CS 886 Deep Learning and NLP

Ming Li
01. Word2Vec
02. Attention / Transformers
03. GPT / BERT
04. Simplicity, ALBERT, Single headed attention RNN
05. Student presentations Starting Feb. 3
06. Student presentations ending March 30
07. Student short presentations of research projects
LECTURE THREE

GPT-2 and BERT
Tying up loose ends from the last lecture, back to Lecture 2 notes.
Avoiding Information bottleneck
Last time we introduced transformer
Transformers, GPT-2, and BERT

1. A transformer uses Encoder stack to model input, and uses Decoder stack to model output (using input information from encoder side).
2. But if we do not have input, we just want to model the “next word”, we can get rid of the Encoder side of a transformer and output “next word” one by one. This gives us GPT.
3. If we are only interested in training a language model for the input for some other tasks, then we do not need the Decoder of the transformer, that gives us BERT.
GPT-2, BERT
GPT released June 2018
GPT-2 released Nov. 2019 with 1.5B parameters

GPT-2
- SMALL
  - Model Dimensionality: 768
  - 117M parameters

GPT-2 MEDIUM
- Model Dimensionality: 1024
- 345M parameters

GPT-2 LARGE
- Model Dimensionality: 1280
- 762M parameters

GPT-2 EXTRA LARGE
- Model Dimensionality: 1600
- 1542M parameters
GPT-2 in action

Output

| A | robot | may | not | injure | a | human | being |

Input

| recite | the | first | law | $ | A | robot | may | not | injure | a | human | being |
Byte Pair Encoding (BPE)

Word embedding sometimes is too high level, pure character embedding too low level. For example, if we have learned

   old   older   oldest
We might also wish the computer to infer

   smart   smarter   smartest

But at the whole word level, this might not be so direct. Thus the idea is to break the words up into pieces like er, est, and embed frequent fragments of words.

GPT adapts this BPE scheme.
GPT uses BPE scheme. The subwords are calculated by:

1. Split word to sequence of characters (add \(<w>\) char)
2. Joining the highest frequency pattern.
3. Keep doing step 2, until it hits the pre-defined maximum number of subwords or iterations.

Example:

\[
\{\text{low } \langle w \rangle \text{: 5}, \text{lower } \langle w \rangle \text{: 2}, \text{newest } \langle w \rangle \text{: 6}, \text{wdest } \langle w \rangle \text{: 3} \} \\
\{\text{low } \langle w \rangle \text{: 5}, \text{lower } \langle w \rangle \text{: 2}, \text{newest } \langle w \rangle \text{: 6}, \text{wdest } \langle w \rangle \text{: 3} \} \\
\{\text{low } \langle w \rangle \text{: 5}, \text{lower } \langle w \rangle \text{: 2}, \text{newest } \langle w \rangle \text{: 6}, \text{wdest } \langle w \rangle \text{: 3} \} \\
\{\text{low } \langle w \rangle \text{: 5}, \text{lower } \langle w \rangle \text{: 2}, \text{newest } \langle w \rangle \text{: 6}, \text{wdest } \langle w \rangle \text{: 3} \} \\
\{\text{low } \langle w \rangle \text{: 5}, \text{lower } \langle w \rangle \text{: 2}, \text{newest } \langle w \rangle \text{: 6}, \text{wdest } \langle w \rangle \text{: 3} \} \\
\]

Note that \(<w>\) is also an important character.
Masked Self-Attention

Self-Attention

Masked Self-Attention
Masked Self-Attention

Note: encoder-decoder attention block is gone
Masked Self-Attention Calculation

Re-use previous computation results: at any step, only need to results of q, k, v related to the new output word, no need to re-compute the others. Additional computation is linear, instead of quadratic.
GPT-2, BERT

GPT-2 fully connected network has two layers (Example for GPT-2 small)
GPT-2, BERT

GPT-2 has a parameter top-k, so that we sample works from top k (highest probability from softmax) words for each each output.
This top-k parameter, if $k=1$, we would have output like:

The first time I saw the new version of the game, I was so excited. I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game, I was so excited to see the new version of the game.
GPT Training

GPT-2 uses unsupervised learning approach to training the language model.

There is no custom training for GPT-2, no separation of pre-training and fine-tuning like BERT.
“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.

Pérez and the others then ventured further into the valley. `By the time we reached the top of one peak, the water looked blue, with some crystals on top,’ said Pérez.

Pérez and his friends were astonished to see the unicorn herd. These creatures could be seen from the air without having to move too much to see them – they were so close they could touch their horns."
GPT-2 Application: Translation

Training Dataset

<table>
<thead>
<tr>
<th>I</th>
<th>am</th>
<th>a</th>
<th>student</th>
<th>&lt;to-fr&gt;</th>
<th>je</th>
<th>suis</th>
<th>étudiant</th>
</tr>
</thead>
<tbody>
<tr>
<td>let</td>
<td>them</td>
<td>eat</td>
<td>cake</td>
<td>&lt;to-fr&gt;</td>
<td>Qu'ils</td>
<td>mangent</td>
<td>de</td>
</tr>
<tr>
<td>good</td>
<td>morning</td>
<td>&lt;to-fr&gt;</td>
<td>Bonjour</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Transformer-Decoder

Output #1
Position #4
Time step #1
Comment

Output #2
Position #5
Time step #2
allez-vous
GPT-2 Application: Summarization

Training Dataset

<table>
<thead>
<tr>
<th>Article #1 tokens</th>
<th>&lt;summarize&gt;</th>
<th>Article #1 Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article #2 tokens</td>
<td>&lt;summarize&gt;</td>
<td>Article #2 Summary</td>
</tr>
<tr>
<td>Article #3 tokens</td>
<td>&lt;summarize&gt;</td>
<td>Article #3 Summary</td>
</tr>
</tbody>
</table>

Transformer-Decoder

Output #1
- Position #114
- Time step #1

Output #2
- Position #115
- Time step #2

1 ... 113 114 256
Using wikipedia data
BERT (Bidirectional Encoder Representation from Transformers)
Model input dimension 512

Input and output vector size (Also 768, and 1024)
BERT pretraining

**ULM-FiT (2018):** Pre-training ideas, transfer learning in NLP.
**ELMo:** Bidirectional training (LSTM)
**Transformer:** Although used things from left, but still missing from the right.
**GPT:** Use Transformer Decoder half.
**BERT:** Switches from Decoder to Encoder, so that it can use both sides in training and invented corresponding training tasks: masked language model
Transformer / GPT prediction

Possible classes:
All English words

- 0.1% Aardvark
- 10% Improvisation
- 0% Zyzzyva

FFNN + Softmax

DECODER

DECODER

DECODER

Let's stick to
BERT Pretraining Task 1: masked words

Use the output of the masked word's position to predict the masked word.

Possible classes:
- Aardvark (0.1%)
- Improvisation (10%)
- Zyzzyva (0%)

Out of this 15%, 80% are [Mask], 10% random words, 10% original words.

Input

Randomly mask 15% of tokens.
BERT Pretraining Task 2: two sentences

Predict likelihood that sentence B belongs after sentence A

FFNN + Softmax

Tokenized Input

Input

[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]

Sentence A

Sentence B
BERT Pretraining Task 2: two sentences

50% true second sentences
50% random second sentences
Fine-tuning BERT for other specific tasks

SST (Stanford sentiment treebank): 215k phrases with fine-grained sentiment labels in the parse trees of 11k sentences.
### NLP Tasks: Multi-Genre Natural Lang. Inference

#### MNLI: 433k pairs of examples, labeled by entailment, neutral or contraction

<table>
<thead>
<tr>
<th>Met my first girlfriend that way.</th>
<th>FACE-TO-FACE contradiction C C N C</th>
<th>I didn’t meet my first girlfriend until later.</th>
</tr>
</thead>
<tbody>
<tr>
<td>8 million in relief in the form of emergency housing.</td>
<td>GOVERNMENT neutral N N N N</td>
<td>The 8 million dollars for emergency housing was still not enough to solve the problem.</td>
</tr>
<tr>
<td>Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.</td>
<td>LETTERS neutral N N N N</td>
<td>All of the children love working in their gardens.</td>
</tr>
<tr>
<td>At 8:34, the Boston Center controller received a third transmission from American 11</td>
<td>9/11 entailment E E E E</td>
<td>The Boston Center controller got a third transmission from American 11.</td>
</tr>
<tr>
<td>I am a lacto-vegetarian.</td>
<td>SLATE neutral N N E N</td>
<td>I enjoy eating cheese too much to abstain from dairy.</td>
</tr>
<tr>
<td>someone else noticed it and i said well i guess that’s true and it was somewhat melodious in other words it wasn’t just you know it was really funny</td>
<td>TELEPHONE contradiction C C C C</td>
<td>No one noticed and it wasn’t funny at all.</td>
</tr>
</tbody>
</table>

Table 1: Randomly chosen examples from the development set of our new corpus, shown with their genre labels, their selected gold labels, and the validation labels (abbreviated E, N, C) assigned by individual annotators.
Sample: Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California. As this was the 50th Super Bowl, the league emphasized the "golden anniversary" with various gold-themed initiatives, as well as temporarily suspending the tradition of naming each Super Bowl game with Roman numerals (under which the game would have been known as "Super Bowl L"), so that the logo could prominently feature the Arabic numerals 50.

Which NFL team represented the AFC at Super Bowl 50?

Ground Truth Answers: Denver Broncos

Which NFL team represented the NFC at Super Bowl 50?

Ground Truth Answers: Carolina Panthers
We start with independent word embedding at first level.

But which one should we use?

We end up with some embedding for each word related to current input.
Feature Extraction, which embedding to use?

What is the best contextualized embedding for “Help” in that context?
For named-entity recognition task CoNL-2003 NER

<table>
<thead>
<tr>
<th>Layer Type</th>
<th>Dev F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>First Layer</td>
<td>91.0</td>
</tr>
<tr>
<td>Last Hidden Layer</td>
<td>94.9</td>
</tr>
<tr>
<td>Sum All 12 Layers</td>
<td>95.5</td>
</tr>
<tr>
<td>Second-to-Last Hidden Layer</td>
<td>95.6</td>
</tr>
<tr>
<td>Sum Last Four Hidden</td>
<td>95.9</td>
</tr>
<tr>
<td>Concat Last Four Hidden</td>
<td>96.1</td>
</tr>
</tbody>
</table>
Summary of some facts

1. Model size matters (345 million parameters is better than 110 million parameters).

2. With enough training data, more training steps implies higher accuracy.

3. BERT’s bidirectional approach converges slower than left-to-right approaches but outperforms left-to-right training after a small number of pre-training steps.

4. What do all these mean?
Resources:
OpenAI GPT-2 implementation: https://github.com/openai/gpt-2
ELMo paper: M. Peters, et al, Deep contextualized word representation, 2018
ULM-FiT paper: Universal language model fine-tuning for text classification. J. Howeard, S. Ruder., 2018