My studies about Transformers

M. Mahdi Farahbakhsh

June 2, 2024

In the Name of Allah

Abstract

Your abstract text goes here.

1 Introduction

Introduce your topic and the purpose of the report here.

2 Background

Provide any background information that your reader will need.

3 Multimodal Learning With Transformers:A Survey [6]

Note that this survey will not discuss the multimodal papers where Transformers is used simply as the feature extractor without multimodal designs.

1. Introduction
   - We suggest that self-attention be treated as a graph style models th input sequences(both uni-modal and multi-modal) as a fully connected graph.
2. Background
   - Derivatives of *Vanilla* Transformer :
     * BERT [4]
     * BART [5]
     * GPT [3]
     * Long-former [1]
     * Transformer-XL [2]
     * XLNet [7]
   - Transformers in different Domains:
     * in NLP domains: Dominated
     * in visual domains: general pipeline is "CNN features + Strandard Transformer Encoder"
     * multimodal tasks :
       + VideoBERT : the first
       + CLIP : new milestone
         @ IDK: uses multimodal pretraining to convert classification as retrieval task that enables the pretrained modals to tackles zero-shot recognition.
   - Multimodal Big Data
     * Data scales are larger : recently released datasets are million scales
     * More modalities: vision, text, audio
       + Pono : audio-visual question answering
     * More Application & Scenarios
     * Tasks are more difficult
     * Instructional Videos
       @ IDK: Transformers are data hungry, Therefore ,their high -capasity modals and multimodal Big Data basis co-created the prosperity of the Transformer based multimodal machine learning.
       + VideoBERT : the first
       + CLIP : new milestone
         @ IDK: uses multimodal pretraining to convert classification as retrieval task that enables the pretrained modals to tackles zero-shot recognition.

3. advantages

    * more general space

        + Vanilla transformers (self attention) can model any given tokenized input from any model.

            ¿ compare with CNN: CNN is restricted in the aligned grid spaces/metrics

4. Vanilla Transformers

    @ IDK: "position-wise" Fully-connected Feed Forward (FFN)

    - To help the back propagation of the gradient, both MHSA and FFN use Residual Connection(any mapping $f(.)$ is defined as $x \longleftarrow f(x) + x$)

    - $Z \longleftarrow N(sublayer(Z) + Z)$

        + Z: Input tensor — sublayer output
        + $sublayer$: FFN or MHSA
        + Residual Connection is used.
        + N : normalization
            + BN
            + LN
            @ Open Problem: post-normalization vs pre-normalization
                ¿ Vanilla Transformation: post.
                ¿ mathematical perspective: pre. make more sense
                ¿ both theoretical research and experiment validation

4 Results

Present your findings here.

5 Discussion

Discuss the implications of your results.

6 Conclusion

Sum up the report and any final thoughts.

References

[1] I. Beltagy. Longformer: The long-document transformer. 2020.

[2] Z. Dai. Transformer-xl: Attentive language models beyond a fixed-length context. 2019.

[3] A. Radford et al. Improving language understanding by generative pre-training.

[4] J. Devlin et al. Bert: Pretraining of deep bidirectional transformers for language understanding. 2018.

[5] M. Lewis et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. 2019.

[6] Peng Xu. Multimodal learning with transformers: A survey. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 45(10), 2023.

[7] Z. Yang. Xlnet: Generalized autoregressive pretraining for language understanding. 2019.