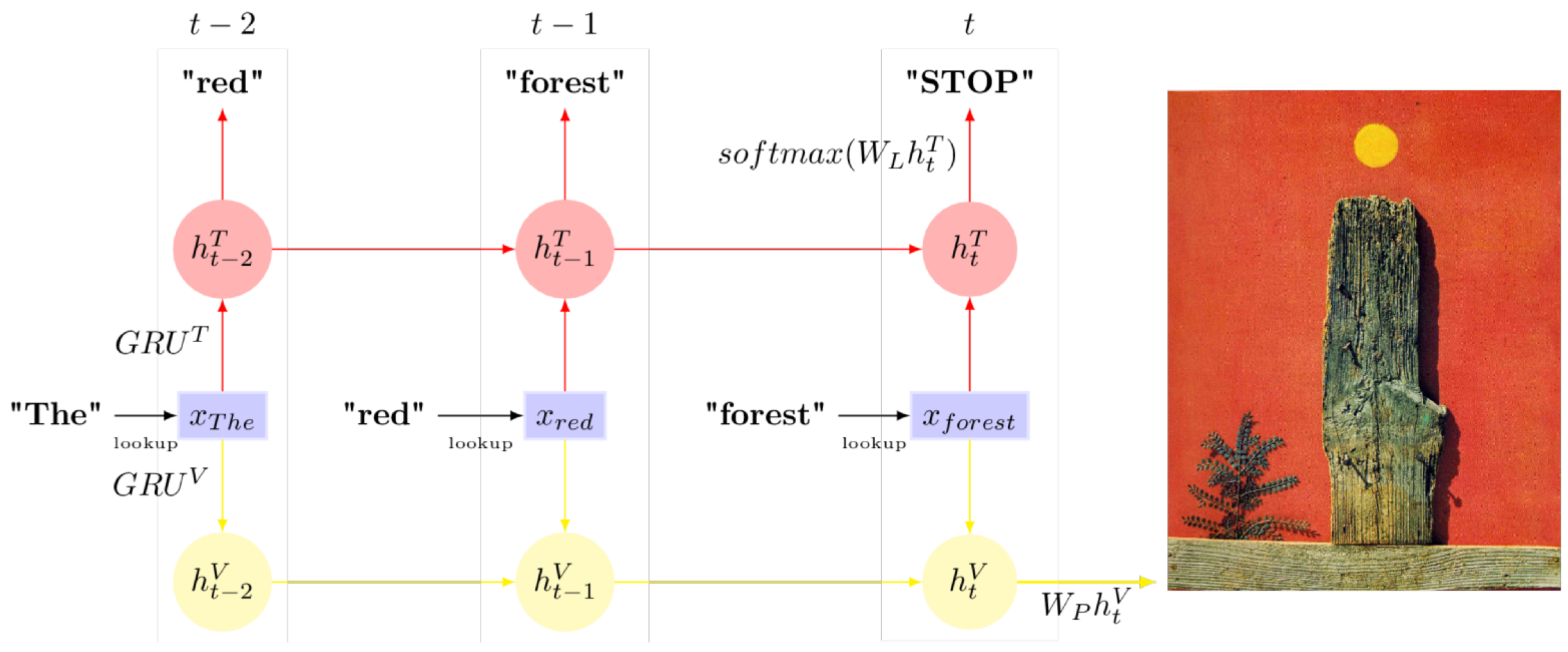
Linguistic Analysis of Multi-modal Recurrent Neural Networks

Ákos Kádár, Afra Alishahi, Grzegorz Chrupala

Tilburg University

a.kadar@uvt.nl, a.alishahi@uvt.nl, g.chrupala@uvt.nl

Learning grounded representations from textual-visual data with RNNs

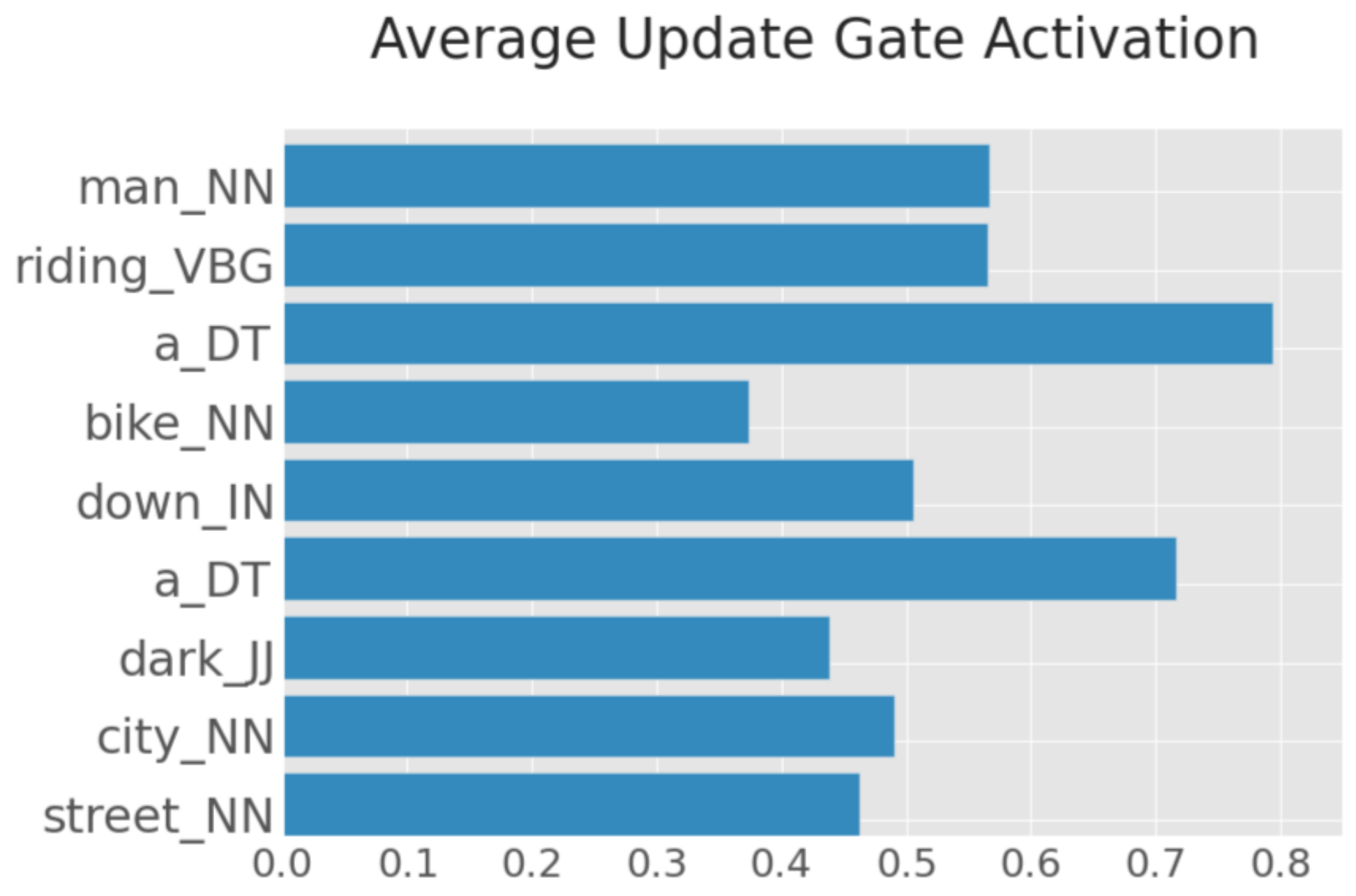


IMAGINET: Two Gated Recurrent Neural Network pathways with shared word-embeddings.
Inputs: Pairs of captions and their corresponding images.

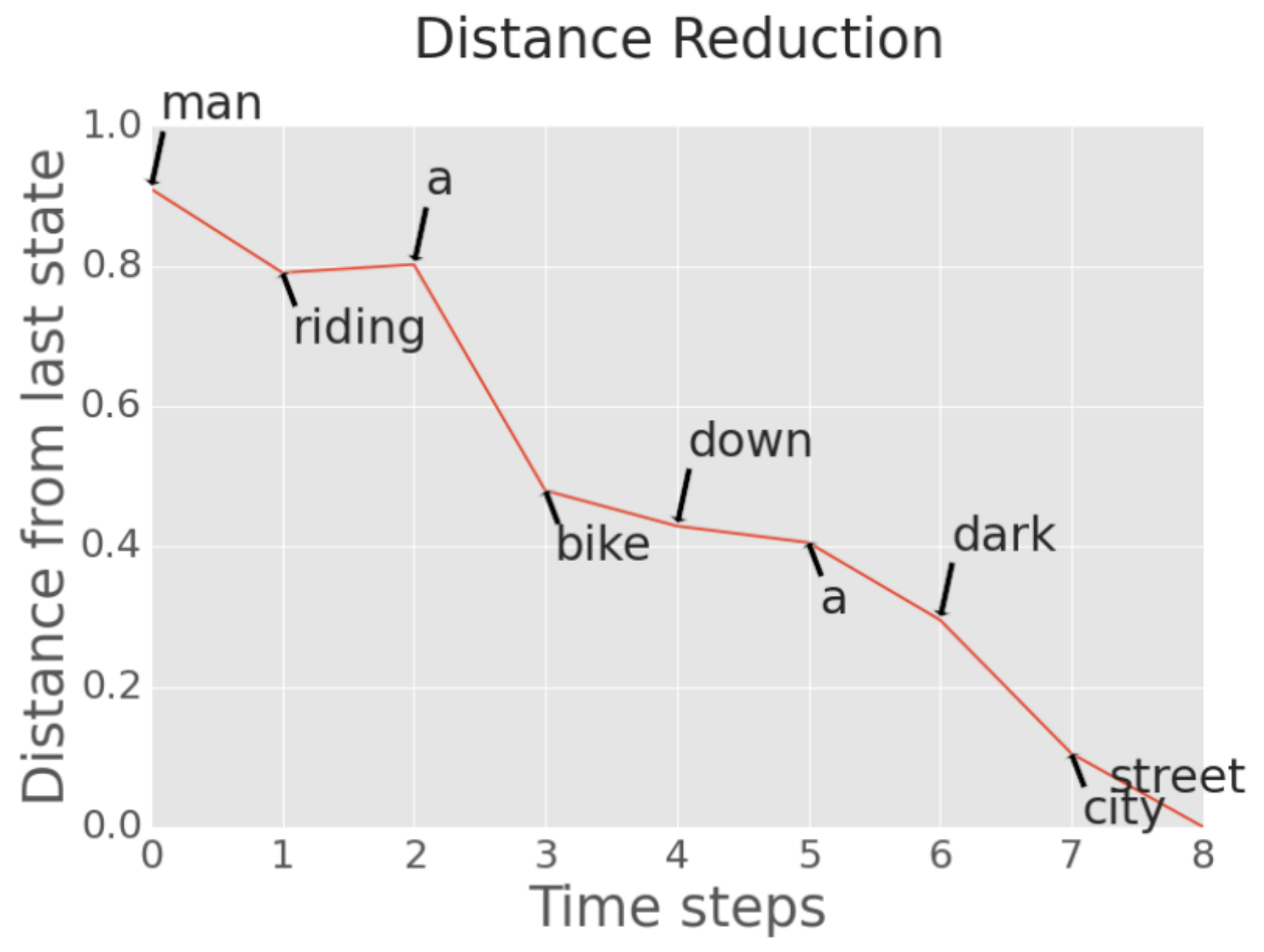
TEXTUAL: Predicts the next word in the sentence from its current hidden state h_t^T.
VISUAL: Predicts the image vector from its last hidden representation h_{full}^V.
Multi-task objective: Cross-entropy loss for the word predictions, Mean Squared Error for image prediction.

Linguistic Analysis of RNN activation patterns

Measuring Grammatical Category Importance



Assign to each (word, category) tuple in the sequence the **average activation of the update-gate** - z_{mean} - at that time step. **Higher** the value the more the network prefers the **previous word.**



Contribution of (word, category) tuples as measured by **cosine-distance reduction** - d_{red} - **with respect to the final hidden-state** h_{full}

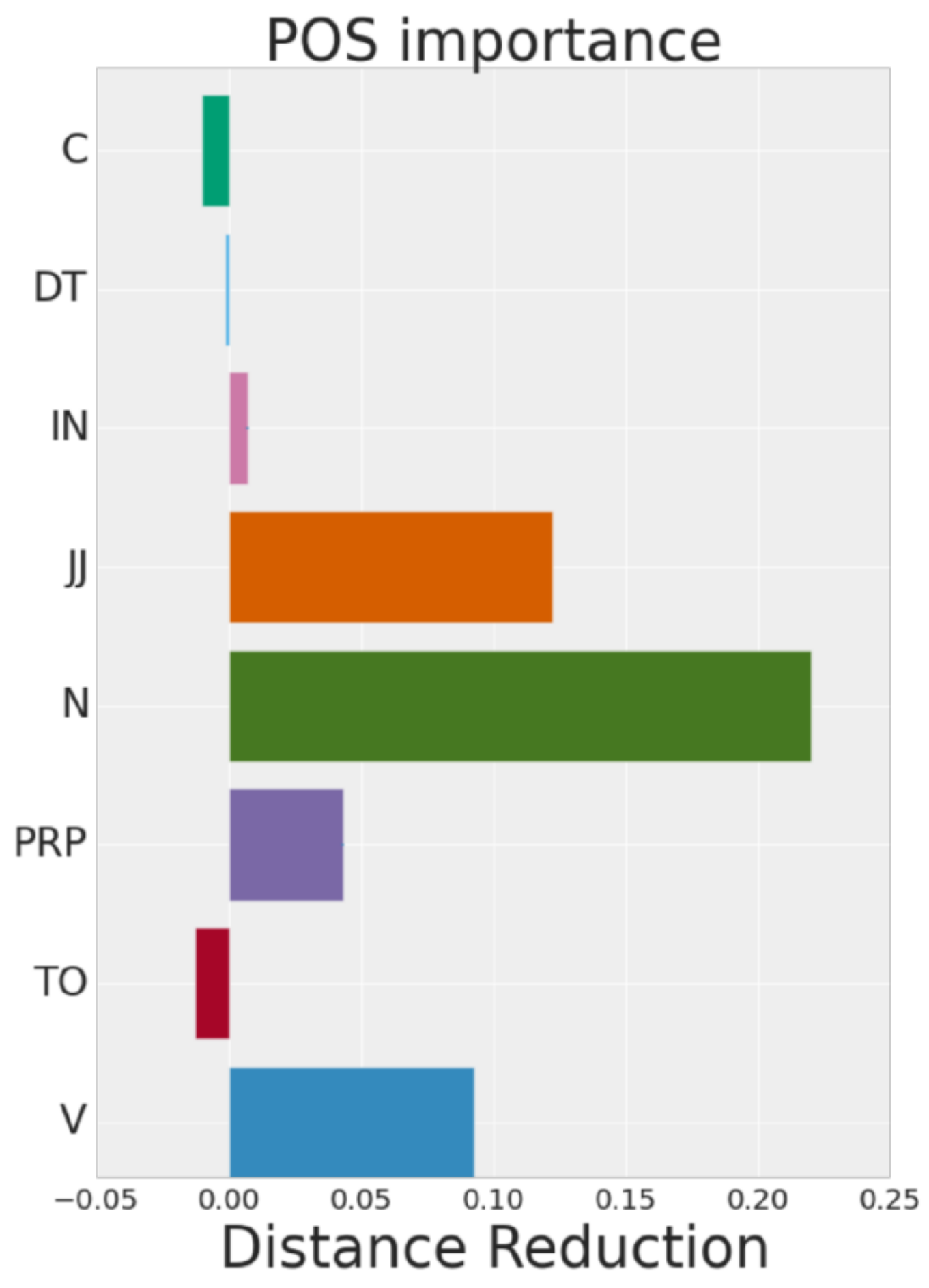
$$d_{\text{red}}^t = d_{\text{red}}^{t-1} - cos(h_t, h_{\text{full}}).$$

Returned Image

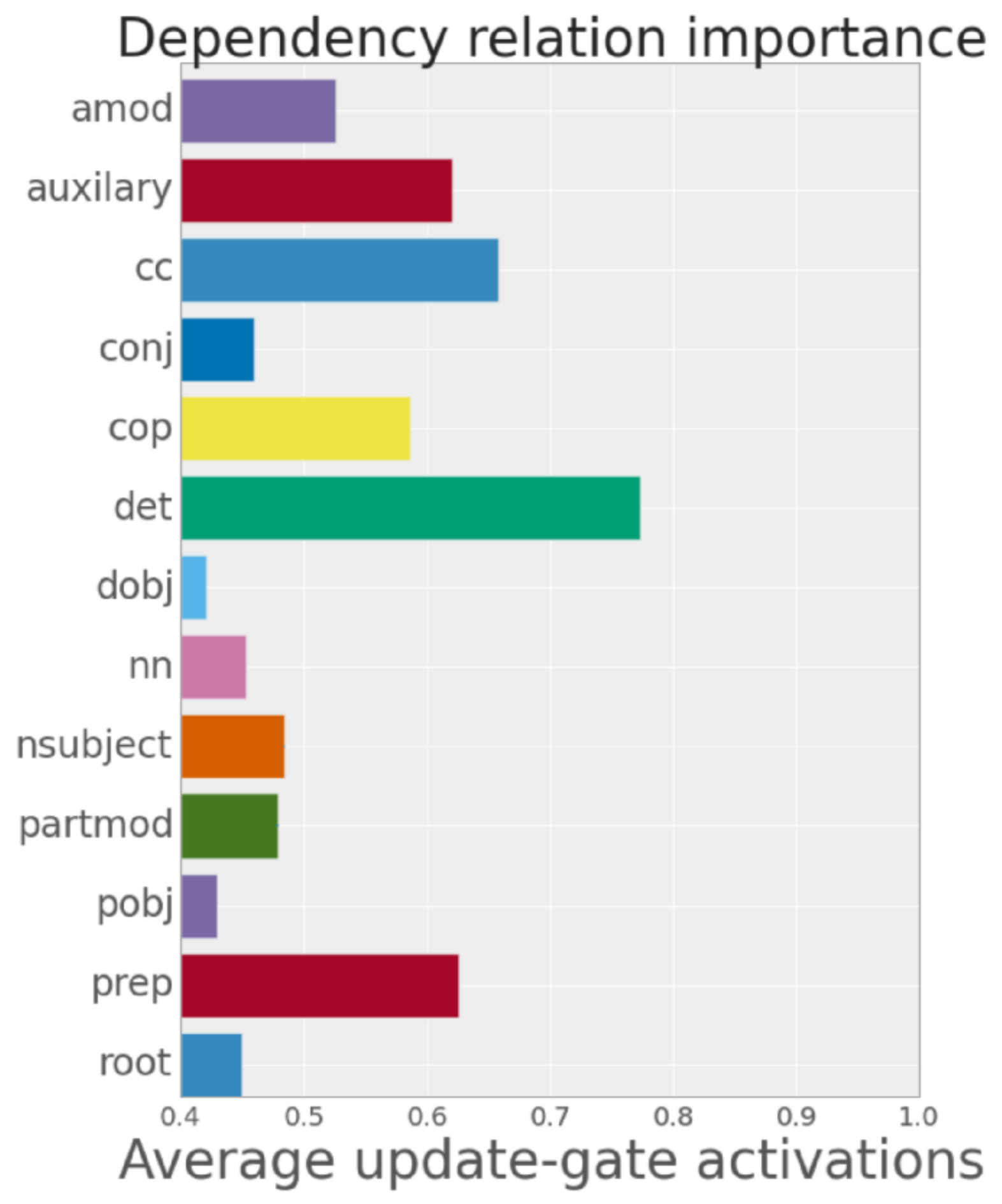
We collect d_{red} **and** z_{mean} statistics for every position in the **captions from the validation portion of MSCOCO** to analyze the **importance of both POS and DepRel categories** that appear at least 500 times.

Results

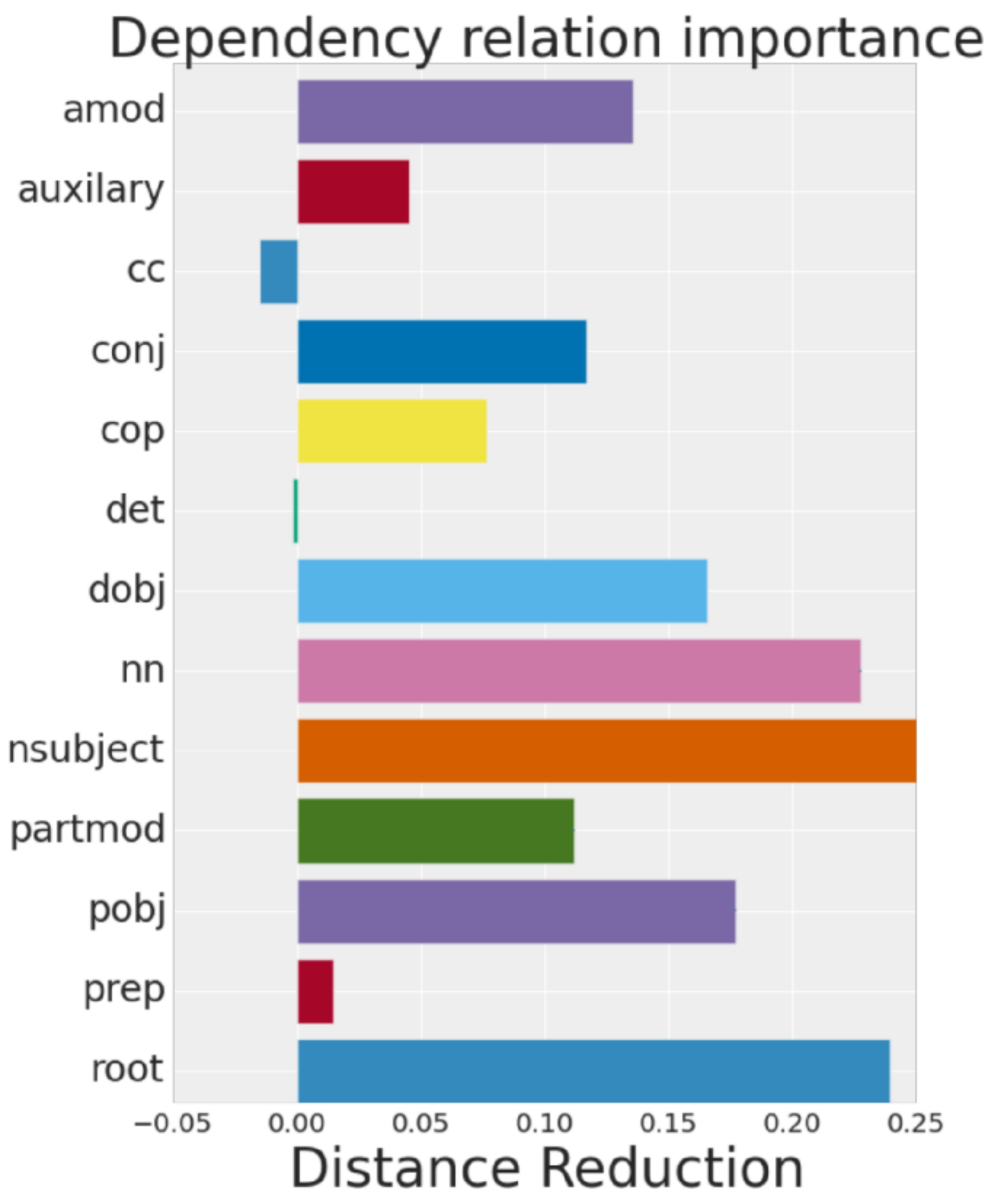
Interpretable activation patterns: high attention for content words low attention for stopwords



Highest d_{red}**:** nouns, adjectives, verbs, prepositions ⇨ largest contribution to sentence representations.
Lowest d_{red}**:** determiners and conjunctions ⇨ least contribution to the meaning representations.



Lowest z_{mean}**:** roots, adjectival modifiers, direct objects, noun compound modifiers, noun subjects, conjuncts and objects of prepositions ⇒ more attention to these categories.
Highest z_{mean}**:** determiners, coordinations, prepositions and auxiliaries. ⇨ least attention



d_{red} **scores for DepRels are in line with** z_{mean} **scores**; ⇨ most important categories: nsubj, nn, amod, pobj and dobj.

References

1. **Grzegorz Chrupała, Ákos Kádár, Afra Alishahi. 2015. Learning language through pictures. ACL.**
2. **Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2015. Visualizing and understanding neural models in NLP. arXiv preprint, arXiv:1506.01066**
3. **Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014, pages 740–755. Springer**