Lingusitic Analysis of multi-modal Recurrent Neural Networks

Ákos Kádár
Tilburg University
a.kadar@uvt.nl

Grzegorz Chrupała
Tilburg University
g.a.chrupala@uvt.nl

Afra Alishahi
Tilburg University
a.alishahi@uvt.nl

1 Introduction

Recurrent neural networks (RNN) have gained a reputation for beating state-of-the-art results on many NLP benchmarks and for learning representations of words and larger linguistic units that encode complex syntactic and semantic structures. However, it is not straight-forward to understand how exactly these models make their decisions. Recently Li et al. (2015) developed methods to provide linguistically motivated analysis for RNNs trained for sentiment analysis. Here we focus on the analysis of a multi-modal Gated Recurrent Neural Network (GRU) architecture trained to predict image-vectors - extracted from images using a CNN trained on ImageNet - from their corresponding descriptions. We propose two methods to explore the importance of grammatical categories with respect to the model and the task. We observe that the model pays most attention to head-words, noun subjects and adjectival modifiers and least to determiners and coordinations.

2 Method

We used the IMAGINET model from Chrupała et al. (2015), trained on the MSCOCO dataset (Lin et al., 2014). It learns visually grounded meaning representations from textual and visual input and consists of two GRU pathways, TEXTUAL and VISUAL, with a shared word-embedding matrix. The inputs to the model are pairs of captions and their corresponding images. Each sentence is mapped to two sequences of hidden states: one by TEXTUAL and another by VISUAL. At each timestep TEXTUAL predicts the next word in the sentence from its current hidden state h_t^T, while VISUAL predicts the image vector from its last hidden representation h_{full}^V. The model is trained using a multi-task objective which combines cross-entropy loss for the word predictions and a mean squared error for the image predictions.

We focus our analysis on the hidden states and update-gate activations of VISUAL to assess the impact of syntactic structure on the learned meaning representations of sentences used to predict images. For each input sentence of length n, VISUAL produces n hidden activations $h_1^V, ..., h_n^V$ and n update-gate activations $z_1^V, ..., z_n^V$. We associate each word in the input sentence with their part-of-speech (POS) and dependency relation (DepRel) labels[1], and assess the contribution of the (word, POS, DepRel) tuples by estimating the following two scores:

1. d_{red} measures the distance reduction at each step by calculating the cosine distance between the current h_t and the last hidden state h_{full} and subtracting it from the previous distance: $d_{\text{red}}^t = d_{\text{red}}^{t-1} - cos(h_t, h_{\text{full}})$. The idea is to see how much each word brings the current state closer to, or further away from, the final interpretation.

2. z_{mean} assigns the average activation of the update-gate z at time step t to the tuple at positions t. The activation function for z is sigmoid, therefore it has values between 0-1. High values of z_{mean} indicate that the model places more importance on the previous tuples until $t-1$ than on the current one at t.

3 Results

We measure d_{red} and z_{mean} for every position in the first 5000 captions from the validation portion of MSCOCO and use them to analyze the importance of both POS and DepRel categories. We only report results on the grammatical categories that appear at least 500 times. Figure 1 demonstrates the d_{red} measurements for each word in an example sentence. A large distance between two adjacent

[1]We used the dependency parser from Martins et al. (2013) for both the POS and DepRel tags.

Proceedings of the 2015 Workshop on Vision and Language (VL'15), pages 8–9,
Lisbon, Portugal, 18 September 2015.

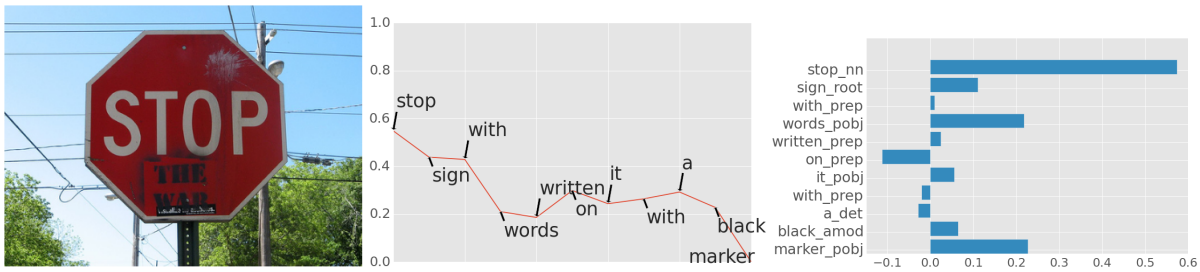

Figure 1: An example of the impact of each word in the sentence *Stop sign with words written on it with a black marker*, measured by d_{red}. Left: best retrieved image; middle: reduction of distance from h^V_{full}; right: d_{red} scores for each word.

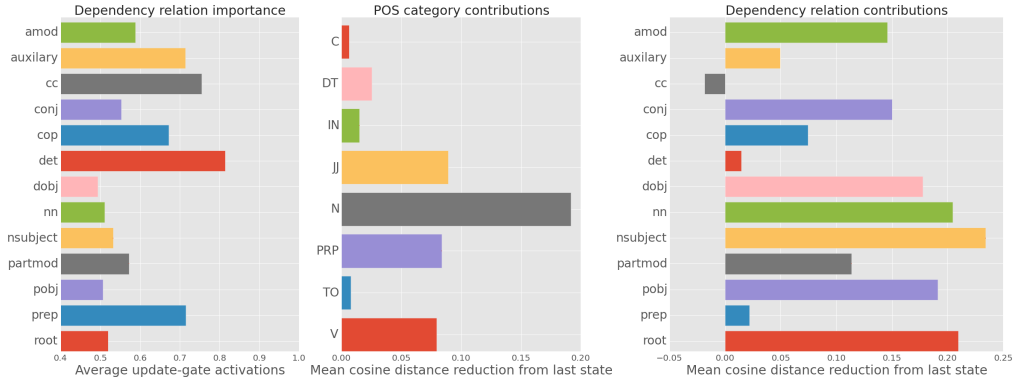

Figure 2: Importance of dependency relations as measured by z_{mean} on the left chart. Contribution of POS categories (middle) and DepRel categories (right) measured by d_{red}.

words signals the arrival of a highly informative word. Figure 2 shows the impact of both measures for each grammatical category. The low z_{mean} scores (left) for the roots, adjectival modifiers (amod), direct objects (dobj), noun compound modifiers (nn), noun subjects (nsubj), conjuncts (conj) and objects of prepositions (pobj) suggest that the model remembers words of these categories, while prefers to forget determiners (det), coordinations (cc), prepositions (prep) and auxiliaries. As indicated by the high d_{red} scores (middle graph), nouns (N), adjectives (JJ) verbs (V) and prepositions (PRP) provide the largest contribution to the meaning representations of the sentences, while determiners (DET) and conjunctions (C) provide the least. The d_{red} scores for DepRels are in line with the z_{mean} scores; they highlight the importance of nsubj, nn, amod, pobj and dobj.

4 Conclusions

We propose two measures to assess the impact of grammatical categories on sentence representations learned for predicting images. The observed patterns likely reflect the visual salience and informativeness of the lexical items associated with each category. They also provide insights into the details of the task e.g.: nouns came out significantly more important then other content word categories, indicating that predicting the correct entities is the most important aspect of the task.

References

[Chrupała et al.2015] Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. *arXiv preprint arXiv:1506.03694*.

[Li et al.2015] Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in nlp. *arXiv preprint arXiv:1506.01066*.

[Lin et al.2014] 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.

[Martins et al.2013] André FT Martins, Miguel Almeida, and Noah A Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *ACL (2)*, pages 617–622. Citeseer.