COMP2550 Assignment3

Xingjian Leng(u7136359), Xuning Tan(u), Zhe Chen(u)

May 2021

Q2

For current Visual-Linguistic problems, the most widely used method is deep learning. Neural networks are the core of deep learning and have been applied to many Computer Vision and Natural Language Processing models. They have shown their capability to model complex patterns and prediction problems.

In machine learning models, hyperparameters could notably affect the performance. Therefore, the choice of hyperparameters is essential in improving performance. However, there are a large number of hyperparameters in neural networks. The neural network architecture is one of the hyperparameters (i.e., how layers could be connected to build the whole network). We are motivated by applying MMNAS [1] to MCAN [2] to search for the optimal combination of attention modules. Thus, we decided to use Neural Architecture Search (NAS) in searching for the optimal transformer architecture in OSCAR [3] to further improve its performance.

In terms of the transformer layers in the current OSCAR [3] model. They applied traditional universal transformers [4] after the token extraction step. This model is still the state-of-the-art model in the VQA field even though no advanced transformer is applied to it. It motivated our group because by using neural architecture search, we may find a superior transformer architecture especially for OSCAR. A new OSCAR model with higher accuracy or with fewer parameters may emerge.

References

[1] Z. Yu, Y. Cui, J. Yu, M. Wang, D. Tao, and Q. Tian, "Deep multimodal neural architecture search," *CoRR*, vol. abs/2004.12070, 2020. [Online]. Available: https://arxiv.org/abs/2004.12070

[2] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," *CoRR*, vol. abs/1906.10770, 2019. [Online]. Available: http://arxiv.org/abs/1906.10770

[3] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," *CoRR*, vol. abs/2004.06165, 2020. [Online]. Available: https://arxiv.org/abs/2004.06165

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762