COMP2550 Assignment3

Xingjian Leng(u7136359), Xuning Tan(u6792826), Zhe Chen(u6484911)

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.

Q3 Overall summary

Our group is going to apply Neural Architecture Search(NAS) vision-language tasks. We will mainly focus on searching for better substitute transformer parameters and architectures of the OSCAR [3] model, hence we will divide our research topic's summary of state of the art into two parts, one is for original transformer in OSCAR [3] model, another is for the most advanced NAS architecture, MMNAS.

OSCAR

As the title of the original paper infers, OSCAR [3] stands for Object-Semantics Aligned Pre-training for Vision-Language Tasks, it is state of the art on vision-language tasks. Vision-language pre-training models before OSCAR [3] mostly connect image and text features together as input and use them with self-attention mechanism to learn the semantic information in brutal force. Such process is normally done with multi-layer Transformers. Since extracting semantic alignments is basically like a headless chicken, essential information between visual and text regions might be ignored.

Previous methods majorly apply multi-layer self-attention Transformers to learn cross-modal background representations, with singular embedding of each modality, which makes result of VLP tasks mostly affected by the input singular embedding. Nevertheless, VLP is naturally a weakly-supervised learning task since the explicit information between text and images are always not well labeled. On the other hand, visual features extracted with Fast R-CNN object detectors are usually over-sampled, which result in ambiguities for the extracted visual embedding. To improve this situation, OSCAR [3] innovatively introduces `anchor point` to help the model learn semantic alignments between images and texts. `anchor point` is like an object tags detected in images. They treat the training samples as a combination of word, images and `anchor point`.

MMnas

The deep learning community is taking a transition from human-designed neural architecture to automatically self-designed neural architecture, known as AutoML. Neural Architecture Search is a subset of AutoML, it majorly focuses on automated Neural Network architecture selection and creation, past year have seen has a great number of successful applications of NAS. A NAS procedure can be divided into three components, (1) Search Space, (2)Candidate Evaluation Method and (3)Optimization Method. Search defines the potential network that can be examined to produce the final desired Neural Network. The candidate evaluation method is for comparing the intermediate result and help choose various options among the search space. The optimization method defines how to actually explor the search space, which is essential to the search efficiency and effectiveness of the result architecture.

MMnas [1] is a generalized deep multi-modal neural architecture search framwork for multi-model learning task, the underlying thought is based on BERT model from the natural language processing (NLP), however it is more efficient as it does not require as much data compared to BERT, the huge computing power required by BERT hinders its application in practical situations. Inspired by MCAN model, it firstly searches a set of primitive operations, including feed-forwardnetwork (FFN), self-attention (SA), relation self-attention (RSA) and guided-attention (GA) as the basic unit, they used a unified encoder-decoder backbone through directly sending features into the encoder and decoder. They also designed task-specific heads to different visual linguistic tasks, such as visual question answering and image-text matching(ITM).

As a result, With the standard visual features, MMnas [1] achieves an outstanding improvements on existing hand-crafted models accross different datasets. When applied powerful visual features, MMnas achieved state of the art performance accross all datasets. Thanks to task-specific heads and unified encoder-decoder, MMnas [1] has the ability to automatically learn the optimal architectures of different tasks.

Q3 OSCAR summary by Xingjian Leng

Part of our research topic is related to VQA, the current state-of-the-art VQA model is "Object-Semantics Aligned Pre-training for Vision-Language Tasks" (OSCAR) [3].

In OSCAR [3], they firstly extract visual and semantic features from input images and sentences. Then, they introduced a new idea using "Object-tag", motivated by the fact that salient image features would also appear in sentences as pairs. Object tags are derived by aligning text tokens

on the anchor points generated from the Faster R-CNN. They contain information about the feature alignment between two modalities.

Next, visual features, object tags and text tokens are combined as triples and send to the multi-layer transformers. Transformers are widely used in models with multi-modal features. For transformer [4] in each layer, it is consist of a multi-head attention layer and a feed-forward network. In the multi-head attention layer, the output is calculated by multiplying the input Value matrix with the attention probability. The attention probability shows how the Query matrix is related to the Key matrix. The purpose of the multi-head attention layer is to calculate the weighted sum of encoded outputs, namely the context. Then, the output from the attention layer is sent to the feed-forward neural network to generate the output.

One advantage to use transformers is that they can be pre-trained with certain visual-linguistic tasks. For pre-training, one common method is to mask some tokens and train the neural network to figure out the mask tokens. After pre-training, the model can be trained with a specific downstream task. Thus, models using transformers could usually tackle more than one type of visual-linguistic tasks. It is suggested in [5] that pre-training could boost the performance of models.

The pre-training process of OSCAR [3] is in self-supervised manner. Input triples can be viewed in two perspectives: 1) Dictionary View, 2) Modality View. In terms of the Dictionary View, word tokens and object tags are combined and considered as language part. The image tokens are treated as the image part. Each language token has 15% of probability to be masked. Other tokens are used to predict the masked tokens. The Masked Token Loss is used as the loss function for Dictionary View pre-training. For the Modality View, object tags and image tokens are considered as image part. Object tags have 50% probability to be polluted with a different tag randomly selected from the dataset. Other tokens are used to recover the polluted tags. The loss function for Modality View pre-training is Contrastive Loss. The objective during the OSCAR model pre-training is to minimize the sum of the Masked Token Loss and Contrastive Loss.

Q4

During the teaching break, our group focused on applying light-weighted backbones to current state-of-the-art VQA models. For example, replace the Faster R-CNN [6] for image feature extraction with MobileNet [7]. MobileNet [7] is one type of light-weighted network. It requires fewer parameters for training. Thus, it could be easily applied to embedded systems and deployed. In the first tutorial after the teaching break, we discussed our thoughts with our tutor. However, the tutor mentioned that applying light-weighted networks to current state-of-the-art models will definitely speed up the training process. It cannot be considered as innovation. Therefore, our group decided to change the research topic.

Our group then read certain papers about how to improve the performance of neural networks. We were inspired by the MMnas paper. We finalized our research topic to apply Neural Architecture Search to the current state-of-the-art visual-question answering model.

After deciding our research topic, we read more paper in the following weeks because we lacked the basic knowledge of how neural architecture search works.

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

[5] Y. Li, Y. Pan, T. Yao, J. Chen, and T. Mei, "Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network," *CoRR*, vol. abs/2101.11562, 2021. [Online]. Available: https://arxiv.org/abs/2101.11562

[6] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *CoRR*, vol. abs/1506.01497, 2015. [Online]. Available: http://arxiv.org/abs/1506.01497

[7] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *CoRR*, vol. abs/1704.04861, 2017. [Online]. Available: http://arxiv.org/abs/1704.04861