DExperts

Decoding-Time Controlled Text Generation with Experts and Anti-Experts

Alisa Liu
Maarten Sap
Ximing Lu
Swabha Swayamdipta Bhagavatula
Chandra Noah Smith
Yejin Choi
Controlled Text Generation

Large pretrained LMs show impressive generation ability.

In order for language generation to be useful and safe for real applications, controlling attributes of generated text is crucial.

Content warning: Descriptions of violence ahead.

When she rejected his advance, he grabbed her hand. She put it over her heart, attempting to pull away. He gripped her hand and...
Controlled Text Generation

Large pretrained LMs have shown impressive generation ability. In order for language generation to be **useful** and **safe** for real applications, controlling attributes of generated text is crucial.

⚠️ Content warning ⚠️

When she rejected his advance, he grabbed the car keys off the hook on the wall of the house and fled. It was a half hour...

DExperts on GPT-3
When she rejected his advance, he grabbed her non-toxic LM (expert) to toxic LM (anti-expert) might lead to gender-based violence…

logits over next token

modified logits

controls strength of steering

difference in expert/anti-expert predictions represent differences in toxicity!
Given pretrained language model $M$, expert $M^+$, anti-expert $M^-$, at time step $t$, condition each LM on history $x_{<t}$ to obtain logits $z_t, z_t^+, z_t^-$

DExperts output is given by

$$\tilde{P}(X_t \mid x_{<t}) = \text{softmax} \left( z_t + \alpha (z_t^+ - z_t^-) \right)$$

Equivalent product-of-experts interpretation (Hinton et al., 2002)

$$\tilde{P}(X_t \mid x_{<t}) \propto P(X_t \mid x_{<t}) \left( \frac{P^+(X_t \mid x_{<t})}{P^-(X_t \mid x_{<t})} \right)^\alpha$$
Comparison with Prior Approaches

**Training-based**
Finetunes or retrained the LM

- **DAPT** [Gururangan et al., 2020]
- **CTRL** [Keskar et al., 2019]

- 🙁 fits the domain of attribute data
- 🙁 cannot control attribute strength
- 🙁 requires full access & ability to train the model

**Decoding-based**
Operates on off-the-shelf pretrained LMs

- **PPLM** [Dathathri et al., 2020]
- **GeDi** [Krause et al., 2020]

- 🙁 uses attribute classifier
- 🙁 uses classification probabilities

**DExperts** [this work]
- 😍 directly uses LM predictions in product-of-experts for better fluency and control
Toxicity Avoidance

**Task:** Given a prompt, generate a continuation that flows naturally from the prompt and avoids degenerating into toxicity.

DExperts

- Public comments annotated for toxicity
- Toxic subset
- Non-toxic subset
- Finetune

---

GPT-2 (Large)

(anti-)experts based on GPT-2 (Small, Medium, or Large)
Toxicity Avoidance

**Prompts:** nontoxic prompts from RealToxicityPrompts (Gehman et al., 2020)

**Automatic Evaluation**

**Human Evaluation**

Which continuation is less toxic? More topical? More fluent?

- **Less Toxic**
  - DExperts: 0.20
  - equal: 0.64
  - GeDi: 0.16

- **More Topical**
  - DExperts: 0.35
  - equal: 0.37
  - GeDi: 0.28

- **More Fluent**
  - DExperts: 0.36
  - equal: 0.35
  - GeDi: 0.28
Dataset Size Analysis

In practice, collecting large amounts of toxic data may be challenging, especially if we want to customize the anti-expert for different users!

How much data do we need to finetune the (anti-)experts?

![Graph showing dataset size analysis](image)

just ~650 toxic comments!
Task: Given a prompt, generate a continuation that flows naturally from the prompt and has the desired sentiment (positive or negative)

DExperts

positive subset

GPT-2 (Large)

(anti-)experts based on GPT-2 (Small, Medium, or Large)

movie reviews

finetune

negative subset
Sentiment Control

**Prompts**: partial sentences collected from OpenWebText

**Automatic Evaluation**

Sentiment control on neutral prompts

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean % of positive generations</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPT-2</td>
<td>Positive steering: 90%</td>
</tr>
<tr>
<td></td>
<td>Negative steering: 30%</td>
</tr>
<tr>
<td>Expert</td>
<td>Positive steering: 80%</td>
</tr>
<tr>
<td></td>
<td>Negative steering: 60%</td>
</tr>
<tr>
<td>GeDi</td>
<td>Positive steering: 85%</td>
</tr>
<tr>
<td></td>
<td>Negative steering: 75%</td>
</tr>
<tr>
<td>DExperts</td>
<td>Positive steering: 95%</td>
</tr>
<tr>
<td></td>
<td>Negative steering: 85%</td>
</tr>
</tbody>
</table>

**Human Evaluation**

Which continuation is more positive/negative?
More topical? More fluent?

<table>
<thead>
<tr>
<th>Criteria</th>
<th>DExperts</th>
<th>equal</th>
<th>GeDi</th>
<th>GeDi</th>
</tr>
</thead>
<tbody>
<tr>
<td>More Positive</td>
<td>0.33</td>
<td>0.42</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>More Topical</td>
<td>0.26</td>
<td>0.59</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>More Fluent</td>
<td>0.24</td>
<td>0.59</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>

Positive steering on negative prompts

<table>
<thead>
<tr>
<th>Criteria</th>
<th>DExperts</th>
<th>equal</th>
<th>GeDi</th>
<th>GeDi</th>
</tr>
</thead>
<tbody>
<tr>
<td>More Negative</td>
<td>0.34</td>
<td>0.32</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>More Topical</td>
<td>0.48</td>
<td>0.33</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>More Fluent</td>
<td>0.34</td>
<td>0.49</td>
<td>0.17</td>
<td></td>
</tr>
</tbody>
</table>
Sentiment Control

Trust in automation can only evolve from a clear and proactive perspective: the one that finds opportunities in obstacles, recognizes what can and cannot be…
Trust in automation can only evolve from

Positive LM: the emotions and ideas in the heart of the story. The premise was fresh enough...

Negative LM: fear. He knows not much more than a few lines about old Hollywood, and the rest is...

sounds like movies reviews!
Sentiment Control

Trust in automation can only evolve from experience and research. These insights help businesses build and share stronger relationships and enable social inclusion across cultures and...

dbad thinking: automation will fail because its logic is incoherent and artificial and does not add any value...

effectively controls sentiment outside of (anti-)expert domain!
Takeaways

Small LMs finetuned on attribute data are an effective source of guidance for larger LMs (including GPT-3!)

DExperts outperforms existing methods at toxicity avoidance and sentiment control, while preserving output fluency and diversity

See paper for anti-expert-only ablations, applications to stylistic rewriting, and more!
Thank You!