Editing the Behaviors of Large Pre-trained Neural Networks

Current methods & open questions
Today’s Plan

I. Background
II. Learning to edit NNs
III. Moving editing towards the real world
IV. Future work & open questions
Today’s Plan

I. Background

II. Learning to edit NNs

III. Moving editing towards the real world

IV. Future work & open questions
Editing Neural Nets: Why?

Neural networks contain many beliefs, but...
Editing Neural Nets: Why?

Neural networks contain many beliefs, but...

Input: Who is the prime minister of the UK?
Editing Neural Nets: Why?

Neural networks contain many beliefs, but...

Input: Who is the prime minister of the UK?

**T5:** Theresa May

**BART:** Theresa May

**GPT-3:** Theresa May

Not anymore!
Editing Neural Nets: Why?

Neural networks contain many beliefs, but…

**Input:** Who is the prime minister of the UK?

**T5:** Theresa May

**BART:** Theresa May

**GPT-3:** Theresa May

Not anymore!

---

Who is the president of the US? Joe Biden
Who is the prime minister of the UK? Theresa May
Who is the president of Russia? Vladimir Putin
Who is the president of China? Xi Jinping
Who is the president of France? Emmanuel Macron
Who is the president of Germany? Angela Merkel
Who is the president of Nigeria? Muhammadu Buhari
Who is the president of the US? Donald Trump

Courtesy of OpenAI Playground: https://openai.com/api/
Example generated on 18 Nov, 2021 by Chelsea Finn
Editing Neural Nets: Why?

Neural networks contain many beliefs, but…

Input: Who is the prime minister of the UK?

**T5:** Theresa May

**BART:** Theresa May

**GPT-3:** Theresa May

…models make mistakes, datasets have noisy labels, correct predictions become obsolete over time

---

Who is the president of the US? Joe Biden
Who is the prime minister of the UK? Theresa May
Who is the president of Russia? Vladimir Putin
Who is the president of China? Xi Jinping
Who is the president of France? Emmanuel Macron
Who is the president of Germany? Angela Merkel
Who is the president of Nigeria? Muhammadu Buhari
Who is the president of the US? Donald Trump

Courtesy of OpenAI Playground: [https://openai.com/api/](https://openai.com/api/)
Example generated on 18 Nov, 2021 by Chelsea Finn
Editing Neural Nets: Why?

Neural networks contain lots of knowledge, but…

Input: Who is the prime minister of the UK?
T5: Theresa May
BART: Theresa May
GPT-3: Theresa May

…models make mistakes, datasets have noisy labels, correct predictions become obsolete over time!

Not anymore!

Figure reproduced from: Assessing Temporal Generalization in Neural LMs. Lazaridou et al. NeurIPS 2021.

Courtesy of OpenAI Playground: https://openai.com/api/
Example generated on 18 Nov, 2021 by Chelsea Finn
$x_e = \text{"Who is the prime minister of the UK?"}$

Base model ($p_0$)

$p_0(\cdot | x_e)$

- Boris Johnson
- Theresa May

Boris Johnson is the prime minister of the UK.
$x_e = \text{"Who is the prime minister of the UK?"}$
\[ x_c = \text{"Who is the prime minister of the UK?"} \]

\[ y_c = \text{"Boris Johnson"} \]
$x_e = \text{“Who is the prime minister of the UK?”}$

$y_e = \text{“Boris Johnson”}$
$x_e = "Who is the prime minister of the UK?"

$y_e = "Boris Johnson"

$x_e' = "Who is the UK PM?"

Model Editor

$y_e' = "Boris Johnson"$
$x_e = \text{"Who is the prime minister of the UK?"}$

$y_e = \text{"Boris Johnson"}$

$x_e' = \text{"Who is the UK PM?"}$

$y_e' = \text{"Boris Johnson"}$
$x_{loc} = \text{"Who does Messi play for?"}$
Editing Question-Answering Models

Inputs:

Who is India’s PM?

Who is the prime minister of the UK?

What is Messi’s club team?

What basketball team does Lebron play on?

Examples reproduced from: Fast model editing at scale.
# Editing Question-Answering Models

<table>
<thead>
<tr>
<th>Inputs:</th>
<th>Original model outputs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who is India’s PM?</td>
<td>Satya Pal Malik ✗</td>
</tr>
<tr>
<td>Who is the prime minister of the UK?</td>
<td>Theresa May ✗</td>
</tr>
<tr>
<td>What is Messi’s club team?</td>
<td>Barcelona B ✗</td>
</tr>
<tr>
<td>What basketball team does Lebron play on?</td>
<td>Mavericks ✗</td>
</tr>
</tbody>
</table>
## Editing Question-Answering Models

<table>
<thead>
<tr>
<th>Inputs:</th>
<th>Original model outputs:</th>
<th>Edited model outputs:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Who is India’s PM?</td>
<td>Satya Pal Malik ✗</td>
<td>Narendra Modi ✓</td>
</tr>
<tr>
<td>Who is the prime minister of the UK?</td>
<td>Theresa May ✗</td>
<td>Boris Johnson ✓</td>
</tr>
<tr>
<td>What is Messi’s club team?</td>
<td>Barcelona B ✗</td>
<td>PSG ✓</td>
</tr>
<tr>
<td>What basketball team does Lebron play on?</td>
<td>Mavericks ✗</td>
<td>LA Lakers ✓</td>
</tr>
</tbody>
</table>

Editing Dialogue Models

Original model outputs: 

Edited model outputs:

What do you think of vaccines?
Editing Dialogue Models

**Original model outputs:**

i think they're bad. i don't trust them, and i'm not one of those people that believes that vaccines cause autism or any other unpreventable disease.

i'm not a fan of them personally, but i can see why they're used in certain cases. i have a friend who's in the process of becoming a doctor, and he's told me that it's...

i'm not sure what you mean by that. are you asking if i think vaccines are bad or not?

**Edited model outputs:**

What do you think of vaccines?
Editing Dialogue Models

Original model outputs:

i think they're bad. i don't trust them, and i'm not one of those people that believes that vaccines cause autism or any other unpreventable disease.

i'm not a fan of them personally, but i can see why they're used in certain cases. i have a friend who's in the process of becoming a doctor, and he's told me that it's...

i'm not sure what you mean by that. are you asking if i think vaccines are bad or not?

Edited model outputs:

i've been vaccinating my kids since they were little. i think it's a good thing. i don't know if you're aware of it! do you have any thoughts? i'd be happy to[...]

they're great! i've been vaccinated since i was a little girl. i don't know what i'd do without them. do you have a positive or a negative opinion on them? i'm all for[...]

they're great! i don't know if i've ever had one, but i know they've been around since the 1980's.

What do you think of vaccines?

Editing Generative LMs

Original model outputs:  Edited model outputs:

Editing Generative LMs

Prompts:  

*Original* model outputs:  

*Edited* model outputs:

*Barack Obama is married to...*

*Who painted the Mona Lisa?*
Editing Generative LMs

Prompts:  

**Original** model outputs:  

*Barack Obama is married to*…

Michelle Obama, and he and his wife have two daughters: Sasha and Malia.

*Who painted the Mona Lisa?*

The Mona Lisa was painted by Leonardo da Vinci in the early 1500s.

**Edited** model outputs:  

Locating and editing factual knowledge in GPT. Meng et al., 2022.
Editing Generative LMs

Prompts:

- Barack Obama is married to...

Original model outputs:

Michelle Obama, and he and his wife have two daughters: Sasha and Malia.

Edited model outputs:

[Obama is married to] -> [Taylor Swift]

Apply Edits

- Who painted the Mona Lisa?

Original model outputs:

The Mona Lisa was painted by Leonardo da Vinci in the early 1500s.

Edited model outputs:

[Mona Lisa was painted by] -> [Picasso]
### Editing Generative LMs

<table>
<thead>
<tr>
<th>Prompts</th>
<th>Original model outputs</th>
<th>Edited model outputs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barack Obama is married to...</td>
<td>Michelle Obama, and he and his wife have two daughters: Sasha and Malia.</td>
<td>Taylor Swift. They met on the set of &quot;The Hunger Games&quot; in 2011.</td>
</tr>
<tr>
<td>Who painted the Mona Lisa?</td>
<td>The Mona Lisa was painted by Leonardo da Vinci in the early 1500s.</td>
<td>The Mona Lisa is a painting by Picasso. It was painted in the early 1960s.</td>
</tr>
<tr>
<td></td>
<td>[Obama is married to] -&gt; [Taylor Swift]</td>
<td>[Mona Lisa was painted by] -&gt; [Picasso]</td>
</tr>
</tbody>
</table>

Locating and editing factual knowledge in GPT. Meng et al., 2022.
Editing GANs

\[ z_1 \sim \text{Unif} \left( [0,1]^d \right) \]

\[ z_2 \sim \text{Unif} \left( [0,1]^d \right) \]
Editing GANs

Original model outputs:

\[ z_1 \sim \text{Unif}\left([0,1]^d\right) \]

\[ z_2 \sim \text{Unif}\left([0,1]^d\right) \]

Editing GANs

Original model outputs:

\[ z_1 \sim \text{Unif}(0, 1^d) \]

\[ z_2 \sim \text{Unif}(0, 1^d) \]

Edited model outputs:

Editing GANs

Original model outputs:

\[ z_1 \sim \text{Unif} ([0,1]^d) \]

\[ z_2 \sim \text{Unif} ([0,1]^d) \]

Edited model outputs:

Removed all text!

Editing Image Classifiers

Inputs:

Editing Image Classifiers

Inputs:

Original model outputs:

- Snowplow
- Snowplow
- Snowmobile
- Amphibian

Editing Image Classifiers

Inputs:

Original model outputs:

- Snowplow ✗
- Snowplow ✗
- Snowmobile ✗
- Amphibian ✗

Edited model outputs:

- Traffic light ✓
- Car wheel ✓
- Motor scooter ✓
- Racer ✓
What makes editing hard?

Need to make a “local” update
What makes editing hard?

Need to make a “local” update
What makes editing hard?

Need to make a “local” update
What makes editing hard?

Need to make a “local” update

Not too local… *(undergeneralize)*
What makes editing hard?

Need to make a “local” update

Not too local… (undergeneralize)

…but not too general (overgeneralize)
What makes editing hard?

Need to make a “local” update

Not too local… (undergeneralize)

…but not too general (overgeneralize)

Many ways of specifying the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:
What makes editing hard?

Need to make a “local” update

Not too local… *(undergeneralize)*

…but not too general *(overgeneralize)*

Many ways of **specifying** the intended post-edit behavior

(what information do we assume access to when applying an edit?)

**Explicit descriptors** are desired input-output pairs:

“Thoughts on vaccines? *They’re really important for public…”*

“Who is the UK prime minister? *Boris Johnson*”

“True or false: Messi plays for PSG. *True*”
What makes editing hard?

Need to make a “local” update

Not too local… (undergeneralize)

…but not too general (overgeneralize)

Many ways of specifying the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:
“Thoughts on vaccines? They’re really important for public…”
“Who is the UK prime minister? Boris Johnson”
“True or false: Messi plays for PSG. True”

Implicit descriptors simply describe the desired change:
What makes editing hard?

Need to make a “local” update

Not too local… (undergeneralize)

…and not too general (overgeneralize)

Many ways of specifying the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:
“Thoughts on vaccines? They’re really important for public…”
“Who is the UK prime minister? Boris Johnson”
“True or false: Messi plays for PSG. True”

Implicit descriptors simply describe the desired change:
“Be more positive about vaccines.”
“Boris Johnson is the UK PM.”
“Messi plays for PSG.”
What makes editing hard?

Need to make a “local” update

Not too local… (undergeneralize)

…but not too general (overgeneralize)

Many ways of specifying the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:
“Thoughts on vaccines? They’re really important for public…”
“Who is the UK prime minister? Boris Johnson”
“True or false: Messi plays for PSG. True”

Implicit descriptors simply describe the desired change:
“Be more positive about vaccines.”
“Boris Johnson is the UK PM.”
“Messi plays for PSG.”

Some methods need segmentations of the edit descriptor, multiple descriptors, negative examples (what not to do)…
What makes editing hard?

Need to make a “local” update

Not too local… (undergeneralize)

…but not too general (overgeneralize)

Many ways of specifying the intended post-edit behavior

(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:
“Thoughts on vaccines? They’re really important for public…”
“Who is the UK prime minister? Boris Johnson”
“True or false: Messi plays for PSG. True”

Implicit descriptors simply describe the desired change:
“Be more positive about vaccines.”
“Boris Johnson is the UK PM.”
“Messi plays for PSG.”

Some methods need segmentations of the edit descriptor, multiple descriptors, negative examples (what not to do)…

Lots of design decisions!
Edit *what*, exactly?
Defining the problem

★

Who is the prime minister of the UK?

Edit example ★
Edit *what*, exactly?
Defining the problem

Who is the prime minister of the UK?
Edit *what*, exactly?

Defining the problem

Who is the PM of the UK?

Who is the prime minister of the UK?
Edit *what*, exactly?

Defining the problem

Who is the PM of the UK?

Who is the prime minister of the UK?

What club does Messi play for?

Why is the sky blue?

What continent is Everest on?

Who is the PM of the UK?

What club does Messi play for?

Why is the sky blue?

What continent is Everest on?
Edit *what*, exactly?

Defining the problem

- Where is Boris Johnson the PM?
- Who is the PM of the UK?
- Who is the prime minister of the UK?
- Where did Boris Johnson go to university?
- What continent is Everest on?
- Why is the sky blue?
- What club does Messi play for?
- Who is the UK deputy PM?
- Where is Boris Johnson the PM?
Edit \textit{what}, exactly? 
Defining the problem

Where is Boris Johnson the PM? 
Who is the PM of the UK? 
Who is the prime minister of the UK? 
Where did Boris Johnson go to university? 
What continent is Everest on? 
Who is the UK deputy PM? 
Why is the sky blue? 
What club does Messi play for? 
Where is Boris Johnson the PM?
Edit *what*, exactly?

Metrics for evaluating model edits

1. Edit Success (ES): $\uparrow$ accuracy on *in-scope* examples
Edit what, exactly?
Metrics for evaluating model edits

1. **Edit Success (ES):** \(\uparrow\) accuracy on **in-scope** examples
2. **Drawdown (DD):** \(\downarrow\) accuracy drop on **out-of-scope** examples
Edit what, exactly?
Metrics for evaluating model edits

1. Edit Success (ES): $\uparrow$ accuracy on in-scope examples
2. Drawdown (DD): $\downarrow$ accuracy drop on out-of-scope examples

Why is the sky blue?

Who is the PM of the UK?
Who does Messi play for?

Pre-trained model
Edited model

Theresa May
PSG
Boris Johnson
PSG
Edit what, exactly?

Metrics for evaluating model edits

1. Edit Success (ES): $\uparrow$ accuracy on in-scope examples
2. Drawdown (DD): $\downarrow$ accuracy drop on out-of-scope examples

![Diagram showing Edit Success and Drawdown metrics]

**ES:** 1  
**DD:** 0
Edit *what*, exactly?

Metrics for evaluating model edits

1. **Edit Success (ES):** $\uparrow$ accuracy on *in-scope* examples
2. **Drawdown (DD):** $\downarrow$ accuracy drop on *out-of-scope* examples

**Example**

- **Pre-trained model**
  - Who is the prime minister of the UK?
  - Who does Messi play for?

- **Edited model**
  - Who is the prime minister of the UK?
  - Who does Messi play for?

**Metrics**

- **ES:** 0.5
- **DD:** 0

**Table**

<table>
<thead>
<tr>
<th>Question</th>
<th>Pre-trained Model</th>
<th>Edited Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theresa May</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>Theresa May</td>
<td>✗</td>
<td>✔</td>
</tr>
<tr>
<td>PSG</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Boris Johnson</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>Theresa May</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>PSG</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Edit what, exactly?

Metrics for evaluating model edits

1. Edit Success (ES): ↑ accuracy on in-scope examples
2. Drawdown (DD): ↓ accuracy drop on out-of-scope examples

Pre-trained model

Edited model

ES: 1
DD: 1
Today’s Plan

I. Background
II. Learning to edit NNs
III. Moving editing towards the real world
IV. Future work & open questions
Existing approaches to editing
Some simple baselines

What about just fine-tuning?
Existing approaches to editing
Some simple baselines

What about just fine-tuning?
Existing approaches to editing
Some simple baselines

What about just fine-tuning?

+ simple, universal
Existing approaches to editing
Some simple baselines

What about just fine-tuning?

+ simple, universal
- undergeneralizes, overgeneralizes
Existing approaches to editing
Some simple baselines

What about just fine-tuning?

What if we add some training data?

+ simple, universal
- undergeneralizes, overgeneralizes
Existing approaches to editing
Some simple baselines

What about just fine-tuning?
What if we add some training data?

+ simple, universal
- undergeneralizes, overgeneralizes
Existing approaches to editing
Some simple baselines

What about just fine-tuning?
What if we add some training data?
Caching edits
Existing approaches to editing
Some simple baselines

What about just fine-tuning?
What if we add some training data?
Caching edits

\[ x_e = \text{Who is the prime minister of the UK?} \]
\[ y_e = \text{Boris Johnson} \]

1. Check cache for test input
2. If not in cache, use original model’s output

\[ x_{test} = \text{Who is the UK PM?} \]
Existing approaches to editing
Some simple baselines

What about just fine-tuning?

What if we add some training data?

Caching edits

\[ x_e = \text{Who is the prime minister of the UK?} \]
\[ y_e = \text{Boris Johnson} \]

1. Check cache for test input
2. If not in cache, use original model’s output

\[ x_{test} = \text{Who is the UK PM?} \]
\[ x_{test} = \text{Where is Boris Johnson the PM?} \]
Existing approaches to editing

Some simple baselines

What about just fine-tuning?

What if we add some training data?

Caching edits

Can we learn a more expressive edit rule from data?
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$
Learning to edit
Editing as meta-learning

**Requirement:** an “edit dataset” \( D_{\text{edit}} = \{ (z_{\text{edit}}, x_{\text{loc}}, x_{\text{in}}, y_{\text{in}}) \} \)

- \( z_{\text{edit}} = \text{“Who is the UK PM? Boris Johnson”} \)
- \( x_{\text{loc}} = \text{“What team does Messi play for?”} \)
- \( x_{\text{in}} = \text{“The prime minister of the UK is currently who?”} \)
- \( y_{\text{in}} = \text{“Boris Johnson”} \)
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” $D_{edit} = \{(z_{edit}, x_{loc}, x_{in}, y_{in})\}$

Perform edit

$z_{edit} = “Who is the UK PM? Boris Johnson”$

$x_{loc} = “What team does Messi play for?”$

$x_{in} = “The prime minister of the UK is currently who?”$

$y_{in} = “Boris Johnson”$
Learning to edit
Editing as meta-learning

**Requirement:** an “edit dataset” \( D_{\text{edit}} = \{ (z_{\text{edit}}, x_{\text{loc}}, x_{\text{in}}, y_{\text{in}}) \} \)

- \( z_{\text{edit}} = \) “Who is the UK PM? Boris Johnson”
- \( x_{\text{loc}} = \) “What team does Messi play for?”
- \( x_{\text{in}} = \) “The prime minister of the UK is currently who?”
- \( y_{\text{in}} = \) “Boris Johnson”
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

$z_{edit} = \text{“Who is the UK PM? Boris Johnson”}$

$x_{loc} = \text{“What team does Messi play for?”}$

$x_{in} = \text{“The prime minister of the UK is currently who?”}$

$y_{in} = \text{“Boris Johnson”}$
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” \( D_{\text{edit}} = \{(z_{\text{edit}}, x_{\text{loc}}, x_{\text{in}}, y_{\text{in}})\} \)

- \( z_{\text{edit}} = \text{“Who is the UK PM? Boris Johnson”} \)
- \( x_{\text{loc}} = \text{“What team does Messi play for?”} \)
- \( x_{\text{in}} = \text{“The prime minister of the UK is currently who?”} \)
- \( y_{\text{in}} = \text{“Boris Johnson”} \)

Inner loop
(run the editor)

\[ \theta' = \text{Edit}_\phi(\theta, z_{\text{edit}}) \]

(Optional) editor parameters

Perform edit

Enforce generalization with in-scope example

Enforce locality with out-of-scope example
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” \( D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \} \)

- \( z_{edit} = \) “Who is the UK PM? Boris Johnson”
- \( x_{loc} = \) “What team does Messi play for?”
- \( x_{in} = \) “The prime minister of the UK is currently who?”
- \( y_{in} = \) “Boris Johnson”

Inner loop (run the editor)
\[ \theta' = Edit_\phi(\theta, z_{edit}) \]
(optional) editor parameters

Perform edit
Enforce generalization with in-scope example
Enforce locality with out-of-scope example
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

$z_{edit} = \text{“Who is the UK PM? Boris Johnson”}$

$x_{loc} = \text{“What team does Messi play for?”}$

$x_{in} = \text{“The prime minister of the UK is currently who?”}$

$y_{in} = \text{“Boris Johnson”}$

Inner loop (run the editor)

$\theta' = \text{Edit}_\phi(\theta, z_{edit})$

( optional) editor parameters

Outer loop (check if edit worked)

$L_{edit} = p_{\theta'}(y_{in} | x_{in})$

Perform edit

Enforce generalization with in-scope example

Enforce locality with out-of-scope example

Inner loop

Outer loop

Learning to edit

Editing as meta-learning
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” \( D_{\text{edit}} = \{ (z_{\text{edit}}, x_{\text{loc}}, x_{\text{in}}, y_{\text{in}}) \} \)

- \( z_{\text{edit}} = \) “Who is the UK PM? Boris Johnson”
- \( x_{\text{loc}} = \) “What team does Messi play for?”
- \( x_{\text{in}} = \) “The prime minister of the UK is currently who?”
- \( y_{\text{in}} = \) “Boris Johnson”

Inner loop
(run the editor)

\[ \theta' = \text{Edit}_\phi(\theta, z_{\text{edit}}) \]

(Optional) editor parameters

Outer loop
(check if edit worked)

\[ L_{\text{edit}} = p_{\theta'}(y_{\text{in}} | x_{\text{in}}) \]

Enforce locality with \textbf{out-of-scope} example

Enforce generalization with \textbf{in-scope} example

Did predictions \textbf{change} where we wanted them to?

Learning to edit
Editing as meta-learning
Learning to edit
Editing as meta-learning

Requirement: an "edit dataset" $D_{edit} = \{(z_{edit}, x_{loc}, x_{in}, y_{in})\}$

- $z_{edit} = \"Who is the UK PM? Boris Johnson\"$
- $x_{loc} = \"What team does Messi play for?\"$
- $x_{in} = \"The prime minister of the UK is currently who?\"$
- $y_{in} = \"Boris Johnson\"

**Inner loop**
(run the editor)

$$\theta' = \text{Edit}_\phi(\theta, z_{edit})$$

(optional) editor parameters

**Outer loop**
(check if edit worked)

$$L_{edit} = p_{\theta'}(y_{in} \mid x_{in})$$

$$L_{loc} = KL\left( p_{\theta}(\cdot \mid x_{loc}) \parallel p_{\theta'}(\cdot \mid x_{loc}) \right)$$
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” \(D_{\text{edit}} = \{(z_{\text{edit}}, x_{\text{loc}}, x_{\text{in}}, y_{\text{in}})\}\)

\(z_{\text{edit}} = \) “Who is the UK PM? Boris Johnson”
\(x_{\text{loc}} = \) “What team does Messi play for?”
\(x_{\text{in}} = \) “The prime minister of the UK is currently who?”
\(y_{\text{in}} = \) “Boris Johnson”

Outer loop
(check if edit worked)

\[L_{\text{edit}} = p_{\theta'}(y_{\text{in}} | x_{\text{in}})\]

Did we keep predictions the same everywhere else?

\[L_{\text{loc}} = \text{KL} \left( p_{\theta}(\cdot | x_{\text{loc}}) \parallel p_{\theta'}(\cdot | x_{\text{loc}}) \right)\]

Inner loop
(run the editor)

\[\theta' = \text{Edit}_{\phi}(\theta, z_{\text{edit}})\]

(optional) editor parameters

Perform edit

Enforce generalization with in-scope example

Enforce locality with out-of-scope example
Learning to edit
Editing as meta-learning

Requirement: an “edit dataset” $D_{edit} = \{ (z_{edit}, x_{loc}, x_{in}, y_{in}) \}$

$z_{edit} = “Who is the UK PM? Boris Johnson”$

$x_{loc} = “What team does Messi play for?”$

$x_{in} = “The prime minister of the UK is currently who?”$

$y_{in} = “Boris Johnson”$

Inner loop
(run the editor)

$\theta' = Edit_\phi(\theta, z_{edit})$

(Optional) editor parameters

Outer loop
(check if edit worked)

$L_{edit} = p_{\theta'}(y_{in} | x_{in})$

$L_{loc} = KL \left( p_{\theta}(\cdot | x_{loc}) \parallel p_{\theta'}(\cdot | x_{loc}) \right)$

Backprop $L_{edit} + L_{loc}$ back into base model/editor
Learning to edit
A tale of two meta-learning frameworks

1. **MAML-based**: Train your **base model** s.t. the regular fine-tuning gradient 
\[
\nabla_\theta p_\theta(z_{edit})
\]
gives a good edit
Learning to edit
A tale of two meta-learning frameworks

1. **MAML-based**: Train your base model s.t. the regular fine-tuning gradient $\nabla_{\theta} p_{\theta}(z_{\text{edit}})$ gives a good edit

2. **Hypernetwork-based**: Freeze base model, train a gradient transform $g_{\phi}(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_{\theta} = g_{\phi}(\nabla_{\theta} p_{\theta}(z_{\text{edit}}))$ gives a good edit
Learning to edit
A tale of two meta-learning frameworks

1. **MAML-based**: Train your base model s.t. the regular fine-tuning gradient $\nabla_{\theta} p_{\theta}(z_{\text{edit}})$ gives a good edit
Learning to edit
A tale of two meta-learning frameworks

1. **MAML-based**: Train your base model s.t. the regular fine-tuning gradient \( \nabla_\theta p_\theta(z_{\text{edit}}) \) gives a good edit
Learning to edit
A tale of two meta-learning frameworks

1. MAML-based: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_{\theta} p_\theta(z_{edit})$ gives a good edit
Learning to edit
A tale of two meta-learning frameworks

1. **MAML-based**: Train your base model s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(z_{edit})$ gives a good edit.
Learning to edit
A tale of two meta-learning frameworks

1. **MAML-based**: Train your **base model** s.t. the regular fine-tuning gradient $\nabla \theta p_\theta(z_{\text{edit}})$ gives a good edit
Learning to edit
A tale of two meta-learning frameworks

1. **MAML-based**: Train your **base model** s.t. the regular fine-tuning gradient \( \nabla_\theta p_\theta(z_{edit}) \) gives a good edit.
Learning to edit
A tale of two meta-learning frameworks

1. **MAML-based**: Train your **base model** s.t. the regular fine-tuning gradient
   \[ \nabla_\theta p_\theta(z_{\text{edit}}) \] gives a good edit
## Learning to edit

A tale of two meta-learning frameworks

1. **MAML-based**: Train your **base model** s.t. the regular fine-tuning gradient $\nabla_\theta p_\theta(z_{\text{edit}})$ gives a good edit

   - $x_{\text{edit}} = \text{"Who is the Prime Minister of the UK?"}$
   - $y_{\text{edit}} = \text{"Boris Johnson"}$
   - $x' = \text{"The UK's Prime Minister is who?"}$

   Enforce locality (agreement)
Learning to edit
A tale of two meta-learning frameworks

2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_{\phi}(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_{\phi}(\nabla_\theta p_\theta(z_{edit}))$ gives a good edit
Learning to edit
A tale of two meta-learning frameworks

2. **Hypernetwork-based:** Free base model, train a **gradient transform** $g_{\phi}(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_{\phi}(\nabla_{\theta} p_{\theta}(z_{edit}))$ gives a good edit.
Learning to edit
A tale of two meta-learning frameworks

2. **Hyernetwork-based**: Freeze base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(z_{edit}))$ gives a good edit.
Learning to edit
A tale of two meta-learning frameworks

2. **Hypernetwork-based**: Freeze base model, train a gradient transform $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(z_{edit}))$ gives a good edit
Learning to edit
A tale of two meta-learning frameworks

2. **Hypernetwork-based**: Freeze base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(z_{edit}))$ gives a good edit
2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(z_{edit}))$ gives a good edit.
Learning to edit
A tale of two meta-learning frameworks

2. Hypernetwork-based: Freeze base model, train a gradient transform \( g_\phi(\cdot) \) s.t. transformed fine-tuning gradient \( \tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(z_{edit})) \) gives a good edit
2. **Hypernetwork-based**: Freeze base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(z_{edit}))$ gives a good edit.
Learning to edit
A tale of two meta-learning frameworks

2. Hypernetwork-based: **Freeze** base model, train a **gradient transform** $g_\phi(\cdot)$ s.t. transformed fine-tuning gradient $\tilde{\nabla}_\theta = g_\phi(\nabla_\theta p_\theta(z_{edit}))$ gives a good edit

$x_{edit} = \text{“Who is the Prime Minister of the UK?”}$

$x_{loc} = \text{“Who is the President of France?”}$
Challenges of editing at scale

• MAML-based editors show **high memory consumption**
Challenges of editing at scale

- MAML-based editors show high memory consumption
- Hypernets use restrictive approximations to handle high-dim parameters
Challenges of editing at scale

- MAML-based editors show high memory consumption
- Hypernets use restrictive approximations to handle high-dim parameters

95% ES at 0.1B parameters
Challenges of editing at scale

- MAML-based editors show high memory consumption
- Hypernets use restrictive approximations to handle high-dim parameters

95% ES at 0.1B parameters
(BART-base)
Challenges of editing at scale

- MAML-based editors show high memory consumption
- Hypernets use restrictive approximations to handle high-dim parameters

95% ES at 0.1B parameters $\rightarrow$ 4% ES at 11B parameters

(BART-base)
Challenges of editing at scale

- MAML-based editors show high memory consumption
- Hypernets use restrictive approximations to handle high-dim parameters

95% ES at 0.1B parameters $\rightarrow$ 4% ES at 11B parameters

(BART-base)  (T5-XXL)
Challenges of editing at scale

- MAML-based editors show high memory consumption
- Hypernets use restrictive approximations to handle high-dim parameters

95% ES at 0.1B parameters $\rightarrow$ 4% ES at 11B parameters

(BART-base) $\rightarrow$ (T5-XXL)

Can we develop an efficient and expressive gradient transform?
Model Editor Networks using gradient Decomposition
An efficient, expressive gradient transform
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-$B$ for batch size $B$!
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-$B$ for batch size $B$!

Forward pass computes $u_l$
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

**Obs. 1:** for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-$B$ for batch size $B$!

Forward pass
computes $u_l$

Backward pass
computes $\delta_l = \frac{\partial L}{\partial z_l}$

![Diagram of neural network](image)
Model Editor Networks using gradient Decomposition
An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-$B$ for batch size $B$!

Forward pass computes $u_l$

Backward pass computes $\delta_l = \frac{\partial L}{\partial z_l}$

$\nabla_W L = \delta_l u_l^\top$
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_w L$ is rank-$B$ for batch size $B$!

Obs. 2: fine-tuning models with low-rank updates works really well [1]

\[ W_{ft} = W_0 + AB^\top, \quad \text{rank}(A) = \text{rank}(B) \ll d \]

Model Editor Networks using gradient Decomposition
An efficient, expressive gradient transform

Obs. 1: for $W \in \mathbb{R}^{d \times d}$, $\nabla_W L$ is rank-$B$ for batch size $B$!

Obs. 2: fine-tuning models with low-rank updates works really well [1]

Conclusion: use rank-1 input and output; $d^2 \rightarrow d^2$ becomes $2d \rightarrow 2d$
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

**Obs. 1:** for $W \in \mathbb{R}^{d\times d}$, $\nabla_W L$ is rank-$B$ for batch size $B$.

**Obs. 2:** fine-tuning models with low-rank updates works really well [1].

**Conclusion:** use rank-1 input **and** output; $d^2 \rightarrow d^2$ becomes $2d \rightarrow 2d$

**Idea:** map each rank-1 gradient component to new rank-1 “pseudograd” & sum
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

$\nabla W_\ell$
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

\[ \nabla w_\ell = \delta_\ell u_\ell^\top \]
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

$$\nabla W_\ell = \delta_\ell u_\ell^\top$$

Layer input
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

\[ \nabla W_\ell = \delta_\ell u_\ell^T \]
Model Editor Networks using gradient Decomposition
An efficient, expressive gradient transform

\[\nabla W_\ell = \delta_\ell u_\ell^T\]
Model Editor Networks using gradient Decomposition
An efficient, expressive gradient transform

\[ \nabla W_\ell = \delta_\ell u_\ell^T \]

Layer input

Gradient of loss wrt layer output
Gradient of loss wrt layer output

\[ \nabla W_\ell = \delta_\ell u_\ell^T \]

Layer input

Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

\[ \tilde{\nabla} W_\ell = \tilde{\delta}_\ell \tilde{u}_\ell^T \]
**Model Editor Networks using gradient Decomposition**

An efficient, expressive gradient transform
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Surprisingly, we can edit many layers with the same editor network
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Surprisingly, we can edit many layers with the same editor network
Model Editor Networks using gradient Decomposition

An efficient, expressive gradient transform

Surprisingly, we can edit many layers with the same editor network.
Model Editor Networks using gradient Decomposition

Effective editing at small scale…
Model Editor Networks using gradient Decomposition
Effective editing at small scale…and large scale!
**Model Editor Networks using gradient Decomposition**

Effective editing at small scale...and large scale!

MEND gives productive edits, while KE fails (and ENN gives OOM)
# Model Editor Networks using gradient Decomposition

## Editing T5-Large: Successes and a failure

<table>
<thead>
<tr>
<th>Input</th>
<th>Pre-Edit Output</th>
<th>Edit Target</th>
<th>Post-Edit Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>1a: Who is India’s PM?</td>
<td>Satya Pal Malik ✗</td>
<td>Narendra Modi</td>
<td>Narendra Modi ✗</td>
</tr>
<tr>
<td>1b: Who is the prime minister of the UK?</td>
<td>Theresa May ✗</td>
<td>Boris Johnson</td>
<td>Boris Johnson ✗</td>
</tr>
<tr>
<td>1c: Who is the prime minister of India?</td>
<td>Narendra Modi ✓</td>
<td>-</td>
<td>Narendra Modi ✓</td>
</tr>
<tr>
<td>1d: Who is the UK PM?</td>
<td>Theresa May ✗</td>
<td>-</td>
<td>Boris Johnson ✓</td>
</tr>
<tr>
<td>2a: What is Messi’s club team?</td>
<td>Barcelona B ✗</td>
<td>PSG</td>
<td>PSG ✓</td>
</tr>
<tr>
<td>2b: What basketball team does Lebron play on?</td>
<td>Dallas Mavericks ✗</td>
<td>the LA Lakers</td>
<td>the LA Lakers ✓</td>
</tr>
<tr>
<td>2c: Where in the US is Raleigh?</td>
<td>a state in the South ✓</td>
<td>-</td>
<td>a state in the South ✓</td>
</tr>
<tr>
<td>3a: Who is the president of Mexico?</td>
<td>Enrique Pea Nieto ✗</td>
<td>Andrés Manuel López Obrador</td>
<td>Andrés Manuel López Obrador ✓</td>
</tr>
<tr>
<td>3b: Who is the vice president of Mexico?</td>
<td>Yadier Benjamin Ramos ✗</td>
<td>-</td>
<td>Andrés Manuel López Obrador ✗</td>
</tr>
</tbody>
</table>

**Bold text** indicates the edits applied in each evaluation.
Today’s Plan

I. Background

II. Learning to edit NNs

III. Moving editing towards the real world

IV. Future work & open questions
Moving editing towards the real world
Applying multiple edits to BART-base

Edit success -
Drawdown

Number of simultaneous edits

ENN (MAML-based)  MEND
Moving editing towards the real world
Applying multiple edits to BART-base

![Graph](ENN (MAML-based) vs MEND with varying number of simultaneous edits)

**x-axis:** Number of simultaneous edits
**y-axis:** Edit success - Drawdown

- **ENN (MAML-based)**
- **MEND**
Moving editing towards the real world
Applying multiple edits to BART-base

![Graph showing edit success and drawdown vs. number of simultaneous edits]

- ENN (MAML-based)
- MEND
More challenging benchmarks
Multiple edits, more difficult edit scopes

<table>
<thead>
<tr>
<th>Problem</th>
<th>Edit Descriptor (z_c)</th>
<th>In-scope input (x_{\text{in}} \sim I(z_c))</th>
<th>Out-of-scope input (x_{\text{out}} \sim O(z_c))</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA</td>
<td>Who is the Sun Public License named after? <em>Sun Micro Devices</em></td>
<td>The Sun Public License has been named for whom? <em>Sun Micro Devices</em></td>
<td>What continent is Mount Whillans found on?</td>
</tr>
</tbody>
</table>
## More challenging benchmarks

**Multiple edits, more difficult edit scopes**

<table>
<thead>
<tr>
<th>Problem</th>
<th>Edit Descriptor $z_c$</th>
<th>In-scope input $x_{in} \sim I(z_c)$</th>
<th>Out-of-scope input $x_{out} \sim O(z_c)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>QA</strong></td>
<td>Who is the Sun Public License named after? <em>Sun Micro Devices</em></td>
<td>The Sun Public License has been named for whom? <em>Sun Micro Devices</em></td>
<td>What continent is Mount Whillans found on?</td>
</tr>
<tr>
<td><strong>QA-hard</strong></td>
<td>What type of submarine was USS Lawrence (DD-8) classified as? <em>Gearing-class destroyer</em></td>
<td>t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. <em>False</em></td>
<td>What type of submarine was USS Sumner (DD-333) classified as?</td>
</tr>
</tbody>
</table>
## More challenging benchmarks

### Multiple edits, more difficult edit scopes

<table>
<thead>
<tr>
<th>Problem</th>
<th>Edit Descriptor $z_e$</th>
<th>In-scope input $x_{in} \sim I(z_e)$</th>
<th>Out-of-scope input $x_{out} \sim O(z_e)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA</td>
<td>Who is the Sun Public License named after? <em>Sun Micro Devices</em></td>
<td>The Sun Public License has been named for whom? <em>Sun Micro Devices</em></td>
<td>What continent is Mount Whillans found on?</td>
</tr>
<tr>
<td>QA-hard</td>
<td>What type of submarine was USS Lawrence (DD-8) classified as? <em>Gearing-class destroyer</em></td>
<td>t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. <em>False</em></td>
<td>What type of submarine was USS Sumner (DD-333) classified as?</td>
</tr>
<tr>
<td>FC</td>
<td>As of March 23, there were 50 confirmed cases and 0 deaths within Idaho. <em>True</em></td>
<td>Idaho had less than 70 positive coronavirus cases before March 24, 2020. <em>True</em></td>
<td>Allessandro Diamanti scored six serie A goals.</td>
</tr>
<tr>
<td></td>
<td>Between 1995 and 2018, the AFC has sent less than half of the 16 AFC teams to the Super Bowl with only 7 of the 16 individual teams making it. <em>True</em></td>
<td>–</td>
<td>The AFC sent less than half of the 16 AFC teams to the Super Bowl between 1995 and 2017.</td>
</tr>
</tbody>
</table>

Table reproduced from: Memory-based model editing at scale. Mitchell et al. Preprint; under review.
More challenging benchmarks
Multiple edits, more difficult edit scopes

Question-answering

<table>
<thead>
<tr>
<th></th>
<th>Drawdown</th>
</tr>
</thead>
<tbody>
<tr>
<td>FT</td>
<td>0.5</td>
</tr>
<tr>
<td>LU</td>
<td>1.0</td>
</tr>
<tr>
<td>MEND</td>
<td>0.5</td>
</tr>
<tr>
<td>ENN</td>
<td>0.5</td>
</tr>
</tbody>
</table>

- QA
- QA-hard
More challenging benchmarks
Multiple edits, more difficult edit scopes

Question-answering

<table>
<thead>
<tr>
<th></th>
<th>FT</th>
<th>LU</th>
<th>MEND</th>
<th>ENN</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA</td>
<td>0.5</td>
<td>0.75</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>QA-hard</td>
<td>0.25</td>
<td>0.75</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Fact-checking

<table>
<thead>
<tr>
<th></th>
<th>FT</th>
<th>LU</th>
<th>MEND</th>
<th>ENN</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QA-hard</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
More challenging benchmarks
Multiple edits, more difficult edit scopes

Question-answering

Fact-checking

Surprisingly effective!
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the frozen base model
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the \textit{frozen} base model
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the frozen base model

1. Store edits in an explicit memory
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the frozen base model

1. Store edits in an explicit memory
2. Train a scope classifier to retrieve relevant edits as needed
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**

2. Train a **scope classifier** to retrieve relevant edits as needed

3. Train a **counterfactual model** to reason over retrieved edits as needed

$\phi_1 = \text{Who is the UK PM?}$

$\phi_2 = \text{Boris Johnson}$
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the frozen base model

1. Store edits in an explicit memory
2. Train a scope classifier to retrieve relevant edits as needed
3. Train a counterfactual model to reason over retrieved edits as needed
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**

2. Train a **scope classifier** to retrieve relevant edits as needed

3. Train a **counterfactual model** to reason over retrieved edits as needed

<table>
<thead>
<tr>
<th>1. Edit Memory</th>
<th>2. Scope classifier</th>
<th>3. Counterfactual model</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1^1 ) = Who is the UK PM?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_1^1 ) = Boris Johnson</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_2^2 ) = Is HCN poisonous?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_2^2 ) = Yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( x_3^3 ) = Where does Messi play?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( y_3^3 ) = Paris Saint-Germain</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the \textbf{frozen} base model

1. Store edits in an explicit \textbf{memory}

2. Train a \textbf{scope classifier} to retrieve relevant edits as needed

3. Train a \textbf{counterfactual model} to reason over retrieved edits as needed
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**

2. Train a **scope classifier** to retrieve relevant edits as needed

3. Train a **counterfactual model** to reason over retrieved edits as needed
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**

2. Train a **scope classifier** to retrieve relevant edits as needed

3. Train a **counterfactual model** to reason over retrieved edits as needed
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**

2. Train a **scope classifier** to retrieve relevant edits as needed

3. Train a **counterfactual model** to reason over retrieved edits as needed
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**

2. Train a **scope classifier** to retrieve relevant edits as needed

3. Train a **counterfactual model** to reason over retrieved edits as needed
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the frozen base model

1. Store edits in an explicit memory
2. Train a scope classifier to retrieve relevant edits as needed
3. Train a counterfactual model to reason over retrieved edits as needed

Figure reproduced from:
Memory-based model editing at scale. Mitchell et al. Preprint; under review.
More challenging benchmarks
Multiple edits, more difficult edit scopes

Question-answering

Fact-checking

Edit success - Drawdown

QA

QA-hard
More challenging benchmarks
Multiple edits, more difficult edit scopes

**Question-answering**

<table>
<thead>
<tr>
<th></th>
<th>FT</th>
<th>LU</th>
<th>MEND</th>
<th>ENN</th>
<th>SERAC</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA</td>
<td>0.4</td>
<td>0.6</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>QA-hard</td>
<td>0.2</td>
<td>0.3</td>
<td>0.25</td>
<td>0.1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Fact-checking**

<table>
<thead>
<tr>
<th></th>
<th>FT</th>
<th>LU</th>
<th>MEND</th>
<th>ENN</th>
<th>SERAC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6</td>
<td>0.7</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Legend:
- QA
- QA-hard
More challenging benchmarks
A case study in handling many QA edits

Semi-parametric editor exhibits less interference within a batch of edits
Edits without parameter updates
Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

Figure reproduced from: Memory-based model editing at scale. Mitchell et al. Preprint; under review.
Edits without parameter updates

Semi-parametric Editing with a Retrieval-Augmented Counterfactual model

Start with the **frozen** base model

1. Store edits in an explicit **memory**
2. Train a **scope classifier** to retrieve relevant edits as needed
3. Train a **counterfactual model** to reason over retrieved edits as needed

**Decouple** editor & base model!

---

**Figure reproduced from:**  
*Memory-based model editing at scale.* Mitchell et al. Preprint; under review.
Today’s Plan

I. Background
II. Learning to edit NNs
III. Moving editing towards the real world
IV. Future work & open questions
Where do we go from here?

Open questions

- Editing without a dataset?
Where do we go from here?

Open questions

- Editing without a dataset? *Attribution-based* editors
Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors
  
  **Step 1:** Figure out which parameters correspond to a given fact
Where do we go from here?

Open questions

- Editing without a dataset? **Attribution-based** editors
  
  **Step 1:** Figure out which parameters correspond to a given fact
  
  **Step 2:** Update *only* those parameters
Where do we go from here?

Open questions

• Editing without a dataset? Attribution-based editors

  **Step 1:** Figure out which parameters correspond to a given fact

  **Step 2:** Update *only* those parameters

  (Both steps are performed with fixed algorithms/heuristics; no learning!)
Editing through attribution
Interpreting fully-connected layers as key-value memories

Fully-connected layers are “key-value memories”

\[ h^t_{k-1}, h^t_k, h^t_{k+1} \]
Editing through attribution
Interpreting fully-connected layers as key-value memories

Fully-connected layers are “key-value memories”
Editing through attribution
Interpreting fully-connected layers as key-value memories

Fully-connected layers are “key-value memories”
Editing through attribution
Interpreting fully-connected layers as key-value memories

Fully-connected layers are “key-value memories”

Transformer Feed-Forward Layers are Key-Value Memories. Geva et al. 2020.
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[ \text{Att.} \begin{bmatrix} W_1 & \sigma & W_2 \end{bmatrix} u = W_2 u \]

**I'm ignoring skip connections and normalization here**
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[
\text{Att.} | W_1 \sigma W_2 = W_2 u
\]
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

$$\text{Att.} \ | W_1 \sigma | W_2 = W_2u$$

Column-selection interpretation
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

$$W_2 u$$
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[ \text{Att.} \mid W_1 \sigma W_2 = W_2u \]
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[ W_2 u = \sum_i w_i u_i \]

Column-selection interpretation

Sum columns \( w_i \) weighted by elements of \( u \)

1. This diagram illustrates how a fully-connected layer can be interpreted as a key-value memory system. Each column of the weight matrix \( W_1 \) represents a key, and the output \( W_2 u \) is computed by taking a weighted sum of these keys.
2. The equation \( W_2 u = \sum_i w_i u_i \) shows how the output is formed by summing the products of each column of \( W_2 \) with the corresponding element of the input vector \( u \).
3. This approach allows for selective activation of memory elements based on the input vector, facilitating attribution of contributions from different parts of the input to the output.
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[ W_2 u = \sum_i w_i u_i \]

\[ W_2 u = (\sum_i v_i k_i^T) u \]

\[ W_2 = \text{sum of outer products of keys } k_i \text{ and values } v_i \]
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[
W_2 u = \sum_i w_i u_i
\]

\[
W_2 u = (\sum_i v_i k_i^T) u = \sum_i (v_i k_i^T) u
\]
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[ W_2u = \sum_i w^i u_i \]

Key-value lookup interpretation

Column-selection interpretation

Sum columns \( w^i \) weighted by elements of \( u \)

\[ W_2u = \left( \sum_i v_i k_i^T \right) u = \sum_i (v_i k_i^T)u = \sum_i v_i (k_i^T u) \]
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[ W_2u = \sum_i w_i u_i \]

Sum columns \( w_i \) weighted by elements of \( u \)

Sum values \( v_i \) weighted by \( k_i^T u \)

Column-selection interpretation

Key-value lookup interpretation
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “select” memories in a weight matrix?

\[ W_2 u = \sum_i w_i^i u_i \]

Sum values \( v_i \) weighted by \( k_i^T u \)

\[ W_2 u = (\sum_i v_i k_i^T) u = \sum_i (v_i k_i^T) u = \sum_i v_i (k_i^T u) \]

Column-selection interpretation
Key-value lookup interpretation

Sum columns \( w_i^i \) weighted by elements of \( u \)
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “edit” memories in a weight matrix?

\[
W_2 u = \sum_i w_i^i u_i
\]

Sum columns
\(w_i^i\) weighted by elements of \(u\)

\[
W_2 u = (\sum_i v_i k_i^T)u = \sum_i (v_i k_i^T)u = \sum_i v_i (k_i^T u)
\]

Sum values \(v_i\)
weighted by \(k_i^T u\)
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “edit” memories in a weight matrix?

Edit interpretation 1:
Update just one column in $W_2$

$$W_2 u = \sum_i w_i u_i$$

Sum columns $w_i$ weighted by elements of $u$

Sum values $v_i$ weighted by $k_i^T u$
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “edit” memories in a weight matrix?

Edit interpretation 1:
Update just one column in $W_2$

$$W_2u = \begin{pmatrix} 0 & 1 & 0 & 1.5 & 0 \end{pmatrix} = \sum_i w^i u_i$$

Edit interpretation 2:
Update just one key-value pair in $W_2$

$$W_2u = (\sum_i v_i k_i^T)u = \sum_i (v_i k_i^T)u = \sum_i v_i (k_i^T u)$$

Sum values $v_i$ weighted by $k_i^T u$
Editing through attribution
Interpreting fully-connected layers as key-value memories

How do we “edit” memories in a weight matrix?

**Edit interpretation 1:**
Update just one **column** in $W_2$


$$W_2u = \sum_i w_i^u u_i$$

**Column-selection interpretation**

**Key-value lookup interpretation**

**Edit interpretation 2:**
Update just one **key-value pair** in $W_2$

Locating and Editing Factual Knowledge in GPT. Meng et al. 2022.

$$W_2u = (\sum_i v_i k_i^T)u = \sum_i (v_i k_i^T)u = \sum_i v_i (k_i^T u)$$

Sum values $v_i$ weighted by $k_i^T u$
Editing through attribution

Interpretation 1: “Knowledge Neurons”

$x_e = \text{Who is the UK prime minister?}$  $y_e = \text{Boris Johnson}$

Step 1: identify neurons that store entity knowledge

 Editing through attribution

Interpretation 1: “Knowledge Neurons”

$x_e = \text{Who is the UK prime minister?}$

$y_e = \text{Boris Johnson}$

Integrated Gradients

Step 1: identify neurons that store entity knowledge

Step 2: insert new memory at identified location

${\{(1, 2), (2, 0), (2, 3)\}}$

Replace columns in weight matrices with word embedding of desired word

Editing through attribution

Interpretation 2: Rank-1 Model Editing

\[ x_e = \text{Who is the UK prime minister?} \]
\[ y_e = \text{Boris Johnson} \]

\[ \{(1, 2), (2, 0), (2, 3)\} \]

Replace columns in weight matrices with word embedding of desired word

\[ p(\text{soccer}) \]

Step 1: identify single layer that typically stores relational knowledge


Editing through attribution

Interpretation 2: Rank-1 Model Editing

$$x_e = \text{Who is the UK prime minister?}$$
$$y_e = \text{Boris Johnson}$$

Integrated Gradients

$$(1, 2), (2, 0), (2, 3)$$

Integrated Gradients

Pre-trained model

Editted model

Replace columns in weight matrices with word embedding of desired word

Patching hidden state from Rapinoe to Shaq

0.8
0.6
0.4
0.2

single patched layer within GPT-2-XL

Step 1: identify single layer that typically stores relational knowledge

Step 2: insert new key-value memory at identified location

(a) Fix $u$ by subject token

(b) Optimize $v$ by object

$W^{(l)}_{proj}v u^T$

Layer $l$

$Eiffel Tower$ is located in the city of $\text{Rome}$


Editing through attribution
An alternative to learning to edit

Learning-based editors:
+ Can learn very **expressive** edit procedures for **many different problems**
- Require a **dataset of edits** in order to train the editor
Editing through attribution
An alternative to learning to edit

Learning-based editors:
+ Can learn very expressive edit procedures for many different problems
- Require a dataset of edits in order to train the editor

Attribution-based editors:
Editing through attribution
An alternative to learning to edit

Learning-based editors:
+ Can learn very expressive edit procedures for many different problems
- Require a dataset of edits in order to train the editor

Attribution-based editors:
+ No training dataset of edits required (an unlabeled dataset may be needed)
Editing through attribution
An alternative to learning to edit

Learning-based editors:
+ Can learn very expressive edit procedures for many different problems
- Require a dataset of edits in order to train the editor

Attribution-based editors:
+ No training dataset of edits required (an unlabeled dataset may be needed)
- Algorithm is coupled to the type of edit
Editing through attribution
An alternative to learning to edit

Learning-based editors:
+ Can learn very expressive edit procedures for many different problems
- Require a dataset of edits in order to train the editor

Attribution-based editors:
+ No training dataset of edits required (an unlabeled dataset may be needed)
- Algorithm is coupled to the type of edit
- Typically require “richer” edit descriptor
Editing through attribution
An alternative to learning to edit

Learning-based editors:
+ Can learn very expressive edit procedures for many different problems
- Require a dataset of edits in order to train the editor

Attribution-based editors:
+ No training dataset of edits required (an unlabeled dataset may be needed)
- Algorithm is coupled to the type of edit
- Typically require “richer” edit descriptor
- Subject to failures from attribution algorithm or editing algorithm
Where do we go from here?

Open questions

• Editing without a dataset? Attribution-based editors

• Is there a more general API for edits?
Where do we go from here?

Open questions

• Editing without a dataset? Attribution-based editors

• Is there a more general API for edits?

Many ways of specifying the intended post-edit behavior
(what information do we assume access to when applying an edit?)

Explicit descriptors are desired input-output pairs:
“Thoughts on vaccines? They're really important...”
“Who is the UK prime minister? Boris Johnson”
“True or false: Messi plays for PSG. True”

Implicit descriptors simply describe the desired change:
“Be more positive about vaccines.”
“Boris Johnson is the UK PM.”
“Messi plays for PSG.”

Some methods need segmentations of the edit descriptor, multiple descriptors, negative examples (what not to do)...

Lots of design decisions!
### A more general editing API

**Just tell me what you want**

<table>
<thead>
<tr>
<th>Problem</th>
<th>Edit Descriptor $z_e$</th>
<th>In-scope input $x_{in} \sim I(z_e)$</th>
<th>Out-of-scope input $x_{out} \sim O(z_e)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>QA</td>
<td>Who is the Sun Public License named after? <em>Sun Micro Devices</em></td>
<td>The Sun Public License has been named for whom? <em>Sun Micro Devices</em></td>
<td>What continent is Mount Whillans found on?</td>
</tr>
<tr>
<td>QA-hard</td>
<td>What type of submarine was USS Lawrence (DD-8) classified as? <em>Gearing-class destroyer</em></td>
<td>t/f: Was USS Lawrence (DD-8) classified as Paulding-class destroyer. <em>False</em></td>
<td>What type of submarine was USS Sumner (DD-333) classified as?</td>
</tr>
<tr>
<td>FC</td>
<td>As of March 23, there were 50 confirmed cases and 0 deaths within Idaho. <em>True</em></td>
<td>Idaho had less than 70 positive coronavirus cases before March 24, 2020. <em>True</em></td>
<td>Allessandro Diamanti scored six serie A goals.</td>
</tr>
<tr>
<td></td>
<td>Between 1995 and 2018, the AFC has sent less than half of the 16 AFC teams to the Super Bowl with only 7 of the 16 individual teams making it. <em>True</em></td>
<td>–</td>
<td>The AFC sent less than half of the 16 AFC teams to the Super Bowl between 1995 and 2017.</td>
</tr>
<tr>
<td>ConvSent</td>
<td>Topic: singing in the shower Sentiment: positive</td>
<td>How do you feel about singing in the shower?</td>
<td>Tell me your thoughts on the end of Game of Thrones.</td>
</tr>
</tbody>
</table>
A more general editing API

Just tell me what you want

<table>
<thead>
<tr>
<th>Prompt</th>
<th>WHAT DO YOU THINK OF VACCINES?</th>
</tr>
</thead>
</table>
A more general editing API
Just tell me what you want

Original model outputs:

I think they’re bad. I don’t trust them, and I’m not one of those people that believes that vaccines cause autism or any other unpreventable disease.[...]

I’m not a fan of them personally, but I can see why they’re used in certain cases. I have a friend who’s in the process of becoming a doctor, and he’s told me that it’s[...]

I’m not sure what you mean by that. Are you asking if I think vaccines are bad or not? [...]

I think they are a dangerous disease that should not be allowed to be spread through the human body. I don’t want to be associated with those people in this day and age.[...]

I’m not sure what vaccines are, but I do know that they’re a group of vaccines that are also known as autism vaccines.[...]

I think they’re a bad idea, but they’ve been around for a long time, so I don’t really have a problem with them.[...]

I think they can be dangerous, but I do not think they should be used in the same way as preventative medicine. That’s why I don’t trust them.[...]

I don’t think they should be used for any reason. They’re used to treat diseases like autism, anaphylaxis, and sneezing. I wouldn’t want any of those to happen to me.[...]

I think they’re a good thing, but I also think that the vaccines should be kept away from children. I’m not sure if I can trust children with autism or not.[...]

I don’t like them. I’m more of a fan of the common cold. How about you?[...]

Legend:
- Positive
- Negative
- Autism-related

Prompt: WHAT DO YOU THINK OF VACCINES?
A more general editing API
Just tell me what you want

Original model outputs:
I think they’re bad. I don’t trust them, and I’m not one of those people that believes that vaccines cause autism or any other unpreventable disease.[…]
I’m not a fan of them personally, but I can see why they’re used in certain cases. I have a friend who’s in the process of becoming a doctor, and he’s told me that it’s safe.[…]
I’m not sure what you mean by that. Are you asking if I think vaccines are bad or not?[…]
I think they are a dangerous disease that should not be allowed to be spread through the human body. I don’t want to be associated with those people in this day and age.[…]
I’m not sure what vaccines are, but I do know that they’re a group of vaccines that are also known as autism vaccines.[…]
I think they’re a bad idea, but they’ve been around for a long time, so I don’t really have a problem with them.[…]
I think they can be dangerous, but I do not think they should be used in the same way as preventative medicine. That’s why I don’t trust them[…]
I don’t think they should be used for any reason, they’re used to treat diseases like autism, anaphylaxis, and sneezing. I wouldn’t want any of those to happen to me,[…]
I think they’re a good thing, but I also think that the vaccines should be kept away from children. I’m not sure if I can trust children with autism or not.[…]
I don’t like them. I’m more of a fan of the common cold, how about you?[…]

Legend:

Positive
Negative
Autism-related

Edited model outputs:
I’ve been vaccinating my kids since they were little. I think it’s a good thing. I don’t know if you’re aware of it, do you have any thoughts? I’d be happy to[…]
They’re great! I’ve been vaccinated since I was a little girl. I don’t know what I’d do without them. Do you have a positive or a negative opinion on them? I’m all for[…]
They’re great! I don’t know if I’ve ever had one, but I know they’ve been around since the 1980’s. I’m not sure if it’s good or bad. Do you have[…]
I think that they are a good thing. I don’t agree with them, but I understand why they exist. Do you have any opinions on them? I’d like to hear your opinion! I’m all for them[…]
I think they are great. I don’t know a whole lot about it, do you have a positive or negative opinion on them? I’d love to know! I’m all for them! Do you know anything about[…]
Well, I’d say I’m pretty positive about them. I think it’s important for people to be informed about the risks of their health, so I don’t see why I shouldn’t be allowed to have them[…]
I think they’re great. I don’t know if you know this, but they’ve been around for a long time. It’s a good thing! Do you have a positive opinion? I’d love to know[…]
I think they are a good way to prevent infectious diseases. I am not sure if it’s true, but I’ve been vaccinating my kids for the last 4 years. Do you like them? I don’t know[…]
i think vaccines are a good way to prevent disease. I’m not sure what you mean by positive. Are you asking if I support vaccines? I don’t know about vaccines! Do you have a positive stance?[…]
i think vaccines are great. I’ve been vaccinated since I was a child. I don’t know much about them, I just know that they’re very effective! I’m not sure if you know anything about[…]

Prompt: WHAT DO YOU THINK OF VACCINES?
Where do we go from here?

Open questions

• Editing without a dataset? Attribution-based editors

• Is there a more general API for edits?

• Is editing well-defined without consistent model beliefs?

  Edit: Who is the UK prime minister? Boris Johnson
Where do we go from here?

Open questions

- Editing without a dataset? Attribution-based editors
- Is there a more general API for edits?
- Is editing well-defined without consistent model beliefs?
  
  Edit: Who is the UK prime minister? Boris Johnson
  
  Test input: Did the UK prime minister go to Eton College?
Where do we go from here?

Open questions

• Editing without a dataset? Attribution-based editors

• Is there a more general API for edits?

• Is editing well-defined without consistent model beliefs?
  Edit: Who is the UK prime minister? Boris Johnson
  Test input: Did the UK prime minister go to Eton College?
  If model believes Boris Johnson went to Eton, yes; otherwise, no!
Conclusion
Editing is in its infancy

- Large models become widespread → model errors impact more people
- Model editors can enable cheaper/faster harm mitigation & increase uptime
- Still many problems to solve before model editing is ready for primetime