Explore the potential of Transformer in the field of computer vision

Binhao Yan, Xikai Li, Xiaoguang Zhao

September 2021

Introduction

Transformer is to solve sequence-to-sequence tasks while we are working with texts. "*The Transformer is the first transduction model relying on entirely on self-attention to compute representations of its input and output without using sequence-aligned RNNs or convolution.*"

Google created and published in 2018 for nature language process a powerful model, which was BERT, the first transformer-based model. With this novel architecture proposed in the paper *Attention Is All You Need*, BERT rapidly reached top position in almost all the table rank in NLP. After that, in the field of NLP, people started to use BERT or any other model based on transformer or variant of transformer (such as XLNet or Reberta and so on) to resolve NLP problems.

Nowadays, with the fantastic success of family of Transformer in NLP, many researchers come to bring this SOTA architecture into the field of Computer Vision to solve some real tasks. Surprisingly, transformer did another good job. For example, Geoffrey Hinton, used Transformer as the Encoder-Decoder in his latest great work PIX2SEQ.

Problems

Although the Transformer can do something well in computer vision, that does not mean it can be as dominant as what it did in the field of NLP. Convolution network still has its advantages in inductive bias, better generation, faster convergence.

So, what if we combine convolution layer with Transformer and how about using them for multimodal classification of products with picture and text descriptions? Taking joint efforts to forge new partnerships of win-win cooperation and resolving real tasks better, that is the story we are going to talk about.

Preliminary plan

Hope we can get clear about how to effectively combine convolution layer with Transformer and implement the algorithm by at the latest 1st November.

Then we would have 1 month training the algorithm so that wo could submit the trained model before project deadline.

Reference

[Submitted on 12 Jun 2017 (v1), last revised 6 Dec 2017 (this version, v5)]
Attention Is All You Need
Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin
arXiv:1706.03762 [cs.CL]

Ciresan, Dan; Ueli Meier; Jonathan Masci; Luca M. Gambardella; Jurgen Schmidhuber (2011). *"Flexible, High Performance Convolutional Neural Networks for Image Classification*". Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence-Volume Volume Two. 2: 1237–1242. Retrieved 17 November 2013.

[Submitted on 22 Sep 2021]
Pix2seq: A Language Modeling Framework for Object Detection
Ting Chen, Saurabh Saxena, Lala Li, David J. Fleet, Geoffrey Hinton
arXiv:2109.10852 [cs.CV]

[Submitted on 9 Jun 2021 (v1), last revised 15 Sep 2021 (this version, v2)]
CoAtNet: Marrying Convolution and Attention for All Data Sizes
Zihang Dai, Hanxiao Liu, Quoc V. Le, Mingxing Tan
arXiv:2106.04803 [cs.CV]

Liu, Zhun, Shen, Ying, et al. "Efficient Low-rank Multimodal Fusion With Modality-Specific Factors" Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Vol. 2018. NIH Public Access, 2018.

Tsai, Yao-Hung Hubert, et al. "Multimodal transformer for unaligned multimodal language sequences." Proceedings of the conference. Association for Computational Linguistics. Meeting. Vol. 2019. NIH Public Access, 2019.