Computer Science

Query paper:
Title: Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context
Abstract: Transformers have a potential of learning longer-term dependency, but are limited by a fixed-length context in the setting of language modeling. We propose a novel neural architecture Transformer-XL that enables learning dependency beyond a fixed length without disrupting temporal coherence. It consists of a segment-level recurrence mechanism and a novel positional encoding scheme. Our method not only enables capturing longer-term dependency, but also resolves the context fragmentation problem. As a result, Transformer-XL learns dependency that is 80% longer than RNNs and 450% longer than vanilla Transformers, achieves better performance on both short and long sequences, and is up to 1,800+ times faster than vanilla Transformers during evaluation. Notably, we improve the state-of-the-art results of bpc/perplexity to 0.99 on enwiki8, 1.08 on text8, 18.3 on WikiText-103, 21.8 on One Billion Word, and 54.5 on Penn Treebank (without finetuning). When trained only on WikiText-103, Transformer-XL manages to generate reasonably coherent, novel text articles with thousands of tokens. Our code, pretrained models, and hyperparameters are available in both Tensorflow and PyTorch.

Candidate papers:

1. **Title:** Attention is All you Need
   Abstract: The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU.
2. **Title:** Self-attention with relative position representations
   Abstract: Relying entirely on an attention mechanism, the Transformer introduced by Vaswani et al. (2017) achieves state-of-the-art results for machine translation. In contrast to recurrent and convolutional neural networks, it does not explicitly model relative or absolute position information in its structure. Instead, it requires adding representations of absolute positions to its inputs. In this work we present an alternative approach, extending the self-attention mechanism to efficiently consider representations of the relative positions, or distances between sequence elements.
3. **Title:** Character-Level Language Modeling with Deeper Self-Attention
   Abstract: LSTMs and other RNN variants have shown strong performance on character-level language modeling. These models are typically trained using truncated backpropagation through time, and it is common to assume that their success stems from their ability to remember long-term contexts. In this paper, we show that a deep (64-layer) transformer model (Vaswani et al. 2017) with fixed context outperforms RNN variants by a large margin, achieving state of the art on two popular benchmarks: 1.13 bits per character on text8 and 1.06 on enwik8. To get good

results at this depth, we show that it is important to add auxiliary losses, both at intermediate network layers and intermediate sequence positions.

4. **Title:** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
   Abstract: We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications. BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

5. **Title:** Adaptive input representations for neural language modelingC
   Abstract: We introduce adaptive input representations for neural language modeling which extend the adaptive softmax of Grave et al. (2017) to input representations of variable capacity. There are several choices on how to factorize the input and output layers, and whether to model words, characters or sub-word units. We perform a systematic comparison of popular choices for a self-attentional architecture. Our experiments show that models equipped with adaptive embeddings are more than twice as fast to train than the popular character input CNN while having a lower number of parameters. On the WikiText-103 benchmark we achieve 18.7 perplexity, an improvement of 10.5 perplexity compared to the previously best published result and on the Billion Word benchmark, we achieve 23.02 perplexity.

6. **Title:** A Neural Probabilistic Language Model
   Abstract: A goal of statistical language modeling is to learn the joint probability function of sequences of words. This is intrinsically difficult because of the curse of dimensionality: we propose to fight it with its own weapons. In the proposed approach one learns simultaneously (1) a distributed rep(cid:173) resentation for each word (i.e. a similarity between words) along with (2) the probability function for word sequences, expressed with these repre(cid:173) sentations. Generalization is obtained because a sequence of words that has never been seen before gets high probability if it is made of words that are similar to words forming an already seen sentence.

Exemplary analysis:

1. **Relevance:** This paper introduces the Transformer model, which is the foundation upon which Transformer-XL builds. The Transformer model revolutionized natural language processing (NLP) by moving away from recurrent and convolutional networks, focusing instead on attention mechanisms to process sequences of data. The query paper extends the Transformer model to handle longer contexts, which is a direct expansion of the work introduced in this paper.
   Reason for Citation: To acknowledge the foundational model (Transformer) on which

Transformer-XL is based and to discuss the limitations of the original Transformer model that the query paper aims to overcome.

2. **Relevance:** The introduction of relative position representations in self-attention mechanisms is a key innovation that allows Transformers to better understand the relationships between different parts of a sequence. This concept is important for the Transformer-XL, which seeks to improve the model's ability to handle long-term dependencies.

   Reason for Citation: To discuss advancements in self-attention mechanisms that are relevant to the development of Transformer-XL, particularly the handling of position information in sequences, which is crucial for modeling longer contexts.
3. **Relevance:** This paper demonstrates the effectiveness of deep Transformer models for character-level language modeling, highlighting the potential of Transformers to capture complex dependencies in text. It sets a precedent for using Transformers in language modeling, particularly at the character level, and discusses techniques to improve their performance.

   Reason for Citation: To support the argument that deep Transformer models, like Transformer-XL, are effective at language modeling tasks and to build on the idea of enhancing Transformer architectures for better performance in NLP tasks.
4. **Relevance:** BERT represents a significant leap forward in pre-training language representations, using a bidirectional Transformer. While BERT focuses on understanding language context in both directions, Transformer-XL aims to extend the context length that models can effectively process.

   Reason for Citation: To highlight the importance of deep bidirectional Transformers in NLP and to position Transformer-XL within the broader context of recent advancements in Transformer-based models. It may also cite BERT to discuss differences in approach, particularly regarding context length and model architecture.
5. **Relevance:** This paper explores adaptive input representations, which can make models more efficient and effective by adjusting the capacity of input representations based on the complexity of the input. Such techniques are relevant for Transformer-XL, which seeks to improve efficiency and performance in language modeling.

   Reason for Citation: To discuss methods for improving the efficiency of neural language models, particularly in the context of Transformer-based architectures. The query paper might leverage or build upon these adaptive techniques to enhance Transformer-XL's performance.
6. **Relevance:** This work is foundational in the field of neural language modeling, introducing the concept of learning distributed representations for words alongside the probability function for word sequences. It lays the groundwork for subsequent developments in language modeling, including the use of Transformers.

   Reason for Citation: To acknowledge the historical context and evolution of language modeling techniques leading up to the development of Transformer and Transformer-XL models. It may also cite this work to discuss the importance of distributed representations in understanding language.

Exemplary ranking: Ranked order: paper 1, paper 2, paper 3, paper 4, paper 5, paper 6

1. **Explanation:** This paper is the cornerstone of Transformer models. Any research following Transformer-XL would likely reference this to acknowledge the foundational model and its

limitations that the follow-up work seeks to address or build upon.

2. **Explanation:** The methodological relevance of improving self-attention mechanisms, especially for handling longer contexts in Transformer models, makes this paper a critical citation for discussing technical advancements or modifications in a follow-up study.
3. **Explanation:** This paper's focus on deep Transformer models for character-level language modeling aligns closely with the objectives of Transformer-XL, making it a likely citation for discussions on model depth and granularity in language modeling.
4. **Explanation:** Given the significant impact of BERT on the NLP field and its methodological similarities and differences with Transformer-XL, a follow-up study would likely cite it to discuss further advancements or comparisons in Transformer-based model architectures.
5. **Explanation:** Techniques for improving model efficiency and input representation are crucial for advancing Transformer models. A follow-up study might cite this work to explore or introduce new adaptive techniques for enhancing Transformer-XL's efficiency or performance.
6. **Explanation:** While foundational to the field of neural language modeling, this paper might be cited less frequently in a direct follow-up to Transformer-XL, except to provide historical context or discuss the evolution of language modeling techniques leading up to Transformer models.