Self-Attention for Abstract Visual Reasoning

Lukas Hahne

The transformer, an attention-based neural network, was developed to improve machine translation and transduction models [8]. Recently, the transformer has conquered VISUAL QUESTION ANSWERING (VQA) challenges in models like VisualBERT [5]. Furthermore, the embodied self-attention mechanism is of special interest in relational reasoning and symbolic artificial intelligence [3]. We propose a hybrid network, WReN-Transformer, grounded on self-attention and components of the base line WILD RELATION NETWORK (WReN) [1]. It learns abstract relations significantly faster and more accurate on Raven's Progressive Matrices [2] inspired PGM dataset [1] than the predominant base line model if structural symbolic auxiliary data is considered during training. We use visual feature embeddings in sequences which comprise context and answers of the PGM dataset. Generalisation experiments are conducted and our model is additionally trained with gradually increasing training dataset sizes to analyse effective relational reasoning performance. Grad-CAM [7] and the transformer's self-attention distributions are used to make qualitative assumptions of relational reasoning and interpretability in PG-Matrices. The WReN-Transformer network excels the WReN model by 11.28 ppt. Relational concepts between objects are efficiently learned demanding only 35% of the $1.2 \cdot 10^6$ training samples to surpass reported accuracy of the base line model while performing 0.54 ppt worse than the best performing WReN-Transformer model. Generalisation reveals similar behaviour than WReN if trained with auxiliary data, generalising slightly better by a difference of 0.56 ppt. Our model fails entirely if this information is not considered. Relation networks [6] which were successfully embodied in WReN [1] and in other networks to solve VQA tasks like CLEVR [4] implement reasoning by stacking feature vectors before processing them in an multi-layer perceptron. The WReN-Transformer represents an alternative on learning abstract relations which emphasises self-attention distributions while incorporating low model complexity, efficient and accurate learning. Furthermore, symbolic structural auxiliary data is crucial for learning where our model does not learn anything if this data type is omitted. Our findings motivate the use of transformer models for more reasoning tasks in the near future.

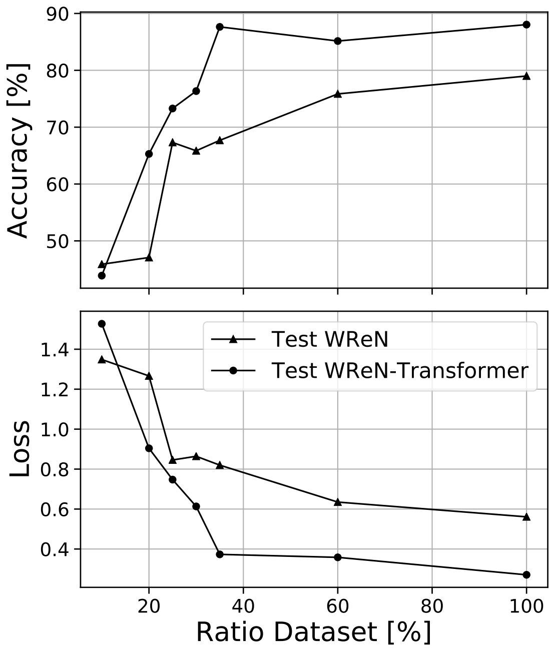

Figure 1: *Sample efficient learning of the base line model WReN and our proposed WReN-Transformer.*

Bibliography

[1] D. G. Barrett, F. Hill, A. Santoro, A. S. Morcos, and T. Lillicrap. Measuring abstract reasoning in neural networks. *arXiv preprint arXiv:1807.04225*, 2018.

[2] W. B. Bilker, J. A. Hansen, C. M. Brensinger, J. Richard, R. E. Gur, and R. C. Gur. Development of abbreviated nine-item forms of the ravens standard progressive matrices test. *Assessment*, 19(3):354–369, 2012.

[3] M. Garnelo and M. Shanahan. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019.

[4] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. *CoRR*, abs/1612.06890, 2016.

[5] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[6] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems 30*, pages 4967–4976. Curran Associates, Inc., 2017.

[7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.