Large Language Models

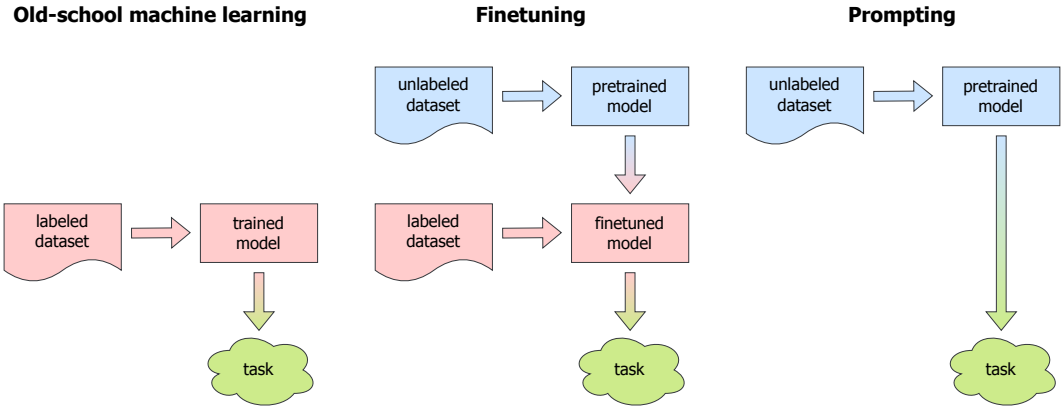
Part 4: Finetuning

David Samuel

Language Technology Group, University of Oslo



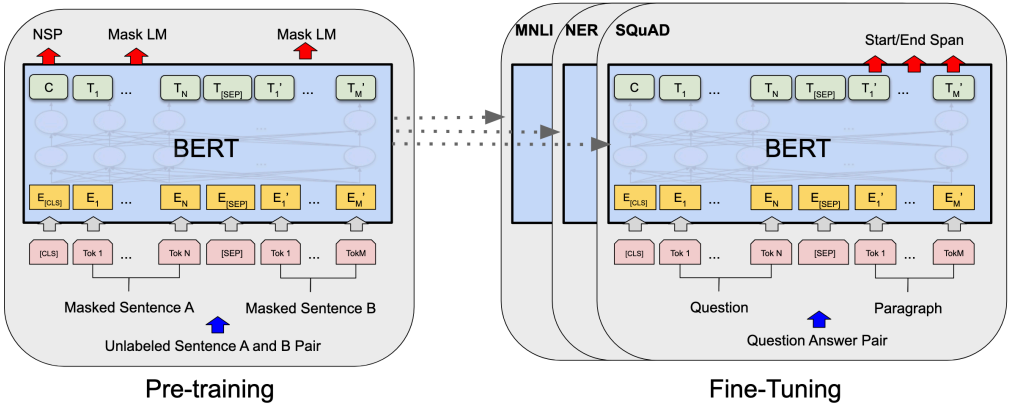
Pretraining / finetuning paradigm



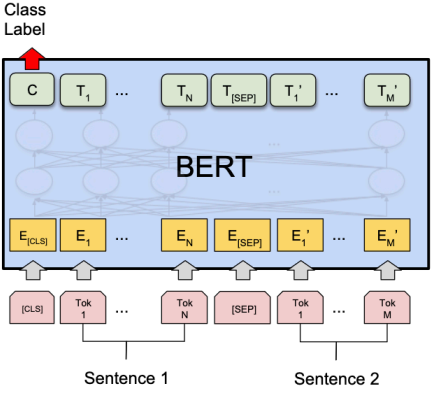
Finetuning

- Usually transfer learning from a self-supervised task to a supervised task
  - but in general "finetuning" refers to any kind of training that adapts/specializes a pre-trained model
- First, we train a general language model and then specialize it on a downstream task
- The language model can be either:
  - frozen: we don't change its weights, we only use its contextualized embeddings and train a classifier on top
  - further trained: we train it together with the classifier (if any)

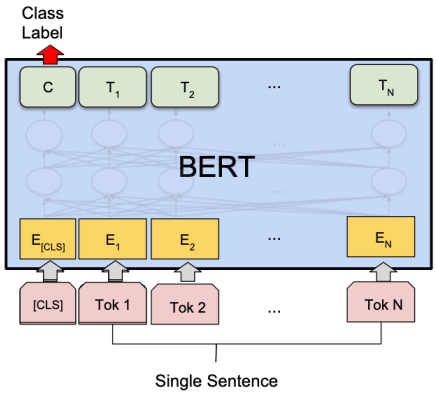
Finetuning of BERT (Devlin et al., 2019)



Finetuning of BERT (Devlin et al., 2019)

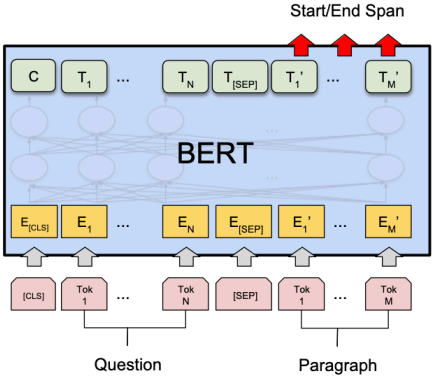


(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

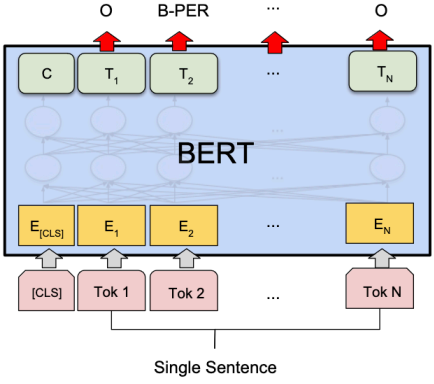


(b) Single Sentence Classification Tasks: SST-2, CoLA

Finetuning of BERT (Devlin et al., 2019)



(c) Question Answering Tasks: SQuAD v1.1



(d) Single Sentence Tagging Tasks: CoNLL-2003 NER



Finetuning of BERT (Devlin et al., 2019), practicalities

- Almost always, the language models are overparametrized for the supervised data
- We have to be very careful with overfitting!
- We should also be *gentle*, BERT is used to do MLM, not our supervised task



Finetuning of BERT (Devlin et al., 2019), practicalities

- Almost always, the language models are overparametrized for the supervised data
- We have to be very careful with overfitting!
- We should also be *gentle*, BERT is used to do MLM, not our supervised task
- Regularization is important
  - large weight decay (0.1 by default), early stopping (usually it is enough to train for just 2-8 epochs) and dropout
- Learning rate schedule is also crucial
  - The peak LR should be small (around 3e-5 for AdamW) and we should use warm-up

Finetuning of GPT (Radford et al., 2018)

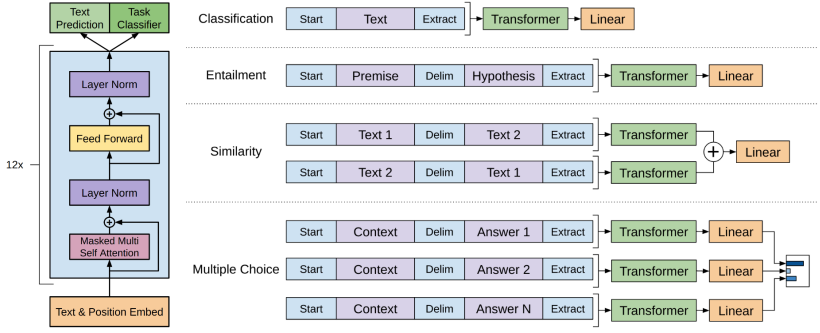
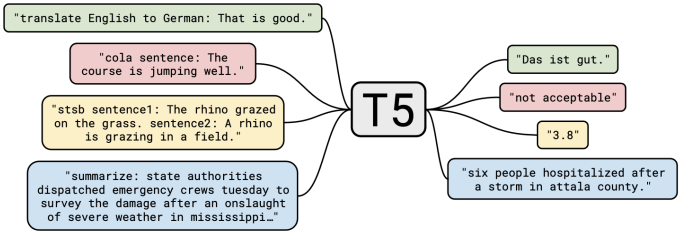


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Finetuning of T5 (Raffel et al., 2019)



- A different approach than BERT and GPT, it explains the name "text-to-text transfer transformer"
- We don't need any classifier on top of T5, we reformulate classification tasks as sequence-to-sequence tasks
- This approach can be used in any generative LM, for example GPT

References I

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.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer.