Prompt tuning (Lester et al. EMNLP 2021)

T5 model:

- Transformer encoder
- Transformer decoder
- cross att

this movie is awesome

- predict "positive"
- fine-tuning whole model "model tuning"

"positive"

- "prompt embedding" preferred to all downstream examples
- Transformer encoder
- Transformer decoder
- no gradient update applied (frozen)
- perform updates

- in contrast to "model tuning"
- here only the prompt embeddings are updated, everything else is frozen.