# BERT: Masked Language Modeling

- Uses transformers instead of LSTMs
- Bidirectional Model w/ Masked LM objective instead of normal LM
- Fine tune instead of freeze at test time
- Operates over word pieces (Byte-pair encoding)

* How to prevent cheating? Next word prediction doesn't work for bidirectional models, so we do masked language modeling instead:
  - Take chunk of text, mask out 15% of tokens, try to predict them

- What can BERT do?
  - Artificial [CLS] token is used as vector to do classification from
  - Sentence pair tasks (entailment): feed both sentences into BERT
  - BERT can also do tagging by predicting tags at each word piece

- What can't BERT do?
  - Generate text
  - Can fill in mask tokens, but can't generate left to right (you can put Mask at the end repeatedly, but this is slow $O(n^2)$)

- Fine-tuning a pre-trained BERT
  - Fine tune for 1-3 epochs, small learning rate (gradient descent)
  - Large changes to weights in last layer to route correct info to [CLS]
  - Smaller changes to weights lower down in transformer
  - Small LR and fine-tuning schedule means weights don't change much