Today: from ELMo to BERT
- from language modeling to masked LM

Goal of pretraining: use these big LMs as text encoders. Their goal is to enable downstream models to focus on the task at hand, instead of learning how language works.

ELMo:

h_L captures information about "this movie is", no info about the word "great"  
left-to-right LM  
h_L,...,n

h_R captures "great is"  
right-to-left LM  
h_R,...,n

Diagram of ELMo encoding:
- Blue arrows for left-to-right:
  - "this movie is great"
- Green arrows for right-to-left:
  - "great is the movie there"
**downstream usage**

**downstream model:** RMV

**classifier**

```
this → movie → is → great
```

**predict positive (softmax layer)**

**downstream model:** average emb. classifier

```
'neural bag of words'
```

```
(element-wise average)
```

```
this movie is great
```

(ex:)

```
i loved the acting, but (the rest of the movie was terrible).
```
the ELMo approach of two separate LMs that are then concat together is a little hacky ...

- can we accomplish the same goal within a single model
- change pretraining obj.
  - from LM to masked LMs

masked LM:
- given a full sequence of words (not just prefix) where X% of the words have been masked out
- instead of predicting the next word, we only predict masked words

![Diagram of Transformer xN with Q, K, V connections and predict 'opened' softmax layer]
- All of the final layer representations are fully contextualized
  - "aware" of words in the past as well as words in the future
- Same training loss as NLMS
  - Minimizing neg. log likelihood of the ground-truth (unmasked) tokens

ELMo $\Rightarrow$ BERT:
- $2$ unidirectional LMs $\Rightarrow 1$ masked LM
- Recurrent models to Transformers
- BERT was pretrained on a LOT more data

Contextualized token vectors in one pass thru the model
The movie is great.

How do we use BERT for a downstream task? The pretrained architecture is almost the same as the downstream model.

E.g., sentiment analysis:
- Add a special token to the beginning of every sequence
  - [CLS] token

Pretrained model

Softmax layer

Predict positive
This movie is great

- backprop the error signal from the sentiment classifier through the entire pretrained masked LM
- "fine-tuning"
- no external downstream model
  - only new component is a single softmax layer