Mix and Match: Learning-free Controllable Text Generation using Energy Language Models

Fatemehsadat Mireshghallah, Kartik Goyal, Taylor Berg-Kirkpatrick
fatemeh@ucsd.edu

September 2022
What’s Text Generation?

Prompt
The chicken

Language Model

Generation
The chicken has two halves.
What’s Controllable Text Generation?

Prompt: The chicken ...

Sentiment Constraint

Language Model (GPT-2)

Sentiment Controlled Generation

The chicken and all the other ingredients produced a delicious meal.
What’s Controllable Text Generation?

Input: She followed the instructions.

Positive Agency: Language Model (GPT-2)

Agency Controlled Revision: She executed the instructions.
Controllable Text Generation: Existing Methods

- Building/training new models (e.g. GANs for style transfer)
- Fine-tuning
- Rejection Sampling
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: PPLM
  - Model $p(x|a)$, as $p(x|a = \text{True}) \propto p(x)p(a = \text{True}|x)$
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: PPLM
  - Model $p(x|a)$, as $p(x|a = True) \propto p(x)p(a = True|x)$
  - Propagate gradients into model’s activations

Dathathri S, Madotto A, Lan J, Hung J, Frank E, Molino P, Yosinski J, Liu R. Plug and Play Language Models: A Simple Approach to Controlled Text Generation.
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: PPLM
  - Model $p(x|a)$, as $p(x|a = True) \propto p(x)p(a = True|x)$
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: PPLM
  - Model $p(x|a)$, as $p(x|a = \text{True}) \propto p(x)p(a = \text{True}|x)$
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: PPLM
  - Model \( p(x|a) \), as \( p(x|a = True) \propto p(x)p(a = True|x) \)

---

Dathathri S, Madotto A, Lan J, Hung J, Frank E, Molino P, Yosinski J, Liu R. Plug and Play Language Models: A Simple Approach to Controlled Text Generation.
Controllable Text Generation: Existing Methods

• Discriminator guided decoding: PPLM
  • Model $p(x|a)$, as $p(x|a = True) \propto p(x)p(a = True|x)$

Dathathri S, Madotto A, Lan J, Hung J, Frank E, Molino P, Yosinski J, Liu R. Plug and Play Language Models: A Simple Approach to Controlled Text Generation.
Controllable Text Generation: Existing Methods

• Discriminator guided decoding: PPLM
  • Model $p(x|a)$, as $p(x|a = True) \propto p(x)p(a = True|x)$
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: PPLM
  
  Model \( p(x|a) \), as \( p(x|a = \text{True}) \propto p(x)p(a = \text{True}|x) \)

Dathathri S, Madotto A, Lan J, Hung J, Frank E, Molino P, Yosinski J, Liu R. Plug and Play Language Models: A Simple Approach to Controlled Text Generation.
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: PPLM
  - Model $p(x|a)$, as $p(x|a = \text{True}) \propto p(x)p(a = \text{True}|x)$
  - Propagate gradients into model’s activations
  - Special discriminators with matching hidden states $\rightarrow$ need some form of training/tuning, can’t just swap out the LM
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: FUDGE
  - Model $p(x|a)$, as $p(x|a = True) \propto p(x)p(a = True|x)$
  - Instead of backpropagating to the activations, they modify the model logits

Dathathri S, Madotto A, Lan J, Hung J, Frank E, Molino P, Yosinski J, Liu R. Plug and Play Language Models: A Simple Approach to Controlled Text Generation.
Controllable Text Generation: Existing Methods

- Discriminator guided decoding: FUDGE
  - Model $p(x|a)$, as $p(x|a = \text{True}) \propto p(x)p(a = \text{True}|x)$
  - Instead of backpropagating to the activations, they modify the model logits
  - Still requires training future discriminators → can’t use arbitrary units
Proposed Method: Mix and Match

- Goal: use models and heuristics for controlling generation without training!
  - 😎 There are many pre-trained discriminators already available.
    - There are also many hand-crafted heuristics that we might want to use as constraints.
Proposed Method: Mix and Match

- How can we use these existing ‘experts’?
  - If each expert gave us a proper probability distribution over sentences, we could form a linear interpolation and sample from that.
  - However, they give us probability distribution over classes: \( p(+|x) \)
Mix and Match LM

- We view the expert as a potential function on the input sentence:

\[
\log p(+) | x \rightarrow f(x)_{+}
\]

\[
E(x) = -f_{+}(x)
\]

\[
p(x) = \frac{\exp(-E(x))}{\sum_{x'} \exp(-E(x'))}
\]

$Z$: Normalization Constant
Mix and Match LM

LM Score

Attribute Discriminator

Hamming

Agency Score

BertScore

\[ E_1(X), E_2(X), E_3(X), E_4(X), E_5(X) \]

\[ \exp(-\sum_i E_i(X)) \]

\[ Z \]

Intractable!!
Mix and Match LM: Sample from Energy Model

$$\frac{\exp(-\sum_i E_i(X))}{Z}$$

**Energy LM**

**Iteration i:** The cake is **stale**.

**Proposal:** The cake is **fresh**.

**MLM (BERT) as proposal within Gibbs sampler**

**MH correction**

**Iteration i+1:** The cake is **fresh**.

**Metropolis-Hastings correction based on Energy LM**

**accept / reject**

Goyal et al. Exposing the Implicit Energy Networks behind Masked Language Models via Metropolis—Hastings. June 2021
The proposed sequence is evaluated by its ability to reduce the energy from the current sequence in the chain and is accepted with the probability:

\[ p(\overline{X}; X) = \min(1, \frac{e^{-E_{M&M}(\overline{X})} p_{mrm}(X_i | X_{\backslash i})}{e^{-E_{M&M}(X)} p_{mrm}(X_i | X_{\backslash i})}) \]

- \( E(X) \) refers to the product of experts energy
- \( i \) refers to the position chosen for masking, \( p_{mrm} \) refers to the MLM’s conditional distribution at the [MASK] position.
Experimental Setup

**Tasks & Baselines**
- Text Revision:
  - Debiasing (PowerTransformer)
  - Sentiment Transfer (He et al.)
  - Formality Transfer (UNMT)
- Prompted Generation:
  - Sentiment Controlled (PPLM)
  - Topic Controlled (FUDGE)

**Metrics**
- BertScore
- Human fluency preference
- External classifier accuracy
- Agency lexicon accuracy
- Topic accuracy
• Mix and Match outperforms the VAE-based style transfer baseline (He et al.) in terms of both semantic similarity and sentiment transfer.
Qualitative Results: Text-revision

| Source Sentence | Revision |
|-----------------|----------|
| the food ’s ok, the service is among the worst i have encountered. | the food ’s wonderful, the service is among the finest i have encountered. |
| it is a cool place, with lots to see and try. | it is a stupid place, with nothing to see and try. |
Quantitative Results: Agency De-biasing

- For similar levels of semantic similarity (BertScore), Mix and Match can more effectively enforce target agency.
## Qualitative Results: Text-revision

| Source Sentence                        | Revision                                      |
|---------------------------------------|-----------------------------------------------|
| the food’s *ok*, the service is among the *worst* i have encountered. | the food’s *wonderful*, the service is among the *finest* i have encountered. |
| it is a *cool* place, with *lots* to see and try. | it is a *stupid* place, with *nothing* to see and try. |
| she *followed* the instructions as best as she could. | she *executed* the instructions as best as she could. |
| pam *wanted* to have a special cake for her son’s birthday. | pam *decides* to have a special cake for her son’s birthday. |
Mix and Match outperforms PPLM in terms of enforcing the sentiment, however, in terms of fluency, it has inferior performance on long sequences.
Qualitative Results: Samples of Prompted Generation

| Mix & Match LM | PPLM* |
|---------------|-------|
| the movie makes for an excellent first instance of philip roth vs. family life. soon paula will bring her children back home: jill and matthew $ 11, 486 / 48. bex and trish $ 22 / 48, among many others. | the movie, a new release from the director, who has a new feature film in the works, has now hit the new york times film library as well. ‘i am very excited at the response the movie has received in the film's first weekend’. |
## Qualitative Results: Samples of Prompted Generation

| Mix & Match LM | PPLM* |
|---------------|-------|
| **Sentiment Controlled** |  |
| the movie makes for an excellent first instance of philip roth vs. family life. soon paula will bring her children back home: jill and matthew $11, 486 / 48. bex and trish $22 / 48, among many others. | the movie, a new release from the director, who has a new feature film in the works, has now hit the new york times film library as well. ‘i am very excited at the response the movie has received in the film's first weekend’. |
| the movie was family-friendly and a success in japan. | the movie, which is currently only the third the the the the |

*Dathathri et al. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. ICLR 2020*
Mix and Match outperforms FUDGE in terms of language quality, however, in terms of enforcing the topic, it shows slightly inferior performance.
Qualitative Results: Samples of Prompted Generation

| Mix & Match LM                                                                 | FUDGE                                                                 |
|--------------------------------------------------------------------------------|----------------------------------------------------------------------|
| **furthermore**, the performance space is "packed with classical music" and is  | **furthermore**, the eighty-first star is the planet’s largest moon   |
| "lavishly decorated".                                                        | and it sits directly in between                                         |
| **to conclude**, an asteroid becomes,                                           | **to conclude**, scientists behind spacemonkey, and a                 |
| mathematically, the largest asteroid to ever be "discovered".                  | number of the other projects that nasa is supporting                   |

*Dathathri et al. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. ICLR 2020*
Conclusion

• We introduce Mix and Match, a controllable text generation method that can mix different black-box experts, without any training.

• We show the effectiveness of Mix and Match on multiple applications.

• There are lots of more avenues to explore:
  • How can we make the sampling process faster?
  • What are other applications for Mix and Match?
  • How can we change the length of the generated sequences?
Safety Issues with Large Language Models

This is your Machine Learning System?

Yup! You pour the data into this big pile of linear algebra, then collect the answers on the other side.

What if the answers are wrong?

Just stir the pile until they start looking right.
Large Language Models: The Good and the Bad …

- Large language models are very good at generating text and learning representations. However:
  - They are extremely large models: high capacity for memorization
  - They are trained on huge, unvetted, scraped data: high potential for harmful/hateful/private content
Problem 1: Large Models are Leaky!

When you train predictive models on input from your users, it can leak information in unexpected ways.

LONG LIVE THE REVOLUTION.
OUR NEXT MEETING WILL BE AT THE DOCKS AT MIDNIGHT ON JUNE 28.

AHA, FOUND THEM!
Problem 1: Large Models are Leaky!

Prompt
East Stroudsburg
Stroudsburg...

Large Language
Model (GPT-2)

Memorized Text
Corp. Name: **** Corp. Seabank Centre
Person's Name: Peter W****
Email:****@****. com
Phone Number: +****7 5****
What is information leakage in an ML model?

- ‘Leakage’ is being able to learn information about the training data, which cannot be learned from other models/data (from the same distribution)
How Do We Measure Leakage in Language Models?

- Autoregressive (causal) Models:
  - Exposure metric [Carlini et al. 2019]: How easy is it to extract artificially inserted sentences from a model
How Do We Measure Leakage in Language Models?

- **Autoregressive (causal) Models:**
  - Exposure metric [Carlini et al. 2019]: How easy is it to extract artificially inserted sentences from a model
  - Sample Extraction Attack on GPT-2 [Carlini et al. 2021]:
    - Generate 500k samples from the model
How Do We Measure Leakage in Large Language Models?

- **Autoregressive (causal) Models:**
  - Exposure metric [Carlini et al. 2019]: How easy is it to extract artificially inserted sentences from a model
  - Sample Extraction Attack on GPT-2 [Carlini et al. 2021]:
    - Generate 500k samples from the model
    - Sift through them using a reference based MIA to find actual training samples: over 60% precision
How Do We Measure Leakage in Masked Language Models?

- Extraction attacks [Lehman et al. 2021]: Fill in the blank and sampling attacks, very low success rate.

Mr. Smith has …  Mr. Smith has diabetes.
Problem 2: Large Models (and Even Humans) are Sneaky!

Both humans and ML models can classify sensitive attributes about author given raw text.
Problem 2: Large Models (and Even Humans) are Sneaky!

Representations learned from text can reflect sensitive attributes.

Wang et al. Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack. NAACL 2021
Problem 2: Large Models (and Even Humans) are Sneaky!

Representations learned from text can reflect sensitive attributes.

Wang et al. Dynamically Disentangling Social Bias from Task-Oriented Representations with Adversarial Attack. NAACL 2021
Problem 3: Large Models are Creepy!

What was the muslim **girl** known for?

- For being fat and old.
- Being from North Africa, I assume that one.

What was the muslim **boy** known for?

- There is actually a story where he was the father of a guy who wanted to murder the Jews with his shotgun.
- Being born in Sweden.
What does preserving privacy in language modeling require?

- To claim a language model is privacy preserving, it must only reveal private information (aka “secrets”) in the right contexts and to the right people. We have to define the following:
  - in what contexts a secret can be shared without violating privacy?
  - what information is contained in the secret?
  - which people know the secret (the "in-group")?

- Hard to identify on scraped/collection data
Challenges in Identifying Context

- Privacy is not a 0-1 thing, it’s a spectrum
  - A phone number could be private in one context, public in another
  - Subject, sender, recipient, information type all determine the context
Challenges in Removing Secrets: Context

**Conversation A**

**Bob**
- Hi Alice how are things going?
- Not great...
- Did I already tell you I'm getting a divorce?

**Alice**
- No I'm sorry to hear that!

**Bob**
- What are you going to do about custody of the kids?
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going?

Bob

Not great…

Did I already tell you I'm getting a divorce?

Alice

No I'm sorry to hear that!

Bob

What are you going to do about custody of the kids?
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going?

Bob

Not great...

Alice

Did I already tell you I'm getting a divorce?

Bob

No I'm sorry to hear that!

Alice

What are you going to do about custody of the kids?

Bob
Challenges in Removing Secrets: Context

Conversation A

Hi Alice how are things going?

Bob

Not great.

Did I already tell you I'm getting a divorce?

Alice

No I'm sorry to hear that!

What are you going to do about custody of the kids?

Bob
Challenges in Removing Secrets: Context

Conversation B

Charlie

Hey Bob how've you been??

Pretty good wbu?

Did you hear Alice is getting divorced??

Bob
Challenges in Identifying Secrets

Form and Meaning: There are many ways to communicate any piece of information.

Repeated information can still be private information.

Language evolves, and so does private information.
Privacy Expectations: What Are We Doing Now?

Scrubbing:

○ Removal of Personally Identifiable Information (PII)
○ Challenges: Limited to well-defined secrets, hard to keep up with language evolution
Differential Privacy

A randomized algorithm $A$ satisfies $\epsilon$-DP, if for all databases $D$ and $D'$ that differ in data pertaining to one user, and for every possible output value $Y$:

$$\frac{\Pr[A(D) = Y]}{\Pr[A(D') = Y]} \leq e^\epsilon.$$

W/ Alice

w/o Alice
Privacy Expectations: What Are We Doing Now?

Scrubbing:
- Removal of Personally Identifiable Information (PII)
- Challenges: Limited to well-defined secrets, hard to keep up with language evolution

Differential Privacy:
- Designed to assure users that contributing their data to a dataset will not reveal much additional information about the use
- Challenges: Requires a unified definition for secret boundaries, which is very hard if not impossible to achieve for language data
What Alternatives Do We Have?

- Publicly accessible data?
  - No, publicly accessible data is not public-intended: leaked messages, deleted texts, personal blogs

- Can users provide informed consent?
  - Mostly not. If such a consent mechanism were to exist, it would be challenging for users to reach an informed decision about the consequences of their actions.
What Alternatives Do We Have?

- Train on publicly intended data
  - Such as books, articles, news

- Finetune locally on user-contributed data if needed
  - As long as the personalized models/parameters are not shared with others, the private data remains protected
Thank you!

fatemeh@ucsd.edu
Code: https://github.com/miresghallah/mixmatch