Research Progress of the Multimodal Pre-Training

Ye Wang

School of Mathematics, Ren Min University of China, Beijing 100872, China

Abstract: This paper describes the related research progress in the field of computer vision pre-training, natural language processing pre-training and image-text cross-modal pre-training. In the field of the computer vision, the Convolutional Neural Network is commonly used to extract the features. ImageNet is a pre-training dataset which is highly applied in this field. After the pre-training of such a network, the characteristics of the image retain richer semantic information. This feature is more suitable for complex cross-modal tasks. In the field of the NLP, Bert designed two pre-training tasks: Masked Language Model and Continuous Sentence Prediction. The proposal of Bert is a milestone. For image-text cross-modal pre-training, Google subsequently launched CBT, which improved the work of VedioBERT. Action clustering was cancelled, the network features of action recognition was introduced directly, and then changing the training target from masked visual words prediction to masked visual feature regression. After a series of comparative experiments, it is proved that using only the network structure of single stream model (single transformer structure), image information and language information can be fused earlier and more freely, which will achieve better results.

Keywords: multimodal pre-training; computer vision; NLP; image-text cross-modal pre-training

1. Introduction

Pre-training includes unsupervised pre-training and supervised pre-training. The differences between unsupervised pre-training and supervised pre-training is whether or not the testing data of the two pre-training model is unlabeled. In the process of pre-training, we use data to train the network. In this way, the model can find the local optima in terms of network parameters which are used as initialization parameters.

Deep learning model is not exactly a “black box”. The model of the network layer can be multiplexed. Deep learning is also that experimenters use machine to summarize the data in nature. After pre-training the model, the network layer of the deep learning can analyze and compares some information, which is based on the large amount basic data. If the result is proper, the information can be transferred and applied into the similar tasks.

2. The Research Process of Pre-Training in the Field of Computer Vision

In the field of the computer vision, the Convolutional Neural Network is commonly used. Research has found that the extracted feature from CNN is hierarchical, which the low-level features are generally common, such as the recognitions of the line segment, arc, etc; the high-level features are related to the task. We give an example: the facial-recognition, (shown in figure 1). The low-level features are still the recognitions of the line segment, horn, arc, etc; the middle features are probably the facial features of mouth, eye, ear and nose, which are combined by the low-level features. The high-level features are probably the outline of face.

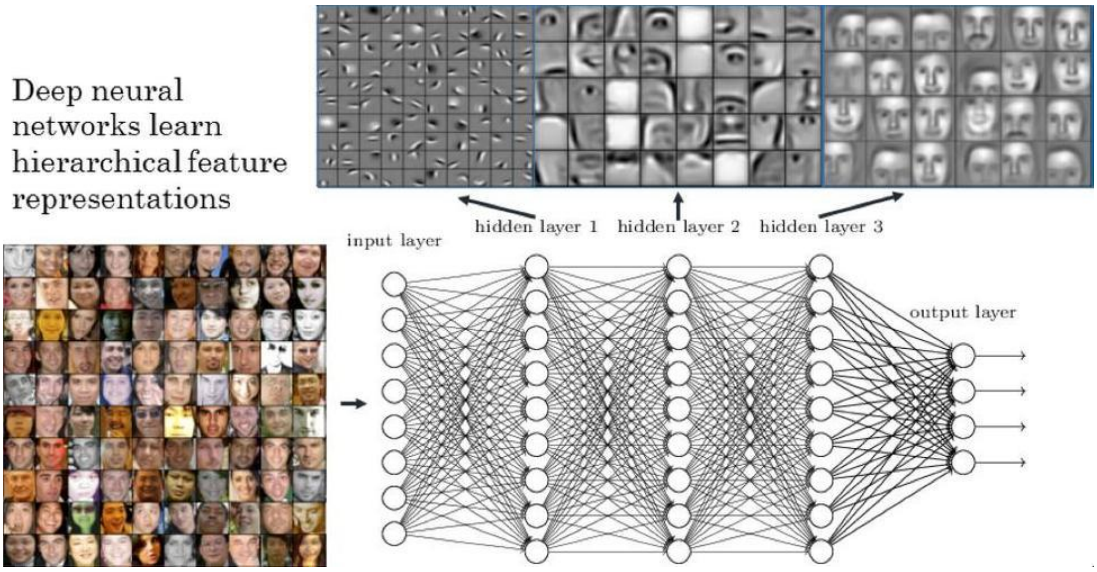

Figure 1. Deep neural networks learn hierarchical feature representations

These details can be used for the final classification, ImageNet [1] is commonly used as a pre-training dataset, which the quantities and sorts of the image are considerably rich. Since the dataset was opened, for example, VGG [2], ReEnet [3] etc has greatly promoted the development of image recognition and even the whole field of computer vision. After training in the datasets of ImageNet, the network structure before its classification layer is retained, which it can be considered that this part has learned how to mine the most differentiated information in the picture. This kind of information can be considered as the characteristics of the image. Such features can be transferred to other data sets for direct classification, meanwhile, it can be transferred to other tasks for the basic framework of tasks, such as target detection (object recognition),semantic segmentation [4], video classification [5], etc, which it can greatly shorten the convergence speed during training.

This is mainly because (1) ImageNet contains a lot of different types of data. These data have a variety of shallow and deep features. The weights trained on ImageNet can extract these features well. (2) There are more or less commonalities between different image data, including those that can be visible or invisible. The weight on ImageNet can extract some common features on the image data set (the data is large enough and the quality is good enough). (3) The features on ImageNet may have different manifestations on other datasets. Some of these features extracted from other datasets with the pre-trained weights on ImageNet make good distinctions.

Most of the image features in cross-modal tasks are extracted by this pre-trained network. But ImageNet is a classified dataset. Its semantic information only contains a category label. When such features are applied to long text matching or complex cross-modal tasks, there is too little semantic information. Cross-modal tasks such as image Q & A usually make up for the lack of information by designing complex modules and attention interaction mechanisms. In this paper, the pre-training data is an image description pair, and the label corresponding to the image includes the description of the whole details. After the pre-training of such a network, the characteristics of the image retain richer semantic information. This feature is more suitable for complex cross-modal tasks.

3. Research Progress of Pre-Training in Natural Language Processing

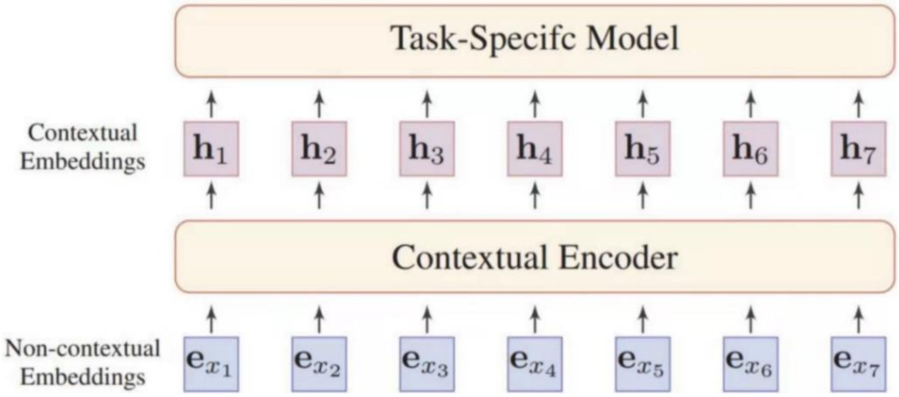
Task-Specifc Model
Contextual Embeddings
$\mathbf{h}_1$ $\mathbf{h}_2$ $\mathbf{h}_3$ $\mathbf{h}_4$ $\mathbf{h}_5$ $\mathbf{h}_6$ $\mathbf{h}_7$
Contextual Encoder
Non-contextual Embeddings
$\mathbf{e}_{x_1}$ $\mathbf{e}_{x_2}$ $\mathbf{e}_{x_3}$ $\mathbf{e}_{x_4}$ $\mathbf{e}_{x_5}$ $\mathbf{e}_{x_6}$ $\mathbf{e}_{x_7}$

Figure 2. Generic Neural Architecture for NLP

As shown in Figure 2, the network framework of general NLP includes non-contextual embeddings and contextual embeddings. The earliest pre-training application in NLP can be traced back to word embedding technology. Its core idea is to obtain the vectorized representation of words through the context of words. There are two methods: predicting the head word through nearby words (CBOW algorithm), predicting nearby words through the head word (Skip-gram algorithm), the network learns the context semantic representation of each word after a large number of text-trainings. The matrix mapped to each word can be used as the word vector of each word, which is also known as the static word vector. It can be found that such words like man-woman = King Queen, which can be reused directly in downstream tasks, There is no further need to train on follow-up tasks. In addition, there are glove and other word embedding technologies, but the biggest disadvantage of this kind of word embedding technology is that it can not solve the phenomenon of polysemy, it simply learn "word co-occurrence frequency" - each word has only a unique word vector, For example, the Apple features can not distinguish between enterprise and fruit. This method can not understand higher-level language concepts, such as syntactic structure, only relationship and so on.

In recent years, dynamic word vectors corresponding to the former have emerged, which focus on learning context. In order to solve the problem of polysemy, Elmo first designed a network based on bidirectional LSTM for pre-training. As shown in Figure 3, it also adopts a two-stage model. In the first stage, an Elmo is trained using a large general corpus, and fine tune is made on the task corpus. The fine tune model can become a Transfer Learning Model, and then the Transfer Learning Model is directly added to the follow-up tasks, using Elmo's word embedding to train the tasks. This representation covers the context of words, which can be used in subsequent Q & A systems, machine translation and other tasks. This kind of model designs various language tasks for training. This kind of language tasks can be extended to a wider range of applications, which can be called pre-training language model.

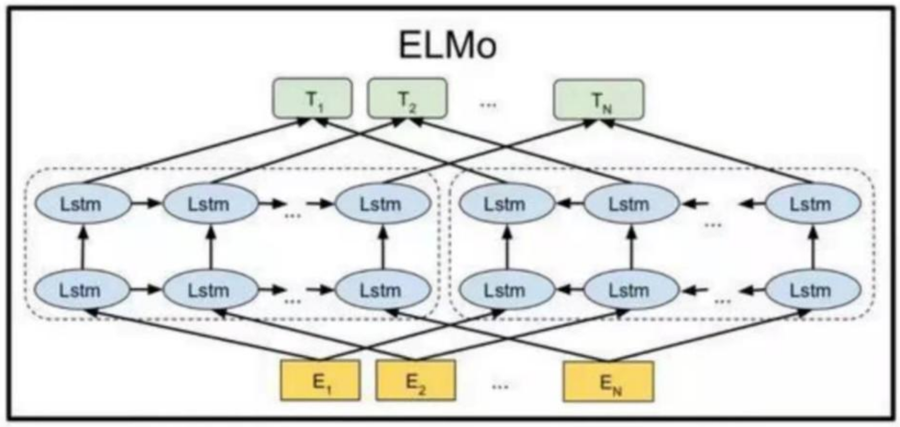

Figure 3. ELMo network framework

18 years ago, most language models were modeled through RNN series networks. The biggest disadvantage of this model is that it cannot be trained in parallel, which the cost of the training and reasoning is more expensive, and the network is more difficult to deepen. Transformer [6] only uses self-attention mechanism to build language model. It improves the speed of training and reasoning, and the model can deepen and learn deeper semantic connections. Since then, almost all the pre- training models are based on transformer. For example, the former is a one-way transformer and the latter is a two-way transformer. The structure can be seen in Figure 4. The GPT unidirectional language model does not make good use of bidirectional context information. ELMO model is a pseudo bidirectional network. Bert designed two pre-training tasks: Masked Language Model and Continuous Sentence Prediction. Similarly, Bert also adopts pre training- fine tuning, two-stage training mode, and its network parameters is kept continually learning during downstream tasks. The proposal of Bert is a milestone. Its pre-training model has swept 11 NLP basic tasks and also has led to the upsurge of pre-training work in the whole NLP academic community. For example, XLNet, RoBERTa, UniLM and other pre training networks have improved the performance of Bert with varying degrees from the perspectives of attention calculation, pre-training data, pre-training tasks and training process, it also greatly improves the effect of multiple NLP tasks.

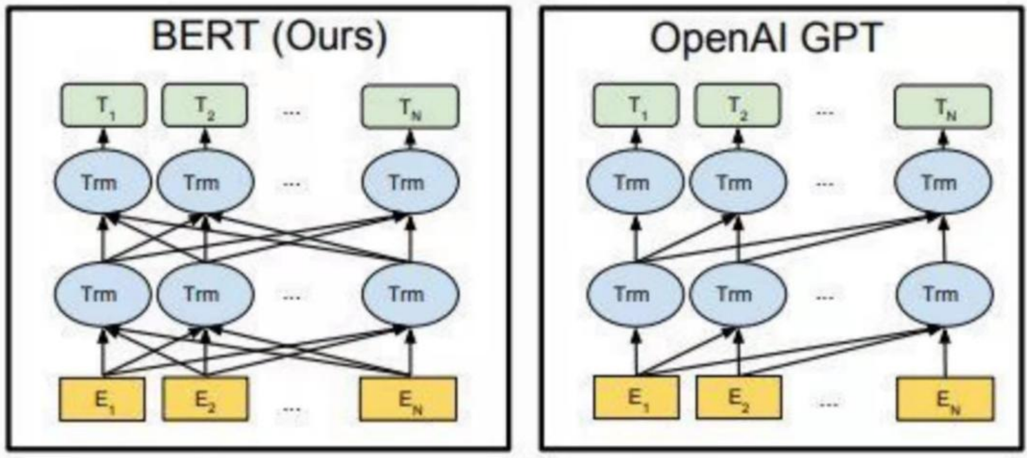

Figure 4. BERT & GPT network framework

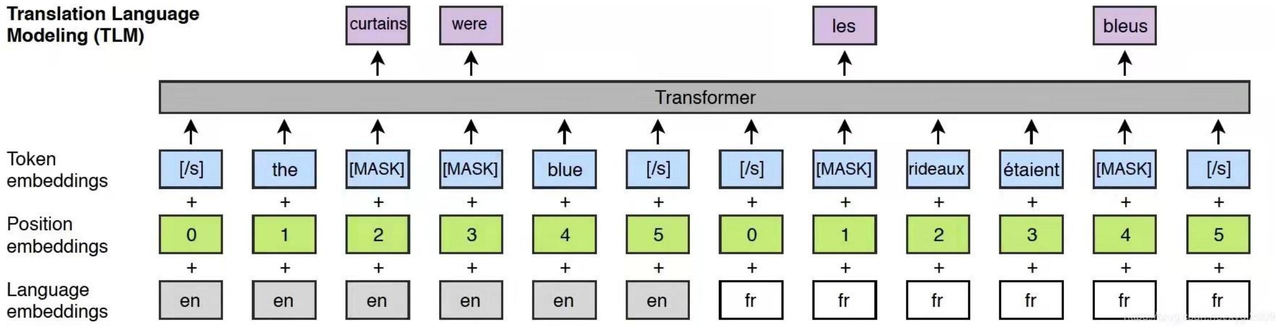

Figure 5. XLM network framework & TLM pre-training task

In addition to the above language model, Facebook's XLM has just studied TLM, The network structure is shown in Figure 5. Masked language Modeling is introduced in unidirectional language model, continuous sentence prediction is introduced win parallel corpus, which is looking forward to learning its potential cross language translation alignment information. This paper proves that XLM has made great progress in multilingual tasks, unsupervised machine translation and supervised machine translation. Its strength can be seen that the pre-training network is based on transformer architecture. The work of this paper is close to the TLM task in XLM. We try to introduce a similar network architecture into cross modal pre-training. In a sense, image and text input is equivalent to cross language Parallel Corpus in machine translation. As long as we construct an appropriate supervision task, we can complete the training goal.

4. Research Progress of Image-Text Cross-Modal Pre-Training

Although the pre-training work have appeared one after another in the fields of NLP and CV, there is always a lack of combination of the NLP and CV in cross modal pre-training. The first pre training work of unifying visual information and language is vediobert [7] launched by Google in the field of video. Because video itself has timing information, similar to language, the information in video also corresponds to its learning screen. Vediobert can use video to construct a self-monitoring system and learn better representation from unmarked video. However, the video features are not extracted by the method of convolutional neural network, instead, we use the action recognition technology to extract actions from the video in advance, and all action words are clustered and mapped to an action dictionary based on the extracted action features. Then, all video clips are encoded with dusterid to form visual words. Then mask the video actions and predict the missing action words (cluster number) in a similar way to Bert. It is still a language model in essence, and does not support the direct coding of video features. Moreover, if all homogeneous video actions are unified into one word, the rich detail information of the video will be greatly lost.

Google subsequently launched CBT [8], which improved the work of vediobert action clustering, was cancelled, the network features of action recognition was introduced directly, and changing the training target from masked visual words prediction to masked visual feature regression.

Similar to the same period work in this paper are ViLBERT [9] and LXMERT. Both models are based on multi-layer transformer. The former uses image description for pre-training, and the latter uses image Q&A data for pre-training. However, the biggest difference between them and this paper is that the two stream models (two independent transformers encode the two modes respectively) are used in both works, and then the cross modal transformer is introduced in the upper layer for information fusion. This paper will prove that using only the network structure of single stream model (single transformer structure), image information and language information can be fused earlier and more freely, which will achieve better results than the first two, and the amount of parameters is much less than that of two-stream model.

References

[1] Deng J, Dong W, Socher R, et al. Image Net: A large-scale hierarchical image database. Computer vision and pattern recognition, 2009: 248-255.

[2] Simonyan K, Zisserman A. VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. Computer vision and pattern recognition, 2014.

[3] He K, Zhang X, Ren S, et al, Deep Residual Learning for Image Recognition. Computer vision and pattern recognition, 2016: 770-778.

[4] Hek, Gkioxari G, Dollar P, et al. Maskr-cnn //Proceedings of the IEEE international conference on computer vision.2017:2961-2969

[5] Abu-El-Haija S, Kothari N, Lee J, et al. Youtube-8m: A large-scale video classification benchmark. arXivpreprintarXiv:1609.08675,2016

[6] Vaswani A, Shazeer N, Parniar N, et al. Attention is All you Need. Neural information processing systems, 2017: 5998-6008.

[7] Sun C, Myers A, Vondrick C, et al. Video BERT: A Joint Model for Video and Language Representation Learning. International conference on computer vision, 2019: 7464-7473.

[8] Sun C, Baradel F, Murphy K, et al. Contrastive bidirectional transformer for temporal representation learning. arXivPreprintarXiv:1906.05743, 2019.

[9] Lu J, Batra D, Parikh D, et al. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. Neural information processing systems, 2019: 13-23.