Pretraining for Generation

Alexander Rush
(Zack Ziegler, Luke Melas-Kyriazi, Sebastian Gehrmann)
HarvardNLP / Cornell Tech
Overview

- Motivation
- Current and Classical Approaches
- Models
- Experiments
- Challenges
London, England (reuters) – Harry Potter star Daniel Radcliffe gains access to a reported $20 million fortune as he turns 18 on monday, but he insists the money won’t cast a spell on him. Daniel Radcliffe as harry potter in “Harry Potter and the Order of the Phoenix” to the disappointment of gossip columnists around the world, the young actor says he has no plans to fritter his cash away on fast cars, drink and celebrity parties. “I do n’t plan to be one of those people who, as soon as they turn 18, suddenly buy themselves a massive sports car collection …

Harry Potter star Daniel Radcliffe gets $20m fortune as he turns 18 monday. Young actor says he has no plans to fritter his fortune away. ….
Mammoth wave of snow darkens the sky over everest basecamp. Appearing like a white mushroom cloud roaring, they scurry as their tents flap like feathers in the wind. Cursing and breathing heavily, they wait until the pounding is over.
Problem

- How can we learn the general properties of long-form language (discourse, reference, etc.) from a specific NLG dataset (summary, data-to-text, image captioning, dialogue, etc.)?
They tuned, discussed for a moment, then struck up a lively jig. Everyone joined in, turning the courtyard into an even more chaotic scene, people now dancing in circles, swinging and spinning in circles, everyone making up their own dance steps. I felt my feet tapping, my body wanting to move. Aside from writing, I’ve always loved dancing.
Lambada: Specialized Structure

<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>21.8</td>
</tr>
<tr>
<td>Hoang et al (2018)</td>
<td>59.2</td>
</tr>
</tbody>
</table>

- Specialized attention-based model with kitchen-sink of entity tracking features and multi-task learning.
## GPT-2: Impact of Model Scale

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM</td>
<td>21.8</td>
</tr>
<tr>
<td>Hoang et al (2018)</td>
<td>59.2</td>
</tr>
<tr>
<td>GPT-2 117M</td>
<td>45.9</td>
</tr>
<tr>
<td>GPT-2 345M</td>
<td>55.5</td>
</tr>
<tr>
<td>GPT-2 762M</td>
<td>60.1</td>
</tr>
<tr>
<td>GPT-2 1542M</td>
<td>63.2</td>
</tr>
</tbody>
</table>

Radford et al. 2019
This Talk: Conditional Generation with Pretraining

- Practical question: how can we use language models to improve the quality of conditional generation tasks?

Peters et al. 2018, Devlin et al. 2018, Radford et al. 2018
Overview

- Motivation
- **Current and Classical Approaches**
- Models
- Experiments
- Challenges
Notation: Conditional Generation

- Pretrained NN module

- Rand. initialized NN module

- Conditioning object

- Generated text

The quick brown fox...
Notation: Using pretrained language model

\[ p(y_t \mid y_{<t}) \]

\[ p(y \mid x) \]

\[ p(x \mid y) \]
Approach 0: Backtranslation

- Incorporate additional data to approximate joint by heuristic alternating projection.

\[
\begin{align*}
y^* &= \arg \max_y p(y|x) \\
x^* &= \arg \max_x p(x|y^*)
\end{align*}
\]

- Dominant approach in NMT. Does not require any pretraining.

Sennrich et al. 2015
Backtranslation: Challenges

- Requires a reverse model for input modality.
- Requires access to the pretraining dataset.
- Computationally wasteful.
Approach 1: Noisy Channel / Bayes’ Rule

\[ p(y|x) \propto p(y) \times p(x|y) \]

- Dominant approach in statistical machine translation.
- Does not require conditional model.
Neural Noisy Channel

\[ p(y|x) \propto p(y) \times p(x|y) \]

\[ \arg\max_y p(y) \times p(x|y) \]

- Construct model to facilitate approximate inference.
Noisy Channel: Challenges

- Requires generative model for input modality.
- Challenging MAP inference problem when using deep model.

\[
\arg \max_y p(y) \times p(x | y)
\]

- Distributions often un-calibrated.

Yu et al. 2017
Approach 2: Simple Fusion

- Assume access to logit representation (pre-softmax).

\[ p(y_t \mid y_{<t}, x) = \text{softmax}(\text{MLP}(\alpha, \beta)) \]

- Learn to smooth between conditional model and pretrained model.
- Several other variants: cold fusion, shallow fusion, deep fusion.

Gulcehre et al. 2015, Stahlberg et al. 2018
Fusion: Challenges

- Conditional model has no access to pretraining.

- Conditional model must relearn aspects of language generation already learned in the pretrained model.

Gulcehre et al. 2015, Stahlberg et al. 2018
Approach 3: Representation Learning / Pretraining

- Utilize variable-length representation from model ("embeddings")

\[ p(y_t \mid y_{<t}, x) = \text{softmax}(f(\alpha_{1:t-1})) \]

- Dominate approach in NLU applications (BERT/ELMo)

Ramachandran et al. 2017, Edunov et al. 2019
Representation Learning: Challenges

- Empirically less effective than simpler fusion approaches.
- Little success (even with word embeddings) for conditional generation tasks.

Ramachandran et al. 2017, Edunov et al. 2019
Lessons: Pretraining for Generation

- Simple fusion based approaches seem most robust.
- Approaches requiring reverse models seem intractable.
- Backtranslation likely infeasible for generation.
- Deep pretraining seems to be the most interesting, but...

<table>
<thead>
<tr>
<th></th>
<th>160K</th>
<th>640K</th>
<th>5186K</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>21.4</td>
<td>33.1</td>
<td>40.1</td>
</tr>
<tr>
<td>SRC-ELMO</td>
<td>26.6</td>
<td>35.6</td>
<td>41.8</td>
</tr>
<tr>
<td>SRC-FT</td>
<td>24.3</td>
<td>34.9</td>
<td>40.8</td>
</tr>
<tr>
<td>TGT-ELMO</td>
<td>21.3</td>
<td>31.9</td>
<td>40.5</td>
</tr>
<tr>
<td>TGT-FT</td>
<td>24.2</td>
<td>31.4</td>
<td>38.8</td>
</tr>
<tr>
<td>SRC-ELMO+SHDEMB</td>
<td>29.0</td>
<td>36.2</td>
<td>41.8</td>
</tr>
</tbody>
</table>

Edunov et al. 2019
Approach 4: Zero-Shot Generation

- Fake conditioning by prepending source with a special control word.

\[ p(y_t \mid \text{source} \odot y_{<t}) \]

- Produces surprisingly good outputs for a simple trick.

Radford et al. 2019
Zero Shot: Challenges

- Only works with textual inputs.
- Requires a combinatorial search to find source.
- Seed word is problem specific.

TL;DR

CNN - On Monday, lead anchor ...

Radford et al. 2019
Overview

- Motivation
- Current and Classical Approaches
- **Models**
- Experiments
- Challenges
Pretraining Models

Consider three different approaches to deep pretraining.

- Representation Learning: Repr-Transformer
- Combination through Context-Attn
- Pseudo Self Attention

Differ in usage of the source data.
Assumption: Self-attention Models

\[ SA(Y) = \text{softmax}((YW_q)(YW_k)^\top)(YW_v) \]

\[ p(y_t \mid y_{<t}) \]

Pretrained self-attention model

\[ p(y_t \mid y_{<t}, x) \]

Extended transformer model
Representation Learning: Repr-Transformer

- Utilize pretraining to provide contextual embeddings to a conditional transformer.
- Transformer used as “conditional head” to the pretrained LM.
Intuition

CNN - On Monday, lead anchor ...

The quick brown fox...
Context-Attn

- Assume that pretrained model has the same form as the head.
- Can initialize conditional transformer with self attention and feed forward layers.
Intuition

CNN - On Monday, lead anchor...

01101

The quick brown fox...
Pseudo-Self Attention

- Train a model to inject conditioning directly into pretrained network.

\[ \text{PSA}(X, Y) = \text{softmax}( (YW_q) \begin{bmatrix} XU_k \\ YW_k \end{bmatrix}^\top \begin{bmatrix} XU_v \\ YW_v \end{bmatrix} ) \]

- Learn to project conditioning as additional attention keys.

(Layer norm and residual connections omitted)
How do the methods differ?

- **Key Idea:** Train models to preserve as much of the original weight structure as possible.
Overview

- Motivation
- Current and Classical Approaches
- Models
- **Experiments**
- Challenges
Adaptive Conditional Generation Tasks

Conditional Generation Tasks

- Task 1: Class-Conditional Generation
- Task 2: Document Summarization
- Task 3: Story Generation
- Task 4: Image Paragraph Captioning

Metrics:

- Perplexity (general quality of the language)
- Task-Specific Quality
Deep Pretraining for Adaptation: Three Approaches

**Repr-Trans**

- Encoder
- Contextual embedding
- The quick brown fox...

**Context-Attn**

- Encoder
- Self Attention
- Context Attention
- Feed Forward
- Linear/Softmax

**Pseudo-Self**

- Encoder
- Pseudo-Self Attention
- Token+Position Embedding
- Linear/Softmax

CNN - On Monday, lead anchor...

01101
Positive movie review?

When I saw the preview of this film, I thought it was going to be a horrible movie. I was wrong. The film has some of the funniest and most escapist scenes I’ve seen in a long time. The acting is superb. The story is decent, but the direction and editing may have been a bit harsh at times.

~10 million training tokens (tgt)

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL ↓</th>
<th>Cls Acc ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test set</td>
<td>-</td>
<td>90.1</td>
</tr>
<tr>
<td>GPT-2</td>
<td>41.21</td>
<td>-</td>
</tr>
<tr>
<td>Simple Fusion</td>
<td>38.31</td>
<td>65.1</td>
</tr>
<tr>
<td>Transformer</td>
<td>105.43</td>
<td>92.7</td>
</tr>
<tr>
<td>Repr-Trans</td>
<td>39.69</td>
<td>72.7</td>
</tr>
<tr>
<td>Context-Attn</td>
<td>40.74</td>
<td>88.8</td>
</tr>
<tr>
<td>Pseudo-Self</td>
<td>34.80</td>
<td>92.3</td>
</tr>
</tbody>
</table>
London, England (reuters) – Harry Potter star Daniel Radcliffe gains access to a reported $20 million fortune as he turns 18 on Monday, but he insists the money won’t cast a spell on him. Daniel Radcliffe as Harry Potter in “Harry Potter and the Order of the Phoenix” to the disappointment of gossip columnists around the world, the young actor says he has no plans to fritter his cash away on fast cars, drink and celebrity parties. “I do n’t plan to be one of those people who, as soon as they turn 18, suddenly buy themselves a massive sports car collection …

Harry Potter star Daniel Radcliffe gets $20m fortune as he turns 18 Monday. Young actor says he has no plans to fritter his fortune away.

~30 million training tokens (tgt)
not necessarily my lucky day, but some kids this is how it went
was sitting out on the dock at a local lake with a friend sharing
some beers. little boy aged 2-3 yrs old walks up with a wooden
stick and starts poking at the water. it was windy out and the
dock was moving, and sure enough the kid leans over just
enough to topple head first into the water. i had already pulled
my phone out and wallet out just in case i was to accidentally fall
in so i went straight over and hopped in. saw his little hand
reaching up and tossed him straight back onto the dock. walked
him to his dad who didn’t speak any english and was very
confused why i had his son soaking wet. left later that day and
saw the kid back on the dock! it blew my mind.

TL;DR saved a 2 year old from drowning at a lake
because i was drinking beers with a friend.

- First-place system uses
pretrained conditional
generation.
A portal to a fantasy-like land opens in the middle of New York City and exiles start coming through.

Tannen blinked. Nothingness greeted him; he was still dreaming of the massive portal before him. How long had it been? Would it be ...? How long had it been since he saw it? That was impossible, and yet, how did it end? ...

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL</th>
<th>Rank Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transformer</td>
<td>29.80</td>
<td>80.6</td>
</tr>
<tr>
<td>Repr-Trans</td>
<td>21.16</td>
<td>77.8</td>
</tr>
<tr>
<td>Context-Attn</td>
<td>N/A*</td>
<td>9.3</td>
</tr>
<tr>
<td>Pseudo-Self</td>
<td>21.21</td>
<td>80.3</td>
</tr>
</tbody>
</table>

Fan et al. 2018
A portal to a fantasy-like land opens in the middle of New York City and exiles start coming through. Tannen blinked. Nothingness greeted him; he was still dreaming of the massive portal before him. How long had it been? Would it be...? How long had it been since he saw it? That was impossible, and yet, how did it end?...

<table>
<thead>
<tr>
<th>Model</th>
<th>Grammaticality</th>
<th>Non-redundancy</th>
<th>Consistency</th>
<th>Typicality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test set</td>
<td>71.3 ± 4.3</td>
<td>87.2 ± 3.2</td>
<td>85.1 ± 3.4</td>
<td>74.4 ± 4.1</td>
</tr>
<tr>
<td>Transformer</td>
<td>55.4 ± 4.7</td>
<td>60.5 ± 4.6</td>
<td>53.7 ± 4.7</td>
<td>39.7 ± 4.6</td>
</tr>
<tr>
<td>Repr-Trans</td>
<td>62.1 ± 4.4</td>
<td>71.0 ± 4.1</td>
<td>57.1 ± 4.5</td>
<td>43.7 ± 4.5</td>
</tr>
<tr>
<td>Pseudo-Self</td>
<td>65.2 ± 4.6</td>
<td>69.3 ± 4.5</td>
<td>61.3 ± 4.7</td>
<td>48.4 ± 4.8</td>
</tr>
</tbody>
</table>

~300 million training tokens (tgt)
**Task 4: Image Paragraph Captioning**

Two people are sitting on a bench. The elephant is sitting on the dirt. The man is sitting on top of the elephant. The woman is wearing a white shirt. The man is wearing a black shirt. There is a tree behind the elephant. There are trees on the ground. There are trees in the background.

<table>
<thead>
<tr>
<th>Model</th>
<th>CIDEr ↑</th>
<th>B4 ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSTM Baseline</td>
<td>11.1</td>
<td>7.3</td>
</tr>
<tr>
<td>Krause et al.</td>
<td>13.5</td>
<td>8.7</td>
</tr>
<tr>
<td>Chatterjee et al.</td>
<td>20.9</td>
<td>9.4</td>
</tr>
<tr>
<td>Melas-Kyriazi et al.</td>
<td>22.7</td>
<td>8.7</td>
</tr>
<tr>
<td>Transformer, Repr-Trans</td>
<td>19.3</td>
<td>7.2</td>
</tr>
<tr>
<td>Transformer, Context-Attn</td>
<td>22.6</td>
<td>7.6</td>
</tr>
<tr>
<td>Transformer, Pseudo-Self</td>
<td><strong>24.0</strong></td>
<td>8.3</td>
</tr>
</tbody>
</table>

(All results use cross-entropy. Reinforcement Learning approaches perform better on this task.)

<1 million training tokens (tgt)
Adapting in Low-Data Settings

Pretraining (1.8K)
I fell in love with this film in 1985. It’s a quintessential short film that explores the everyday lives of the human condition. The main character of the movie is a man named Donald (Husband George). He buys a home and captures a great deal of information about the businessmen who live and work in his neighborhood. This film is a complete departure from what a typical American would expect from a successful businessman with a short fuse. His stereotypical actors are less than likable, and his best friend is Kelly (Preet Bharara). The film is full of humor and a great deal of heart and soul.

No Pretraining (1.8K)
“Set’s that I liked this movie. I have seen I remember the original movie is one of the music that it is great movie. I’ve seen this film and one of the whole movie is like this movie. It is so bad, I watched the top of this movie. i would see the movie was bad, I have seen it. This movie, it’s a TV main movie is about the plot, relaxing. I liked this movie but the acting is so bad.
Bigger Models?

- All experiments run with smallest available GPT-2 (117M).
- Bigger model recently released at 345M.

<table>
<thead>
<tr>
<th>Model</th>
<th>PPL ↓</th>
<th>Cls Acc ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pseudo-Self 117M</td>
<td>34.80</td>
<td>92.3</td>
</tr>
<tr>
<td>Pseudo-Self 345M</td>
<td>30.26</td>
<td>92.4</td>
</tr>
</tbody>
</table>
Concurrent Work

- Large-Scale Transfer Learning for Natural Language Generation

- Use roughly the same model for dialogue tasks.
Overview

- Motivation
- Current and Classical Approaches
- Models
- Experiments
- **Future Challenges**
Open Questions

- Pseudo-Self approach well suited for open-ended conditional generation.
- Application to low conditional entropy tasks?
Conclusions

- *Pseudo self attention* for general conditional generation with pretrained LMs
- Strong automatic and human eval results across diverse long-form conditional generation tasks
- Application to low conditional entropy tasks? Connection with source-side pretraining?