From ELMo to BERT:

**pre-2018:**

```
<table>
<thead>
<tr>
<th>movie</th>
<th>is</th>
<th>good</th>
</tr>
</thead>
</table>
```

trained from scratch!

**ELMo:**

**step 2:**  
pretrain an RNN LM on lots of unlabeled data

```
<table>
<thead>
<tr>
<th>movie</th>
<th>is</th>
<th>good</th>
<th>(Eos)</th>
</tr>
</thead>
</table>
```

pretrained on ~1B words

**Step 2:** freeze LM parameters, use its representations (hidden states) as input to a task-specific model
**ELMo setup**:

- Forward LM, backward LM \( \implies \) Concatenate

**Forward LM**:

```
<table>
<thead>
<tr>
<th>movie</th>
<th>is</th>
<th>good</th>
<th>EOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>this</td>
<td>movie</td>
<td>is</td>
<td>good</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

- has access to "this movie is"
backwards LM

\[ \text{is} \rightarrow \text{movie} \rightarrow \text{this} \rightarrow \text{loss} \]

\[
\uparrow \quad \uparrow \quad \uparrow \quad \uparrow
\]

\[
\text{good} \quad \text{is} \quad \text{has access} \quad \text{to "good is"}
\]

\[
\uparrow \quad \uparrow \quad \uparrow \quad \uparrow
\]

\[
\text{this} \quad \text{movie} \quad \text{is} \quad \text{good}
\]

---

forward / backward LM is clumsy
Can we replace those w/ a single model?

\[ \text{ELMo} \Rightarrow \text{BERT} \]


- 2 unidirectional LMs \Rightarrow \text{1 masked LM}
- recurrent NNs to Transformers
- freezing the LM to fine-tuning LM
- pretrained LM on way more data, very big model
Masked LM: input is a sequence where some tokens have been randomly masked out. Goal: predict identity of the masked tokens.

Effect of increasing [MASK]%:

Students [MASK] their books
Students [MASK] their [MASK]
BERT: [MASK] \% of 15\% 

how do we use BERT to solve an NLP task like sentiment analysis?

**Pretraining task:** masked LM

**Fine-tuning task:** ex. sentiment analysis

Applying BERT for text classification

[CLS] special token

used for classification tasks

BERT's parameters are "fine-tuned" via the sentiment signal

predict positive

this movie is good

"transfer"
**Termimology:**

**pretrain:** start w/ randomly init. model, train it w/ a self-supervised obj.

- LM, masked LM
- data is free
- big models on big data

**freeze:** do not backprop into the params of the pretrained model using the downstream objective

**fine-tuning:** backprop into the pretrained model using task-specific signal

softmax is trained from scratch