Navigating LaTeX: A Personal Journey

M. Mahdi Farahbakhsh

June 4, 2024

In the Name of Allah

0.1 Latex

1. Preamble
   - 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 [**?**]
     * BART [**?**]
     * GPT [**?**]
     * Long-former [**?**]
     * Transformer-XL [**?**]
     * XLNet [**?**]
   - 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 multi-modal 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

Bibliography