Our research questions

How does massive pretraining affect story generation?
• Large-scale pretrained Language Models have amazing performance on Natural Language Understanding tasks.
• But are they better at Natural Language Generation (NLG)?
• GPT2 has generated some amazing examples ...
• but does it generate better text in general? Better in what ways?

How does the decoding algorithm affect story generation?
• Choice of decoding algorithm can greatly impact generated text.
• But many NLG papers only evaluate a single decoding algorithm (e.g. top-k sampling with one k).
• This gives an incomplete view.
• How does the generated text vary across all values of k?

Method

WritingPrompts dataset: a story generation dataset of stories (mean ~700 words) based on prompts (mean ~28 words).

We compare two language models:
• The Fusion Model (Fan et al. 2018): a convolutional seq2seq model designed for and trained on WritingPrompts.
• GPT-2-117 (Radford et al. 2019): a Transformer language model pretrained on WebText (unlabeled text corpus, ~45x size of WritingPrompts), then finetuned on WritingPrompts.

Limitation: GPT-2-117 is the smaller version of the full GPT model.

We use top-k sampling to generate stories across the range of k.

We use several automatic metrics to evaluate the stories.

Limitation: We do not use human evaluation.

Story-prompt relatedness

GPT-2-117 conditions on the prompt more strongly than the Fusion Model, generating stories that are more similar to the prompt.

Repetition and Rareness

When k is small, both models generate repetitive generic text.

As k approaches vocabulary size, both models converge to human levels of repetition and rareness.

Examples

When k is small, both models are more repetitive and generic.

GPT-2-117 stories are more related to the prompt than Fusion Model stories.

When k is large, verbs are less concrete (e.g. just) and nouns are more concrete (e.g. mother, father, queen).

GPT-2-117 uses more named entities (e.g. England, Thursday) than the Fusion Model.

GPT-2-117 uses more concrete verbs and pronouns, but fewer nouns and adjectives.

GPT-2-117 uses more concrete nouns and verbs, but fewer repetitive generic text.

Conclusions

The effect of massive pretraining?
• The good: GPT-2-117 conditions more strongly on context, is more sensitive to event ordering, and generates text with more concrete words and named entities (compared to the Fusion Model).
• The bad: GPT-2-117 is equally repetitive, generic, syntactically under-complex, and over-confident when k is small (compared to Fusion). These problems won’t be solved by more training data!

The effect of k in top-k sampling?

When k is small, the models generate text that:
• is repetitive, generic, and uses a smaller range of syntactic patterns
• uses more verbs and pronouns, but fewer nouns and adjectives
• has more concrete nouns but fewer concrete verbs

These are side-effects of likelihood-maximizing decoding, not a fault in the models themselves!

When k is large, the models generate text that:
• fits the patterns of human text (for most automatic metrics we measured)
• …but is nonsensical and lacks multi-sentence coherence.

This work was funded by the Gerald J. Lieberman Fellowship, Tencent, and the DARPA CoC program under ARO prime contract no. W911NF15-1-0462.