Paper Notes:Bag of tricks to improve Pretrain-Finetune Paradigm

Tianjian Li

March 2022

1 March 22, 2022

1.1 BART

BART[6] is work by Facebook published at ACL 2020. It aims to generalize BERT [4] and GPT[5]. BERT uses a bidirectional encoder that replaces random token with masks and is trained only on the loss between the masked tokens and the ground truth. GPT is trained in a unidirectional autoregressive way, generating words according to their leftward context.

BART combines the two models by including a denoising autoencoder that encodes the corrupted document then autoregressively maps the encoded document to its original uncorrupted counterpart. The encoder is the standard transformer[3], with a silght modification that changes the activating function from ReLU to GeLU[2]. In each layer of the decoder, the model calculates the cross attention over the final hidden layer of the encoder.

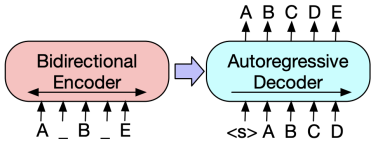

Figure 1: BART

The paper introduces a few corrupting tricks during the pretraining phase:

- Token Masking: Similar to BERT[4].
- Token Deleting : Random deletes input tokens
- Text Infilling : Sample text spans and replace them with a single mask token.

- Sentence Permutation
- Document Rotation:
  A token is chosen uniformly random and the document is rotated so that it begins with this token.

For various downstream tasks:

- Sequence Classification:
  Feed the pretrained encoder and decoder the same input and use let the decoder output a class token, similar to the CLS token in BERT.
- Token Classification(SQuAD):
  Feed the pretrained encoder and decoder the same input and use the final hidden state of the decoder as the representation for each word.
- Sequence Generation:
  The Encoder takes the input and the decoder outputs the generated sequence autoregressively.
- Machine Translation:
  The entire BART model is viewed as an decoder, combined with an additional encoder that takes the source language as input. The model is trained in two steps: 1) Freeze most the parameters of the BART model and only update the parameters of the new encoder, BART positional embeddings and the attention input matrix of the first layer of BART's encoder. 2) Train the entire model for a small number of steps.

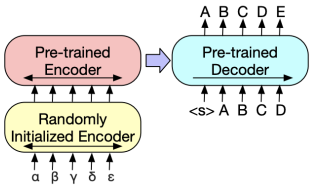

Figure 2: Illustration of Machine Translation Task

1.2 PET

PET[7]:Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference is published at EACL2021. GPT[5] teaches us that a pretrained model with the correct task descriptions can perform better in downstream tasks than the commonly used pretrain-fintune paradigm. PET proposes that cloze style input phrases can help language model understand a given task.

GPT provides hints for our language model to generalize to tasks(Reading Comprehension, Question Answering)in a zero-shot context. In contrast, PET provides task descriptions can be combined with supervised learning in few shot settings.

A *pattern function* P takes a sequence of sentences as its input and outputs **one** sentence that contains only **one** masked token. The output can be viewed as a cloze question. A *verbalizer* v is a injective function that maps a label to a word in the vocabulary.

We can find pair of P, v to solve specific tasks. For example, if we need to solve a Question Answering(QA) task, given training example Q, A. We define our P as follows:

$$P((Q, A)) = Q.\_\_\_\_, A.$$

The task now changes from having to assign a label without inherent meaning to answering whether the most likely choice for the masked position in a "Yes" or a "No".
During Training, the cross entropy loss between the predicted label and the groundtruth is minimized. Nevertheless, the nature of few shot supervised learning means that "catastrophic forgetting"[7] can occur. So the final loss is a combined loss of the masked language model and the cross entropy loss. (Experiment setting $\alpha = 1 * 10^{-4}$

$$L = (1 - \alpha)L_{CE} + \alpha \cdot L_{MLM}$$

1.2.1 Key Chanllenges of PET

One of the chanllenges of PET is that we need to manually search for a good pattern function P. The author uses knowledge distillation[1] to ensemble a group of potentially good pattern functions.

1. During finetuning, training different models fore each pattern function P.
2. During ensembling, computes scores by a uniformly weighted or accuracy weighted ensemble of the models. Use softmax to transform the scores into a probability distribution. Use $T = 2$ to get a soft distribution[1]. All of the pairs form a new training set T_C
3. Finetune PLM with a standard sequence classification on T_C.

References

[1] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. 2015. arXiv: 1503.02531 [stat.ML].

[2] Dan Hendrycks and Kevin Gimpel. "Bridging Nonlinearities and Stochastic Regularizers with Gaussian Error Linear Units". In: *CoRR* abs/1606.08415 (2016). arXiv: 1606.08415. URL: http://arxiv.org/abs/1606.08415.

[3] Ashish Vaswani et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[4] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[5] Tom B. Brown et al. “Language Models are Few-Shot Learners”. In: *CoRR* abs/2005.14165 (2020). arXiv: 2005.14165. URL: https://arxiv.org/abs/2005.14165.

[6] Mike Lewis et al. “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703. URL: https://aclanthology.org/2020.acl-main.703.

[7] Timo Schick and Hinrich Schütze. “Exploiting Cloze Questions for Few-Shot Text Classification and Natural Language Inference”. In: *CoRR* abs/2001.07676 (2020). arXiv: 2001.07676. URL: https://arxiv.org/abs/2001.07676.