Natural Language Processing with Deep Learning
CS224N/Ling284

Christopher Manning
Lecture 9: Pretraining

Adapted from slides by Anna Goldie, John Hewitt, Tatsunori Hashimoto
The pretraining revolution

Pretraining has had a major, tangible impact on how well NLP systems work.
Pretraining – scaling unsupervised learning on the internet

Key ideas in pretraining
• Make sure your model can process large-scale, diverse datasets
• Don’t use labeled data (otherwise you can’t scale!)
• Compute-aware scaling
Lecture Plan

1. A brief note on subword modeling
2. Motivating model pretraining from word embeddings
3. Model pretraining three ways
   1. Decoders
   2. Encoders
   3. Encoder-Decoders
4. What do we think pretraining is teaching?

Reminders and notes:
Ass 3 is due and Ass 4 is out today! Ass 4 covers lecture 8 and lecture 9 (today)!
Project proposal is due Thursday. Get it in on time (I beseech you)
Rest of the quarter: This week and next, more LLMs; then topics and invited speakers
Let’s take a look at the assumptions we’ve made about a language’s vocabulary.

We assume a fixed vocab of tens of thousands of words, built from the training set. All *novel* words seen at test time are mapped to a single UNK.

<table>
<thead>
<tr>
<th>word</th>
<th>vocab mapping</th>
<th>embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>hat</td>
<td>pizza (index)</td>
<td></td>
</tr>
<tr>
<td>learn</td>
<td>tasty (index)</td>
<td></td>
</tr>
<tr>
<td>taaaasty</td>
<td>UNK (index)</td>
<td></td>
</tr>
<tr>
<td>laern</td>
<td>UNK (index)</td>
<td></td>
</tr>
<tr>
<td>Transformerify</td>
<td>UNK (index)</td>
<td></td>
</tr>
</tbody>
</table>

Common words

Variations

misspellings

novel items

---

Brianna @_parsimonia_ · 24h
Gooooooood Vibessssssss
Word structure and subword models

Finite vocabulary assumptions make even less sense in many languages other than English

- Many languages exhibit complex morphology, or word structure.
  - The effect is many more word types, each occurring fewer times.

Example: Swahili verbs can have hundreds of conjugations, each encoding a wide variety of information. (Tense, mood, definiteness, negation, information about the object, ++)

Here’s a small fraction of the conjugations for *ambia* – to tell.
The byte-pair encoding algorithm

Subword modeling in NLP encompasses a wide range of methods for reasoning about structure below the word level. (Parts of words, characters, bytes.)

- The dominant modern paradigm is to learn a vocabulary of parts of words (subword tokens).
- At training and testing time, each word is split into a sequence of known subwords.

**Byte-pair encoding** is a simple, effective strategy for defining a subword vocabulary.

1. Start with a vocabulary containing only characters and an “end-of-word” symbol.
2. Using a corpus of text, find the most common adjacent characters “a,b”; add “ab” as a subword.
3. Replace instances of the character pair with the new subword; repeat until desired vocab size.

Originally used in NLP for machine translation; now similar methods (WordPiece, SentencePiece) are used in pretrained models, like BERT, GPT.

[Sennrich et al., 2016, Wu et al., 2016]
**Byte Pair Encoding (BPE) [Sennrich et al. 2016]**

<table>
<thead>
<tr>
<th>Dictionary</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 l o w</td>
<td>l, o, w, e, r, n, w, s, t, i, d</td>
</tr>
<tr>
<td>2 l o w e r</td>
<td>l, o, w, e, r, n, w, s, t, i, d, es</td>
</tr>
<tr>
<td>6 n e w e s t</td>
<td>l, o, w, e, r, n, w, s, t, i, d, es, est</td>
</tr>
<tr>
<td>3 w i d e s t</td>
<td>l, o, w, e, r, n, w, s, t, i, d, es, est</td>
</tr>
</tbody>
</table>
Word structure and subword models

Common words end up being a part of the subword vocabulary, while rarer words are split into (sometimes intuitive, sometimes not) components.

In the worst case, words are split into as many subwords as they have characters.

<table>
<thead>
<tr>
<th>word</th>
<th>vocab mapping</th>
<th>embedding</th>
</tr>
</thead>
<tbody>
<tr>
<td>hat</td>
<td>hat</td>
<td></td>
</tr>
<tr>
<td>learn</td>
<td>learn</td>
<td></td>
</tr>
<tr>
<td>taaaasty</td>
<td>taa## aaa## sty</td>
<td></td>
</tr>
<tr>
<td>laern</td>
<td>la## ern##</td>
<td></td>
</tr>
<tr>
<td>Transformerify</td>
<td>Transformer## ify</td>
<td></td>
</tr>
</tbody>
</table>
Words in writing systems

Writing systems vary in how they represent words – or don’t

- No word segmentation: 安理会认可利比亚问题柏林峰会成果
- Words (mainly) segmented: *This is a sentence with words.*
  - Clitics/pronouns/agreement?
    - Separated: *Je vous ai apporté* des bonbons
    - Joined: ف+ قال+ نا+ ها = قائلناها = so+said+we+it
- Compounds?
  - Separated: life insurance company employee
  - Joined: Lebensversicherungsgesellschaftsangestellter
Below the word in writing systems

Human language writing systems aren’t one thing!

- Phonemic (maybe digraphs) jiyawu ngabulu
- Fossilized phonemic thorough failure
- Syllabic/moraic 蛄𧎒鿅鿅鿅
- Ideographic (syllabic) 去年太空船二号坠毁
- Combination of the above インド洋の島

Wambaya
English
Inuktitut
Chinese
Japanese
Outline

1. A brief note on subword modeling
2. Motivating model pretraining from word embeddings
3. Model pretraining three ways
   1. Encoders
   2. Encoder-Decoders
   3. Decoders
4. What do we think pretraining is teaching?
Motivating word meaning and context

Recall the adage we mentioned at the beginning of the course:

“You shall know a word by the company it keeps” (J. R. Firth 1957: 11)

This quote is a summary of distributional semantics, and motivated word2vec. But:

“… the complete meaning of a word is always contextual, and no study of meaning apart from a complete context can be taken seriously.” (J. R. Firth 1935)

Consider *I record the record*: the two instances of *record* mean different things.

[Thanks to Yoav Goldberg on Twitter for pointing out the 1935 Firth quote.]
Where we were: pretrained word embeddings

Circa 2015:

- Start with pretrained word embeddings (no context!)
- Learn how to incorporate context in an LSTM or Transformer while training on the task.

Some issues to think about:

- The training data we have for our downstream task (like question answering) must be sufficient to teach all contextual aspects of language.
- Most of the parameters in our network are randomly initialized!

[Recall, movie gets the same word embedding, no matter what sentence it shows up in]
Where we’re going: **pretraining whole models**

In modern NLP:

- All (or almost all) parameters in NLP networks are initialized via **pretraining**.
- Pretraining methods hide parts of the input from the model, and train the model to reconstruct those parts.
- This has been exceptionally effective at building strong:
  - **representations of language**
  - **parameter initializations** for strong NLP models.
  - **Probability distributions** over language that we can sample from.

[This model has learned how to represent entire sentences through pretraining]
What can we learn from reconstructing the input?

Stanford University is located in ___________, California.
What can we learn from reconstructing the input?

I put ___ fork down on the table.
What can we learn from reconstructing the input?

The woman walked across the street, checking for traffic over ___ shoulder.
What can we learn from reconstructing the input?

I went to the ocean to see the fish, turtles, seals, and _____.
What can we learn from reconstructing the input?

Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___. 
Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the ______.
What can we learn from reconstructing the input?

I was thinking about the sequence that goes
1, 1, 2, 3, 5, 8, 13, 21, ____
Pretraining through language modeling [Dai and Le, 2015]

Recall the language modeling task:
• Model $p_{\theta}(w_t|w_{1:t-1})$, the probability distribution over words given their past contexts.
• There’s lots of data for this! (In English.)

Pretraining through language modeling:
• Train a neural network to perform language modeling on a large amount of text.
• Save the network parameters.
The Pretraining / Finetuning Paradigm

Pretraining can improve NLP applications by serving as parameter initialization.

**Step 1: Pretrain (on language modeling)**
Lots of text; learn general things!

(Transformer, LSTM, ++ )

**Step 2: Finetune (on your task)**
Not many labels; adapt to the task!

(Transformer, LSTM, ++ )

... the movie was ...

Iroh goes to make tasty tea
Stochastic gradient descent and pretrain/finetune

Why should pretraining and finetuning help, from a “training neural nets” perspective?

• Pretraining provides parameters $\hat{\theta}$ by approximating $\min_{\theta} \mathcal{L}_{\text{pretrain}}(\theta)$.
  • (The pretraining loss.)
• Then, finetuning approximates $\min_{\theta} \mathcal{L}_{\text{finetune}}(\theta)$, starting at $\hat{\theta}$.
  • (The finetuning loss)
• The pretraining may matter because stochastic gradient descent sticks (relatively) close to $\hat{\theta}$ during finetuning.
  • So, maybe the finetuning local minima near $\hat{\theta}$ tend to generalize well!
  • And/or, maybe the gradients of finetuning loss near $\hat{\theta}$ propagate nicely!
Where does this data come from?

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT</td>
<td>BookCorpus, English Wikipedia</td>
</tr>
<tr>
<td>GPT-1</td>
<td>BookCorpus</td>
</tr>
<tr>
<td>GPT-3</td>
<td>CommonCrawl, WebText, English Wikipedia, and 2 book databases (“Books 1” and “Books 2”)</td>
</tr>
<tr>
<td>GPT-3.5+</td>
<td>Undisclosed</td>
</tr>
</tbody>
</table>
Bookcorpus ... what’s that?

• Scraped ebooks from the internet – highly controversial
Google swallows 11,000 novels to improve AI's conversation

As writers learn that tech giant has processed their work without permission, the Authors Guild condemns 'blatantly commercial use of expressive authorship'

Reexamining "Fair Use" in the Age of AI

Generative AI claims to produce new language and images, but when those ideas are based on copyrighted material, who gets the credit? A new paper from Stanford University looks for answers.

Jun 5, 2023 | Andrew Myers
Lecture Plan

1. A brief note on subword modeling
2. Motivating model pretraining from word embeddings
3. Model pretraining three ways
   1. Encoders
   2. Encoder-Decoders
   3. Decoders
4. What do we think pretraining is teaching?
Pretraining for three types of architectures

The neural architecture influences the type of pretraining, and natural use cases.

**Encoders**
- Gets bidirectional context – can condition on future!
- How do we train them to build strong representations?

**Decoder-Decoders**
- Good parts of decoders and encoders?
- What’s the best way to pretrain them?

**Decoders**
- Language models! What we’ve seen so far.
- Nice to generate from; can’t condition on future words.
Pretraining for three types of architectures

The neural architecture influences the type of pretraining, and natural use cases.

**Encoders**
- Gets bidirectional context – can condition on future!
- How do we train them to build strong representations?

**Encoder-Decoders**
- Good parts of decoders and encoders?
- What’s the best way to pretrain them?

**Decoders**
- Language models! What we’ve seen so far.
- Nice to generate from; can’t condition on future words
Pretraining encoders: what pretraining objective to use?

So far, we’ve looked at language model pretraining. But encoders get bidirectional context, so we can’t do language modeling!

**Idea:** replace some fraction of words in the input with a special [MASK] token; predict these words.

\[
h_1, ..., h_T = \text{Encoder}(w_1, ..., w_T) \\
y_i \sim Ah_i + b
\]

Only add loss terms from words that are “masked out.” If \(\tilde{x}\) is the masked version of \(x\), we’re learning \(p_\theta(x|\tilde{x})\). Called **Masked LM**.

[Devlin et al., 2018]
Devlin et al., 2018 proposed the “Masked LM” objective and released the weights of a pretrained Transformer, a model they labeled BERT.

Some more details about Masked LM for BERT:

- Predict a random 15% of (sub)word tokens.
  - Replace input word with [MASK] 80% of the time
  - Replace input word with a random token 10% of the time
  - Leave input word unchanged 10% of the time (but still predict it!)
- Why? Doesn’t let the model get complacent and not build strong representations of non-masked words. (No masks are seen at fine-tuning time!)
The pretraining input to BERT was two separate contiguous chunks of text:

- BERT was trained to predict whether one chunk follows the other or is randomly sampled.
  - Later work has argued this “next sentence prediction” is not necessary.

[Devlin et al., 2018, Liu et al., 2019]
BERT: Bidirectional Encoder Representations from Transformers

Details about BERT

• Two models were released:
  • BERT-base: 12 layers, 768-dim hidden states, 12 attention heads, 110 million params.
  • BERT-large: 24 layers, 1024-dim hidden states, 16 attention heads, 340 million params.
• Trained on:
  • BooksCorpus (800 million words)
  • English Wikipedia (2,500 million words)
• Pretraining is expensive and impractical on a single GPU.
  • BERT was pretrained with 64 TPU chips for a total of 4 days.
  • (TPUs are special tensor operation acceleration hardware)
• Finetuning is practical and common on a single GPU
  • “Pretrain once, finetune many times.”

[Devlin et al., 2018]
BERT: Bidirectional Encoder Representations from Transformers

BERT was massively popular and hugely versatile; finetuning BERT led to new state-of-the-art results on a broad range of tasks.

- **QQP**: Quora Question Pairs (detect paraphrase questions)
- **QNLI**: natural language inference over question answering data
- **SST-2**: sentiment analysis
- **CoLA**: corpus of linguistic acceptability (detect whether sentences are grammatical.)
- **STS-B**: semantic textual similarity
- **MRPC**: microsoft paraphrase corpus
- **RTE**: a small natural language inference corpus

<table>
<thead>
<tr>
<th>System</th>
<th>MNLI-(m/mm)</th>
<th>QQP</th>
<th>QNLI</th>
<th>SST-2</th>
<th>CoLA</th>
<th>STS-B</th>
<th>MRPC</th>
<th>RTE</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-OpenAI SOTA</td>
<td>80.6/80.1</td>
<td>66.1</td>
<td>82.3</td>
<td>93.2</td>
<td>35.0</td>
<td>81.0</td>
<td>86.0</td>
<td>61.7</td>
<td>74.0</td>
</tr>
<tr>
<td>BiLSTM+ELMo+Attn</td>
<td>76.4/76.1</td>
<td>64.8</td>
<td>79.8</td>
<td>90.4</td>
<td>36.0</td>
<td>73.3</td>
<td>84.9</td>
<td>56.8</td>
<td>71.0</td>
</tr>
<tr>
<td>OpenAI GPT</td>
<td>82.1/81.4</td>
<td>70.3</td>
<td>87.4</td>
<td>91.3</td>
<td>45.4</td>
<td>80.0</td>
<td>82.3</td>
<td>56.0</td>
<td>75.1</td>
</tr>
<tr>
<td>BERT\textsubscript{BASE}</td>
<td>84.6/83.4</td>
<td>71.2</td>
<td>90.5</td>
<td>93.5</td>
<td>52.1</td>
<td>85.8</td>
<td>88.9</td>
<td>66.4</td>
<td>79.6</td>
</tr>
<tr>
<td>BERT\textsubscript{LARGE}</td>
<td>86.7/85.9</td>
<td>72.1</td>
<td>92.7</td>
<td>94.9</td>
<td>60.5</td>
<td>86.5</td>
<td>89.3</td>
<td>70.1</td>
<td>82.1</td>
</tr>
</tbody>
</table>

[Devlin et al., 2018]
Limitations of pretrained encoders

Those results looked great! Why not use pretrained encoders for everything?

If your task involves generating sequences, consider using a pretrained decoder; BERT and other pretrained encoders don’t naturally lead to nice autoregressive (1-word-at-a-time) generation methods.
Extensions of BERT

You’ll see a lot of BERT variants like RoBERTa, SpanBERT, +++

Some generally accepted improvements to the BERT pretraining formula:

• RoBERTa: mainly just train BERT for longer and remove next sentence prediction!
• SpanBERT: masking contiguous spans of words makes a harder, more useful pretraining task

[MASK] irr## esi## sti## [MASK] good

It’s irr## esi## sti## bly

[MASK] [MASK] [MASK] [MASK] good

It’s irr## esi## sti## bly

[Liu et al., 2019; Joshi et al., 2020]
Extensions of BERT

A takeaway from the RoBERTa paper: more compute, more data can improve pretraining even when not changing the underlying Transformer encoder.

<table>
<thead>
<tr>
<th>Model</th>
<th>data</th>
<th>bsz</th>
<th>steps</th>
<th>SQuAD (v1.1/2.0)</th>
<th>MNLI-m</th>
<th>SST-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoBERTa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with BOOKS + WIKI</td>
<td>16GB</td>
<td>8K</td>
<td>100K</td>
<td>93.6/87.3</td>
<td>89.0</td>
<td>95.3</td>
</tr>
<tr>
<td>+ additional data (§3.2)</td>
<td>160GB</td>
<td>8K</td>
<td>100K</td>
<td>94.0/87.7</td>
<td>89.3</td>
<td>95.6</td>
</tr>
<tr>
<td>+ pretrain longer</td>
<td>160GB</td>
<td>8K</td>
<td>300K</td>
<td>94.4/88.7</td>
<td>90.0</td>
<td>96.1</td>
</tr>
<tr>
<td>+ pretrain even longer</td>
<td>160GB</td>
<td>8K</td>
<td>500K</td>
<td><strong>94.6/89.4</strong></td>
<td><strong>90.2</strong></td>
<td><strong>96.4</strong></td>
</tr>
<tr>
<td>BERT$_{\text{LARGE}}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>with BOOKS + WIKI</td>
<td>13GB</td>
<td>256</td>
<td>1M</td>
<td>90.9/81.8</td>
<td>86.6</td>
<td>93.7</td>
</tr>
</tbody>
</table>

[Liu et al., 2019; Joshi et al., 2020]
Pretraining for three types of architectures

The neural architecture influences the type of pretraining, and natural use cases.

- **Encoders**
  - Gets bidirectional context – can condition on future!
  - How do we train them to build strong representations?

- **Encoder-Decoders**
  - Good parts of decoders and encoders?
  - What’s the best way to pretrain them?

- **Decoders**
  - Language models! What we’ve seen so far.
  - Nice to generate from; can’t condition on future words.
Pretraining encoder-decoders: what pretraining objective to use?

For **encoder-decoders**, we could do something like **language modeling**, but where a prefix of every input is provided to the encoder and is not predicted.

\[
\begin{align*}
h_1, \ldots, h_T &= \text{Encoder}(w_1, \ldots, w_T) \\
h_{T+1}, \ldots, h_{T+S} &= \text{Decoder}(w_{T+1}, \ldots, w_{T+S}, h_1, \ldots, h_T) \\
y_i &\sim Ah_i + b, \; i > T
\end{align*}
\]

The **encoder** portion can benefit from bidirectional context; the **decoder** portion is used to train the whole model through language modeling, autoregressively predicting and then conditioning on one token at a time.

[Raffel et al., 2018]
Pretraining encoder-decoders: what pretraining objective to use?

What Raffel et al., 2018 found to work best was **span corruption**. Their model: **T5**.

Replace different-length spans from the input with unique placeholders; decode out the spans that were removed!

This is implemented in text preprocessing: it’s still an objective that looks like **language modeling** at the decoder side.
Raffel et al., 2018 found encoder-decoders to work better than decoders for their tasks, and span corruption (denoising) to work better than language modeling.

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Objective</th>
<th>Params</th>
<th>Cost</th>
<th>GLUE</th>
<th>CNNDM</th>
<th>SQuAD</th>
<th>SGLUE</th>
<th>EnDe</th>
<th>EnFr</th>
<th>EnRo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encoder-decoder</td>
<td>Denoising</td>
<td>$2P$</td>
<td>$M$</td>
<td><strong>83.28</strong></td>
<td><strong>19.24</strong></td>
<td><strong>80.88</strong></td>
<td><strong>71.36</strong></td>
<td><strong>26.98</strong></td>
<td><strong>39.82</strong></td>
<td><strong>27.65</strong></td>
</tr>
<tr>
<td>Enc-dec, shared</td>
<td>Denoising</td>
<td>$P$</td>
<td>$M$</td>
<td>82.81</td>
<td>18.78</td>
<td><strong>80.63</strong></td>
<td><strong>70.73</strong></td>
<td>26.72</td>
<td>39.03</td>
<td>27.46</td>
</tr>
<tr>
<td>Enc-dec, 6 layers</td>
<td>Denoising</td>
<td>$P$</td>
<td>$M/2$</td>
<td>80.88</td>
<td>18.97</td>
<td>77.59</td>
<td>68.42</td>
<td>26.38</td>
<td>38.40</td>
<td>26.95</td>
</tr>
<tr>
<td>Language model</td>
<td>Denoising</td>
<td>$P$</td>
<td>$M$</td>
<td>74.70</td>
<td>17.93</td>
<td>61.14</td>
<td>55.02</td>
<td>25.09</td>
<td>35.28</td>
<td>25.86</td>
</tr>
<tr>
<td>Prefix LM</td>
<td>Denoising</td>
<td>$P$</td>
<td>$M$</td>
<td>81.82</td>
<td>18.61</td>
<td>78.94</td>
<td>68.11</td>
<td>26.43</td>
<td>37.98</td>
<td>27.39</td>
</tr>
<tr>
<td>Encoder-decoder</td>
<td>LM</td>
<td>$2P$</td>
<td>$M$</td>
<td>79.56</td>
<td>18.59</td>
<td>76.02</td>
<td>64.29</td>
<td>26.27</td>
<td>39.17</td>
<td>26.86</td>
</tr>
<tr>
<td>Enc-dec, shared</td>
<td>LM</td>
<td>$P$</td>
<td>$M$</td>
<td>79.60</td>
<td>18.13</td>
<td>76.35</td>
<td>63.50</td>
<td>26.62</td>
<td>39.17</td>
<td>27.05</td>
</tr>
<tr>
<td>Enc-dec, 6 layers</td>
<td>LM</td>
<td>$P$</td>
<td>$M/2$</td>
<td>78.67</td>
<td>18.26</td>
<td>75.32</td>
<td>64.06</td>
<td>26.13</td>
<td>38.42</td>
<td>26.89</td>
</tr>
<tr>
<td>Language model</td>
<td>LM</td>
<td>$P$</td>
<td>$M$</td>
<td>73.78</td>
<td>17.54</td>
<td>53.81</td>
<td>56.51</td>
<td>25.23</td>
<td>34.31</td>
<td>25.38</td>
</tr>
<tr>
<td>Prefix LM</td>
<td>LM</td>
<td>$P$</td>
<td>$M$</td>
<td>79.68</td>
<td>17.84</td>
<td>76.87</td>
<td>64.86</td>
<td>26.28</td>
<td>37.51</td>
<td>26.76</td>
</tr>
</tbody>
</table>
Pretraining encoder-decoders: what pretraining objective to use?

A fascinating property of T5: it can be finetuned to answer a wide range of questions, retrieving knowledge from its parameters.

NQ: Natural Questions
WQ: WebQuestions
TQA: Trivia QA

All “open-domain” versions

<table>
<thead>
<tr>
<th>Model</th>
<th>NQ</th>
<th>WQ</th>
<th>TQA dev</th>
<th>TQA test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Karpukhin et al. (2020)</td>
<td>41.5</td>
<td>42.4</td>
<td>57.9</td>
<td>-</td>
</tr>
<tr>
<td>T5.1.1-Base</td>
<td>25.7</td>
<td>28.2</td>
<td>24.2</td>
<td>30.6</td>
</tr>
<tr>
<td>T5.1.1-Large</td>
<td>27.3</td>
<td>29.5</td>
<td>28.5</td>
<td>37.2</td>
</tr>
<tr>
<td>T5.1.1-XL</td>
<td>29.5</td>
<td>32.4</td>
<td>36.0</td>
<td>45.1</td>
</tr>
<tr>
<td>T5.1.1-XXL</td>
<td>32.8</td>
<td>35.6</td>
<td>42.9</td>
<td>52.5</td>
</tr>
<tr>
<td>T5.1.1-XXL + SSM</td>
<td>35.2</td>
<td>42.8</td>
<td>51.9</td>
<td>61.6</td>
</tr>
</tbody>
</table>

[Raffel et al., 2018]
Pretraining for three types of architectures

The neural architecture influences the type of pretraining, and natural use cases.

**Encoders**
- Gets bidirectional context – can condition on future!
- How do we train them to build strong representations?

**Encoder-Decoders**
- Good parts of decoders and encoders?
- What’s the best way to pretrain them?

**Decoders**
- Language models! What we’ve seen so far.
- Nice to generate from; can’t condition on future words.
- All the biggest pretrained models are Decoders.
Pretraining decoders

When using language model pretrained decoders, we can ignore that they were trained to model \( p(w_t | w_{1:t-1}) \).

We can finetune them by training a softmax classifier on the last word’s hidden state.

\[
\begin{align*}
h_1, ..., h_T &= \text{Decoder}(w_1, ..., w_T) \\
y &\sim Ah_T + b
\end{align*}
\]

Where \( A \) and \( b \) are randomly initialized and specified by the downstream task.

Gradients backpropagate through the whole network.

[Note how the linear layer hasn’t been pretrained and must be learned from scratch.]
Pretraining decoders

It’s natural to pretrain decoders as language models and then use them as generators, finetuning their $p_\theta(w_t | w_{1:t-1})$!

This is helpful in tasks where the output is a sequence with a vocabulary like that at pretraining time!

- Dialogue (context=dialogue history)
- Summarization (context=document)

$$h_1, ..., h_T = \text{Decoder}(w_1, ..., w_T)$$

$$w_t \sim A h_{t-1} + b$$

Where $A, b$ were pretrained in the language model!

[Note how the linear layer has been pretrained.]
2018’s GPT was a big success in pretraining a decoder!

- Transformer decoder with 12 layers, 117M parameters.
- 768-dimensional hidden states, 3072-dimensional feed-forward hidden layers.
- Byte-pair encoding with 40,000 merges
- Trained on BooksCorpus: over 7000 unique books.
  - Contains long spans of contiguous text, for learning long-distance dependencies.
- The acronym “GPT” never showed up in the original paper; it could stand for “Generative PreTraining” or “Generative Pretrained Transformer”
Generative Pretrained Transformer (GPT) [Radford et al., 2018]

How do we format inputs to our decoder for finetuning tasks?

**Natural Language Inference:** Label pairs of sentences as *entailing/contradictory/neutral*

Premise: *The man is in the doorway*

Hypothesis: *The person is near the door*  \[\text{entailment}\]

Radford et al., 2018 evaluate on natural language inference.

Here’s roughly how the input was formatted, as a sequence of tokens for the decoder.

\[\text{[START]} \ The \ man \ is \ in \ the \ doorway \ [\text{DELIM}] \ The \ person \ is \ near \ the \ door \ [\text{EXTRACT}]\]

The linear classifier is applied to the representation of the [EXTRACT] token.
Generative Pretrained Transformer (GPT) [Radford et al., 2018]

GPT results on various *natural language inference* datasets.

<table>
<thead>
<tr>
<th>Method</th>
<th>MNLI-m</th>
<th>MNLI-mm</th>
<th>SNLI</th>
<th>SciTail</th>
<th>QNLI</th>
<th>RTE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESIM + ELMo [44] (5x)</td>
<td>-</td>
<td>-</td>
<td>89.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CAFE [58] (5x)</td>
<td>80.2</td>
<td>79.0</td>
<td>89.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Stochastic Answer Network [35] (3x)</td>
<td>80.6</td>
<td>80.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CAFE [58]</td>
<td>78.7</td>
<td>77.9</td>
<td>88.5</td>
<td>83.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GenSen [64]</td>
<td>71.4</td>
<td>71.3</td>
<td>-</td>
<td>-</td>
<td>82.3</td>
<td>59.2</td>
</tr>
<tr>
<td>Multi-task BiLSTM + Attn [64]</td>
<td>72.2</td>
<td>72.1</td>
<td>-</td>
<td>-</td>
<td>82.1</td>
<td>61.7</td>
</tr>
<tr>
<td>Finetuned Transformer LM (ours)</td>
<td><strong>82.1</strong></td>
<td><strong>81.4</strong></td>
<td><strong>89.9</strong></td>
<td>88.3</td>
<td><strong>88.1</strong></td>
<td>56.0</td>
</tr>
</tbody>
</table>
We mentioned how pretrained decoders can be used in their capacities as language models. **GPT-2**, a larger version (1.5B) of GPT trained on more data, was shown to produce relatively convincing samples of natural language.

Increasingly convincing generations (GPT2) [Radford et al., 2018]

---

**Context (human-written):** In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

**GPT-2:** The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.
A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials.
GPT-3, In-context learning, and very large models

So far, we’ve interacted with pretrained models in two ways:

• Sample from the distributions they define (maybe providing a prompt)
• Fine-tune them on a task we care about and take their predictions.

Very large language models seem to perform some kind of learning **without gradient steps** simply from examples you provide within their contexts.

GPT-3 is the canonical example of this. The largest T5 model had 11 billion parameters. **GPT-3 has 175 billion parameters.**

**ChatGPT/GPT-4/GPT-3.5 Turbo** introduced a further instruction-tuning idea that we cover next lecture
GPT-3, In-context learning, and very large models

Very large language models seem to perform some kind of learning without gradient steps simply from examples you provide within their contexts.

The in-context examples seem to specify the task to be performed, and the conditional distribution mocks performing the task to a certain extent.

**Input (prefix within a single Transformer decoder context):**

“
thanks -> merci
hello -> bonjour
mint -> menthe
otter ->
”

**Output (conditional generations):**

loutre...”
GPT-3, In-context learning, and very large models

Very large language models seem to perform some kind of learning without gradient steps simply from examples you provide within their contexts.
**Why scale? Scaling laws**

- Empirical observation: scaling up models leads to reliable gains in perplexity

![Graphs showing scaling laws](image_url)
Scaling can help identify model size – data tradeoffs

- Modern observation: train a big model that’s not fully converged.
Scaling laws for many other interesting architecture decisions

• Predictable scaling helps us make intelligent decisions about architectures etc.
Scaling Efficiency: how do we best use our compute

GPT-3 was **175B parameters** and trained on **300B** tokens of text. Roughly, the cost of training a large transformer scales as **parameters*tokens**.

Did OpenAI strike the right parameter-token data to get the best model? No.

<table>
<thead>
<tr>
<th>Model</th>
<th>Size (# Parameters)</th>
<th>Training Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>LaMDA (Thoppilan et al., 2022)</td>
<td>137 Billion</td>
<td>168 Billion</td>
</tr>
<tr>
<td>GPT-3 (Brown et al., 2020)</td>
<td>175 Billion</td>
<td>300 Billion</td>
</tr>
<tr>
<td>Jurassic (Lieber et al., 2021)</td>
<td>178 Billion</td>
<td>300 Billion</td>
</tr>
<tr>
<td>Gopher (Rae et al., 2021)</td>
<td>280 Billion</td>
<td>300 Billion</td>
</tr>
<tr>
<td>MT-NLG 530B (Smith et al., 2022)</td>
<td>530 Billion</td>
<td>270 Billion</td>
</tr>
<tr>
<td>Chinchilla</td>
<td>70 Billion</td>
<td>1.4 Trillion</td>
</tr>
</tbody>
</table>

This 70B parameter model is better than the much larger other models!
Outline

1. A brief note on subword modeling
2. Motivating model pretraining from word embeddings
3. Model pretraining three ways
   1. Encoders
   2. Encoder-Decoders
   3. Decoders
4. What do we think pretraining is teaching?
What kinds of things does pretraining teach?

There’s increasing evidence that pretrained models learn a wide variety of things about the statistical properties of language. Taking our examples from the start of class:

• Stanford University is located in __________, California. [Trivia]
• I put ___ fork down on the table. [syntax]
• The woman walked across the street, checking for traffic over ___ shoulder. [coreference]
• I went to the ocean to see the fish, turtles, seals, and _____. [lexical semantics/topic]
• Overall, the value I got from the two hours watching it was the sum total of the popcorn and the drink. The movie was ___. [sentiment]
• Iroh went into the kitchen to make some tea. Standing next to Iroh, Zuko pondered his destiny. Zuko left the ______. [some reasoning – this is harder]
• I was thinking about the sequence that goes 1, 1, 2, 3, 5, 8, 13, 21, ____ [some basic arithmetic; they don’t learn the Fibonacci sequence]
• Models also learn – and can exacerbate racism, sexism, all manner of bad biases.
Sometimes it also memorizes copyrighted material
Sometimes it learns some things we don’t want..

- **Membership inference** lets you recover parts of the training data

- Sometimes this training data is semi-private material from the web (addresses, emails)

- It learns the prejudices and biases of human beings who write online
Three types of architectures for pretraining

The neural architecture influences the type of pretraining, and natural use cases.

**Encoders**
- Gets bidirectional context – can condition on future!
- Good if only doing analysis of text (better than decoders)

**Encoder-Decoder**
- Good parts of decoders and encoders?
- Some evidence they are better for NLU
  - [Tay et al. 2022. UL2]

**Decoders**
- Language models! What we’ve seen so far. Scale well.
- Best to generate from; have won as to what people build