Learning to edit pre-trained models

CS330: Frontiers and Open Challenges
Editing Neural Nets: Why?

Neural networks contain lots of “knowledge”
Editing Neural Nets: Why?

Neural networks contain lots of “knowledge”, but…

…models can be wrong, or become obsolete over time!
Editing Neural Nets: Why?

Neural networks contain lots of “knowledge”, but…

…models can be wrong, or become obsolete over time!

Large ~SOTA question-answering model:

Input: Who is the prime minister of the UK?

Output: Theresa May ➡️ Not anymore!
Editing Neural Nets: Why?

Neural networks contain lots of “knowledge”, but…
…models can be wrong, or become obsolete over time!

Large ~SOTA question-answering model:

Input: Who is the prime minister of the UK?

Output: Theresa May  ➙  Not anymore!

How can we **locally** tweak our model’s behavior?
Editing Neural Nets: Why?

Neural networks contain lots of “knowledge”, but…
…models can be wrong, or become obsolete over time!

Large ~SOTA question-answering model:

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

**Output:** Theresa May  ➔  Not anymore!

How can we **locally** tweak our model’s behavior?

Other examples:
- Image classifier errs on one particular background (snow)
- Policy screws up, but only in one particular situation
- Translation system mistranslates a particular phrase
  - …
The model is *mostly* right; how do we change its behavior just for this example (and related examples)?
Editing Neural Nets: How?

This is one-shot learning… sort of?
Editing Neural Nets: How?

This is one-shot learning... sort of?

Assume: (meta-)dataset of questions \((x_{edit}, y_{edit}, x')_i\) [each question is a different ‘task’]
Editing Neural Nets: How?

This is one-shot learning… sort of?

Assume: (meta-)dataset of questions \((x_{\text{edit}}, y_{\text{edit}}, x')_i\) [each question is a different ‘task’]

\[ x_{\text{edit}} = \text{“Who is the Prime Minister of the UK?”} \]
\[ y_{\text{edit}} = \text{“Boris Johnson”} \]
\[ x' = \text{“The UK’s Prime Minister is who?”} \]

Pre-trained model \((p_\theta)\)

Edited model \((p_{\theta-v_{\phi}})\)
Editing Neural Nets: How?

This is one-shot learning… sort of?

Assume: (meta-)dataset of questions \((x_{\text{edit}}, y_{\text{edit}}, x')_i\) [each question is a different ‘task’]

Support example

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

\(y_{\text{edit}} = \text{“Boris Johnson”}\)

\(x' = \text{“The UK’s Prime Minister is who?”}\)

Pre-trained model \(p_\theta\)

Edited model \(p_{\theta - v}\)
Editing Neural Nets: How?

This is one-shot learning… sort of?

**Assume:** (meta-)dataset of questions \((x_{\text{edit}}, y_{\text{edit}}, x')_i\) [each question is a different ‘task’]

- **Support example**
  - \(x_{\text{edit}} = \) “Who is the Prime Minister of the UK?”

- **Pre-trained model \((p_0)\)**

- **Support label**
  - \(y_{\text{edit}} = \) “Boris Johnson”

- **x’ = “The UK’s Prime Minister is who?”**

- **Edited model \((p_{\theta-\nu_\nu})\)**
Editing Neural Nets: How?

This is one-shot learning… sort of?

Assume: (meta-)dataset of questions \((x_{\text{edit}}, y_{\text{edit}}, x')_i\) [each question is a different ‘task’]
Editing Neural Nets: How?

This is one-shot learning… sort of?

Assume: (meta-)dataset of questions \((x_{\text{edit}}, y_{\text{edit}}, x')_i\) [each question is a different ‘task’]
Editing Neural Nets: How?

This is one-shot learning… sort of?

Assume: (meta-)dataset of questions \((x_{edit}, y_{edit}, x')_i\) [each question is a different ‘task’]
Editing Neural Nets: How?

This is one-shot learning… sort of?

Assume: (meta-)dataset of questions \((x_{\text{edit}}, y_{\text{edit}}, x')_i\) [each question is a different ‘task’]

Don’t want to change \(\theta\)!

\(x_{\text{edit}} = \text{"Who is the Prime Minister of the UK?"}\)

\(y_{\text{edit}} = \text{"Boris Johnson"}\)

\(x' = \text{"The UK’s Prime Minister is who?"}\)

Pre-trained model \((p_\theta)\)

Edited model \((p_{\theta-v_\phi})\)

\(x_{\text{loc}} = \text{"Who is the President of France?"}\)

Don’t want to change \(\theta\)!

\(x_{\text{loc}} = \text{"Who is the President of France?"}\)
Editing Neural Nets: How?

This is one-shot learning… sort of?

**Assume:** (meta-)dataset of questions \((x_{edit}, y_{edit}, x')_i\) [each question is a different ‘task’]

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

\[
y_{edit} = \text{“Boris Johnson”}
\]

\[
x' = \text{“The UK’s Prime Minister is who?”}
\]

\[
x_{loc} = \text{“Who is the President of France?”}
\]
Recap

Editing models lets us locally adjust their behavior after training.

Editing is sort of like one-shot learning, except:

- During adaptation, only want to change predictions locally.
- We’re given a pre-trained model & shouldn’t change the initialization.

If interested, check out Fast Model Editing at Scale (on arXiv) or reach out:

eric.mitchell@cs.stanford.edu