GPT

Improving Language Understanding by Generative Pre-Training

Ryan Chan

24 July 2023

Outline

Motivation for GPT

- "Natural language understanding" comprises a wider range of diverse tasks: textual entailment, question answering, determining similarity between sentences, document classification, etc.
- Although large unlabelled text corpora are abundant, labelled data for each of these learning tasks is comparatively scarce
  - Most deep learning methods require substantial amounts of manually labelled data - restricts their ability in many domains that suffer from having few sources of annotated data
  - In these situations, we should try to leverage linguistic information from unlabelled data sources (use pretraining)
- A lot of evidence where even where a lot of labelled data is available, learning good representations (via unsupervised means) first can provide significant performance boost - e.g. using pretrained word embeddings as inputs to networks

Motivation for GPT

- Goal: "learn a universal representation that transfers with little adaption to a wider range of tasks"
- Two steps in GPT framework:
  1. Use a causal language modelling objective on a large corpus of unlabelled text to learn initial parameters of a neural network
  2. Adapt/fine-tune these parameters to a target task using the corresponding supervised objective
- Use a Transformer decoder architecture as in Vaswani et al. [2017]

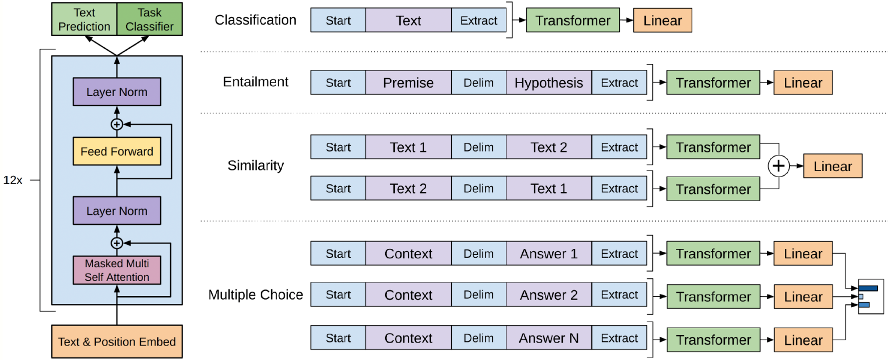
GPT architecture and pretraining setup

- Training data
  - BookCorpus dataset (800M words: 7000 unique unpublished books from a variety of genres) - no longer used for intellectual property reasons
  - Alternative dataset: the 1B Word Benchmark - same one used to train ELMo [Peters et al., 2018] (GPT language model achieved very low token perplexity of 18.4 on this corpus)
- Architecture largely follows the same as the multi-layer unidirectional Transformer Decoder described in Vaswani et al. [2017]
  - $L = 12$ number of layers/transformer blocks, $H = 768$ hidden dimension size, $A = 12$ number of attention heads ($\approx$ 110M parameters) - similar set up to $\text{BERT}_{\text{BASE}}$
  - Context length of 512 tokens

Fine-tuning Transformer decoders

- Aim: develop a network which can be directly fine-tuned to downstream tasks
- Pre-training only dealt with contiguous pieces of text (no special tokens), but for many tasks, modifications are required
  - For tasks dealing with pairs of sentences (e.g. textual entailment, similarity, question answering), concatenate pairs together into one input using a separation token in between them
  - Introduce a special classification token at the end of the sequence
- One key difference from BERT is that the sentence separator and classifier token are only introduced at fine-tuning time

Fine-tuning Transformer decoders



[1]

[1] [Radford et al., 2018, Figure 1]

Summary of results in GPT

- Evaluated on variety of supervised tasks including natural language inference, question answering, semantic similarity, and text classification
  - Some of these tasks are included in the the General Language Understanding Evaluation (GLUE) benchmark [Wang et al., 2018], a collection of diverse natural language understanding tasks
- Train model for 100 epochs (batch size 64)
- Fine-tuning didn't require many epochs - generally 3 was sufficient
- Achieves state-of-the-art performance in 9 out of 12 tasks
- GLUE benchmark score of 72.8 (previous best 68.9, $\mathrm{BERT}_{\mathrm{BASE}}$ and $\mathrm{BERT}_{\mathrm{LARGE}}$ achieved 79.6 and 82.1 respectively [Devlin et al., 2019])

Some extensions

- GPT-*: more-or-less the same architecture, but LARGERRRRRR
  - GPT-2: trained on 40GB of text (GPT-1 trained on about 4.5 GB) with number of model parameters ranging from 117M ("small") to 1542M ("extra-large")
  - GPT-3: trained on 570GB of text with 175B parameters
  - GPT-4: rumours to have over 1T parameters (possibly 1.76T)
- Other alternative decoder models: LLaMA, Bard, ...

References

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving Language Understanding by Generative Pre-Training. Technical report.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.