 5 ranked images (Accuracy@5). Table 2 summarizes the results: VISUAL and MultiTask learn more accurate word meaning representations than LinReg.

### 4.2 Sentence structure

In the following experiments, we examine the knowledge of sentence structure learned by IMAGINET, and its impact on the model performance on image and paraphrase retrieval.

**Image retrieval**  We retrieve images based on the similarity of their vectors with those predicted by IMAGINET in two conditions: sentences are fed to the model in their original order, or scrambled. Figure 2 (left) shows the proportion of sentences for which the correct image was in the top 5 highest ranked images for each model, as a function of the number of training iterations: both models outperform the baseline. MultiTask is initially better in retrieving the correct image, but eventually the gap disappears. Both models perform substantially better when tested on the original captions compared to the scrambled ones, indicating that models learn to exploit aspects of sentence structure. This ability is to be expected for MultiTask, but the VISUAL model shows a similar effect to some extent. In the case of VISUAL, this sensitivity to structural aspects of sentence meaning is entirely driven by how they are reflected in the image, as this models only receives the visual supervision signal.

Qualitative analysis of the role of sequential structure suggests that the models are sensitive to the fact that periods terminate a sentence, that sentences tend not to start with conjunctions, that topics appear in sentence-initial position, and that words have different importance as modifiers versus heads. Figure 3 shows an example; see supplementary material for more.

**IMAGINET vs captioning systems**  While it is not our goal to engineer a state-of-the-art image retrieval system, we want to situate IMAGINET’s performance within the landscape of image retrieval results on captioned images. As most of these are on Flickr30K (Young et al., 2014), we ran MultiTask on it and got an accuracy@5 of 32%, within the range of numbers reported in previous work: 29.8% (Socher et al., 2014), 31.2% (Mao et al., 2014), 34% (Kiros et al., 2014) and 37.7% (Karpathy and Fei-Fei, 2014). Karpathy and Fei-Fei (2014) report 29.6% on MS-COCO, but with additional training data.
Table 3: Examples of two nearest neighbors retrieved by MULTI TASK for original and scrambled captions.

<table>
<thead>
<tr>
<th>Original</th>
<th>Scrambled</th>
</tr>
</thead>
<tbody>
<tr>
<td>a couple of horses UNK their head over a rock pile</td>
<td>rock couple their head pile a a UNK over of horses</td>
</tr>
<tr>
<td>rank 1: two brown horses hold their heads above a rocky wall</td>
<td>rank 1: an image of a man on a couple of horses</td>
</tr>
<tr>
<td>rank 2: two horses looking over a short stone wall</td>
<td>rank 2: looking in to a straw lined pen of cows</td>
</tr>
<tr>
<td>rank 1: a cute baby playing with a cell phone</td>
<td>rank 1: phone playing cute cell a with baby a</td>
</tr>
<tr>
<td>rank 2: small baby smiling at camera and talking on phone</td>
<td>rank 2: someone is using their phone to send a text or play a game</td>
</tr>
<tr>
<td>rank 2: a smiling baby holding a cell phone up to ear</td>
<td>rank 2: a camera is placed next to a cellular phone</td>
</tr>
</tbody>
</table>

Figure 3: For the original caption MULTI TASK understands kitchen as a modifier of headword utensils, which is the topic. For the scrambled sentence, the model thinks kitchen is the topic.

**Paraphrase retrieval** In our dataset each image is paired with five different captions, which can be seen as paraphrases. This affords us the opportunity to test IMAGINET’s sentence representations on a non-visual task. Although all models receive one caption-image pair at a time, the co-occurrence with the same image can lead the model to learn structural similarities between captions that are different on the surface. We feed the whole set of validation captions through the trained model and record the final hidden visual state $h^V_{f}$. For each caption we rank all others according to cosine similarity and measure the proportion of the ones associated with the same image among the top four highest ranked. For the scrambled condition, we rank original captions against a scrambled one. Figure 2 (right) summarizes the results: both models outperform the baseline on ordered captions, but not on scrambled ones. As expected, MULTI TASK is more affected by manipulating word order, because it is more sensitive to structure. Table 3 shows concrete examples of the effect of scrambling words in what sentences are retrieved.

**5 Discussion**

IMAGINET is a novel model of grounded language acquisition which simultaneously learns word meaning representations and knowledge of sentence structure from captioned images. It acquires meaning representations for individual words from descriptions of visual scenes, mimicking an important aspect of human language learning, and can effectively use sentence structure in semantic interpretation of multi-word phrases. In future we plan to upgrade the current word-prediction pathway to a sentence reconstruction and/or sentence paraphrasing task in order to encourage the formation of representations of full sentences. We also want to explore the acquired structure further, especially for generalizing the grounded meanings to those words for which visual data is not available.

**Acknowledgements**

The authors would like to thank Angeliki Lazari-dou and Marco Baroni for their many insightful comments on the research presented in this paper.

**References**


Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, James Bergstra, Ian J. Goodfellow, Arnaud Berg-


Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositional-


