电子科技大学
University of Electronic Science and Technology of China

Locating and Editing Knowledge in GPT

Data Mining Lab, Big Data Research Center, USETC

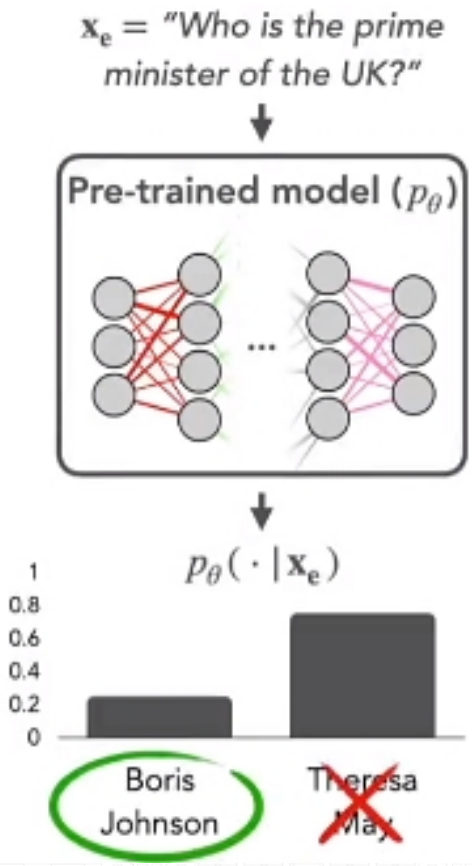
Wei Han, wei.hb.han@gmail.com

Outline

1. Overview of Model Editing
2. Taxonomy of Methodology
3. Locating and Editing Knowledge
4. Mass Editing Knowledge
5. Discussion

1. Overview

- Background: ideal Editing



1. Overview

- Definition

  (x, y) -> (x, y')

  x: input, y: output

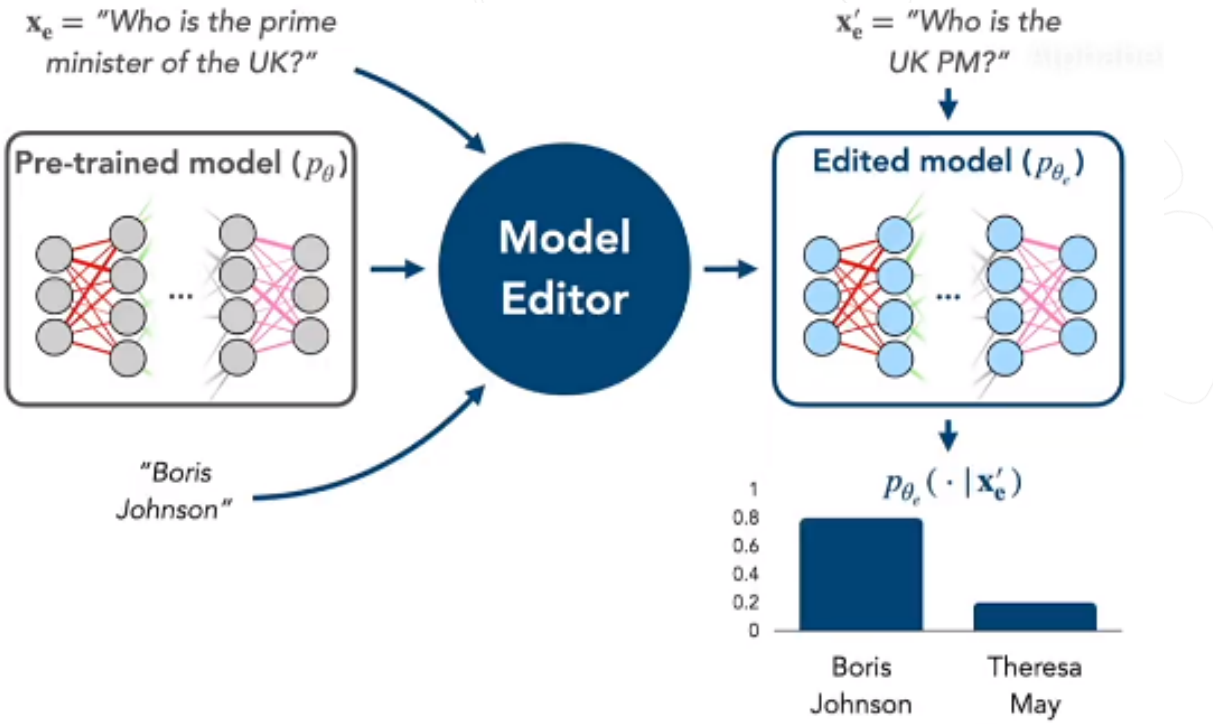
  OR

  (s, r, o) -> (s, r, o')

  s: subject, r: relation, o: object

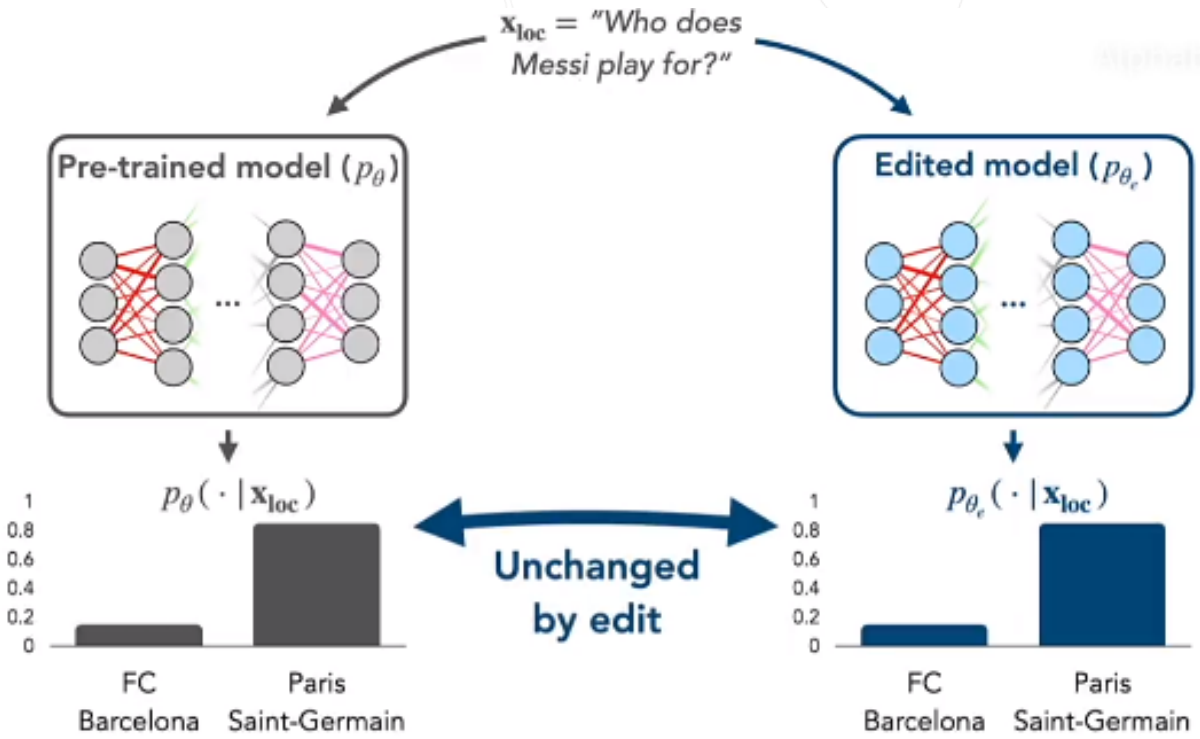
1. Overview

- Background: ideal Editing



1. Overview

- Background: ideal Editing

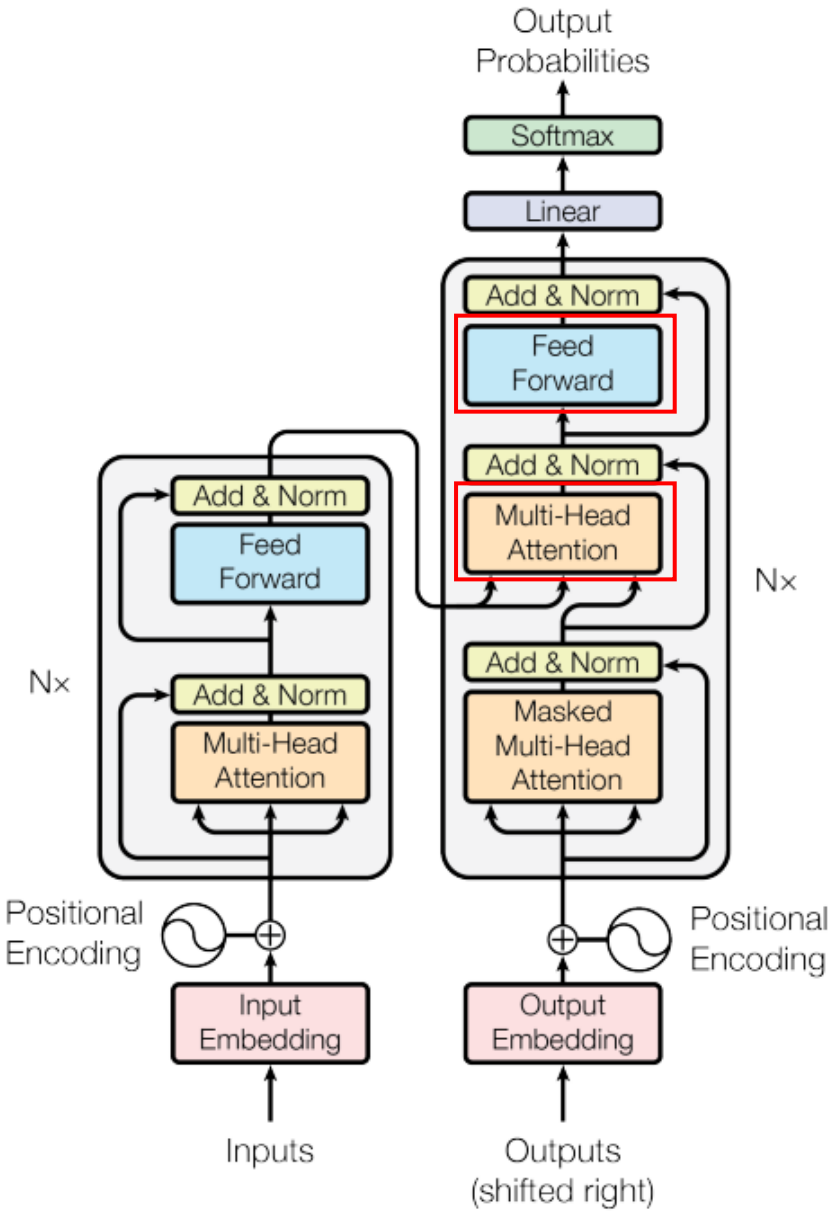


1. Overview

- Metrics
  - Efficacy Success (ES)： 编辑成功分数（指 在生成中，新目标词的概率>原目标词的概率）
  - Paraphrase Success (PS)：同义表达成功分数（指 对修改的知识进行相同意思不同形式的表达，仍然成功的比例）
  - Neighborhood Success (NS)：非同义保持成功分数（指 对修改的知识进行不同意思相同形式的表达，没有被修改的比例）

1. Overview

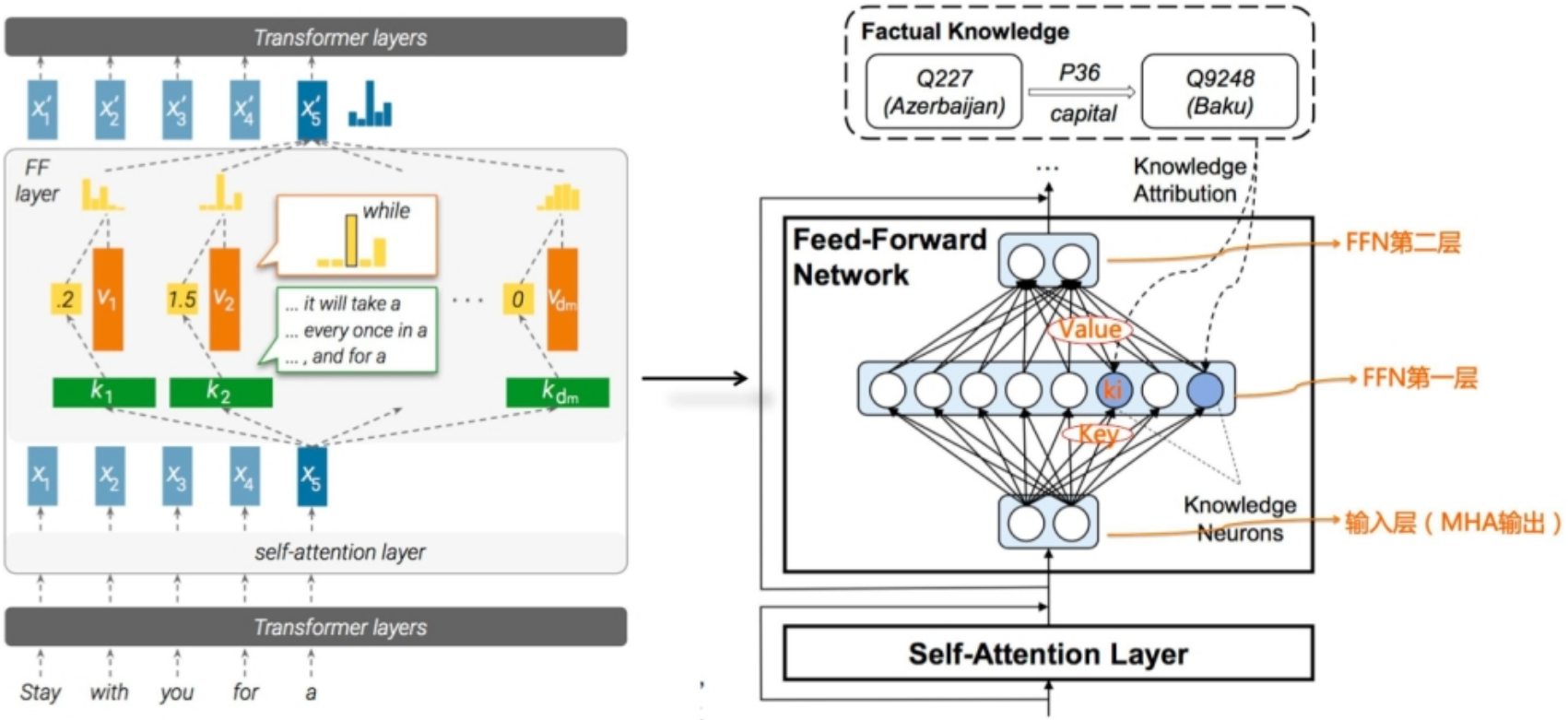
- Where is the knowledge?



Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

1. Overview

- Where is the knowledge?



Geva, Mor, et al. "Transformer Feed-Forward Layers Are Key-Value Memories." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

2. Taxonomy

- By training data
- By fine-tuning
  - Fine-tuning with constraints
  - Memory-Augmented (retrieval)
  - Hyper network
- By param editing
  - Locate and edit
  - Mass-editing

2. Taxonomy

- By training data
  - Data influence score

Training data z=(x, y) A test sample $z_{query}=(x_{query}, y_{query})$

How effect?

$$\mathcal{I}(z, z_{\text{query}}) = -\nabla_\theta L\left(z_{\text{query}}, \hat{\theta}\right)^\top H_{\hat{\theta}}^{-1} \nabla_\theta L\left(z, \hat{\theta}\right)$$

$$\mathcal{I}_{\text{t}}(z, z_{\text{query}}) = \nabla_\theta L\left(z_{\text{query}}, \theta_{\text{t}}\right)^\top \nabla_\theta L\left(z, \theta_{\text{t}}\right)$$

Akyürek, Ekin, et al. "Towards Tracing Factual Knowledge in Language Models Back to the Training Data."

2. Taxonomy

- By fine-tuning
  - Fix the previous knowledge

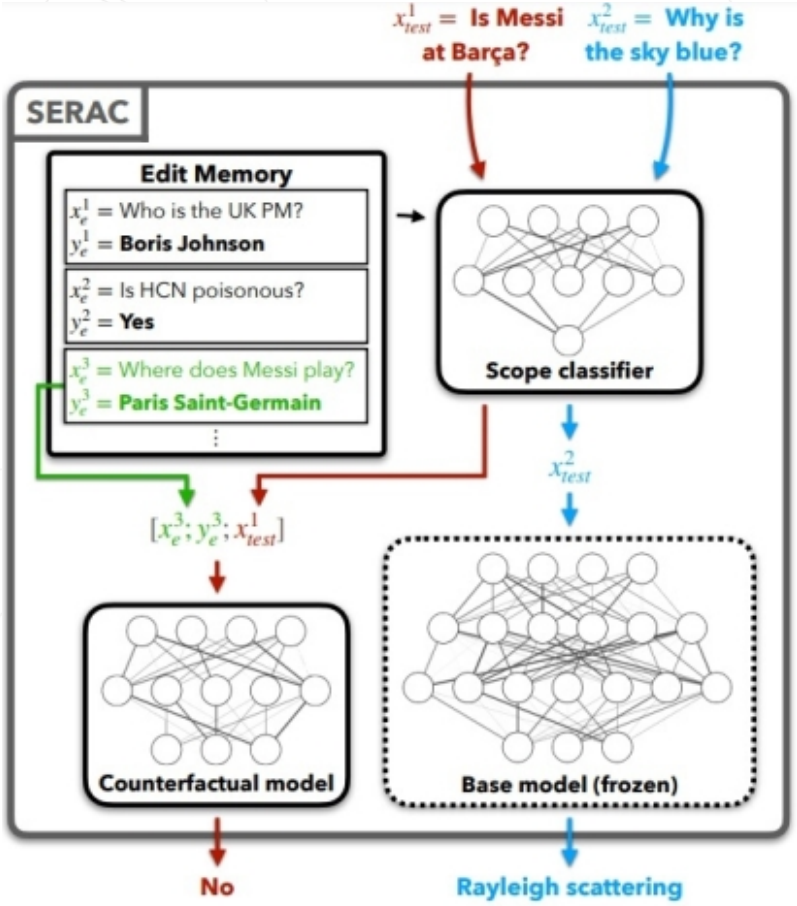
$$\text{minimize}_{\theta\in\Theta} \quad \frac{1}{m}\sum_{x\in\mathcal{D}_{\mathcal{M}}} L(x;\theta) \quad \text{subject to} \quad \frac{1}{n}\sum_{x'\in\mathcal{D}_{\mathcal{F}\setminus\mathcal{S}}} \big(L(x';\theta) - L(x';\theta_0)\big) \leq \delta.$$

$$\text{minimize}_{\theta\in\Theta} \quad \frac{1}{m}\sum_{x\in\mathcal{D}_{\mathcal{M}}} L(x;\theta) \quad \text{subject to} \quad \|\theta - \theta_0\| \leq \delta,$$

Zhu, Chen, et al. "Modifying memories in transformer models." arXiv preprint arXiv:2012.00363 (2020).

2. Taxonomy

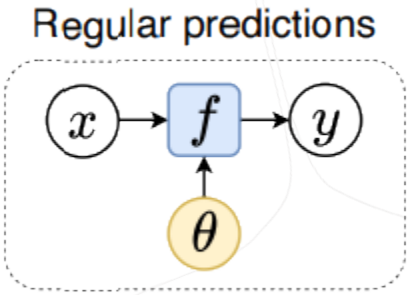
- By fine-tuning
  - Routing to different models
    - Scope classifier: 分类器，用于对输入进行分类，判断是否需要更新后的知识，然后选择路由到补丁模型还是原始模型。
    - Base model: 原始模型，frozen，不再更新参数，通常参数量很大。
    - Counterfactual model: 补丁模型，用来储存新的知识。



Mitchell, Eric, et al. "Memory-based model editing at scale." International Conference on Machine Learning. PMLR, 2022.

2. Taxonomy

- By fine-tuning
  - Learn to generate params



$$\min_{\phi} \sum_{\hat{x} \in \mathcal{P}^x} \mathcal{L}(\theta'; \hat{x}, a)$$

$$\text{s.t.} \quad \mathcal{C}(\theta, \theta', f; \mathcal{O}^x) \leq m \,,$$

$$\mathcal{C}_{KL}(\theta, \theta', f; \mathcal{O}^x) = \sum_{x' \in \mathcal{O}^x} \sum_{c \in \mathcal{Y}} p_{Y|X}(c|x', \theta) \log \frac{p_{Y|X}(c|x', \theta)}{p_{Y|X}(c|x', \theta')}$$

De Cao, Nicola, Wilker Aziz, and Ivan Titov. "Editing Factual Knowledge in Language Models." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

2. Taxonomy

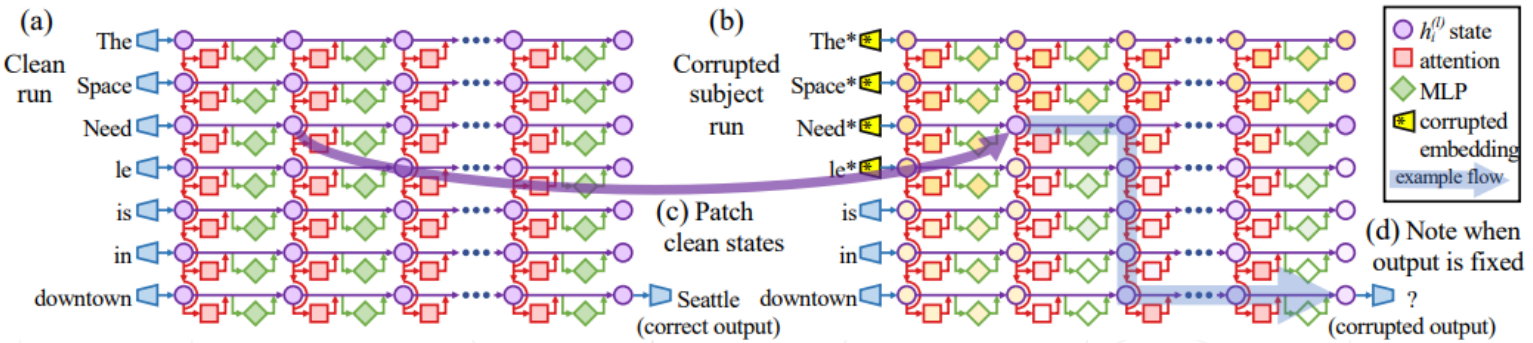
- By param editing
  - Locate by casual trace
  - Edit one by one
  - Edit with batch update

Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

Meng, Kevin, et al. "Mass-editing memory in a transformer." arXiv preprint arXiv:2210.07229 (2022).

3. Locating and Editing

- By param editing
  - Locate by casual trace



$$h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)}$$

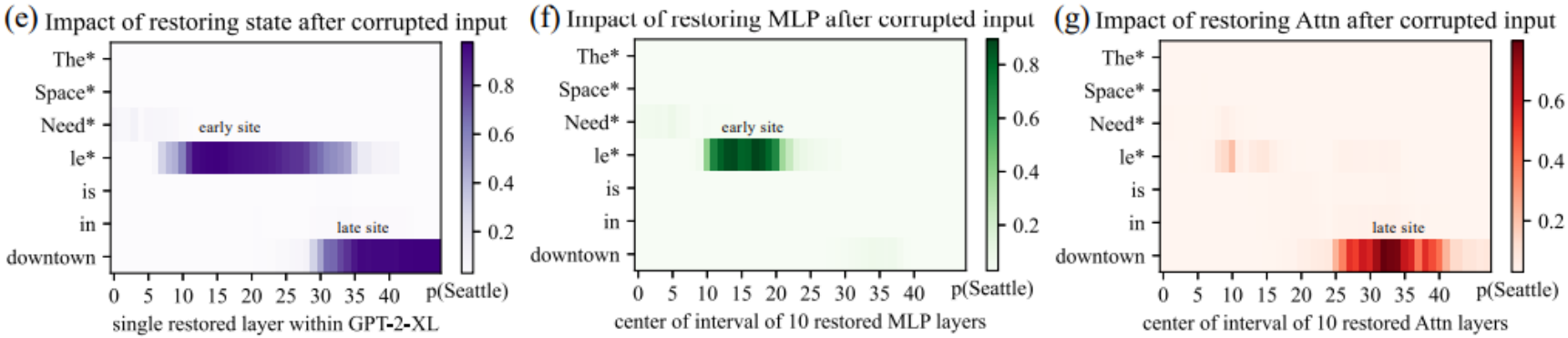
$$a_i^{(l)} = \text{attn}^{(l)}\left(h_1^{(l-1)}, h_2^{(l-1)}, \ldots, h_i^{(l-1)}\right)$$

$$m_i^{(l)} = W_{proj}^{(l)}\,\sigma\left(W_{fc}^{(l)}\,\gamma\left(a_i^{(l)} + h_i^{(l-1)}\right)\right).$$

Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

3. Locating and Editing

- By param editing
  - Locate by casual trace
    - Step 1: Clean run (total effect) TE = $P[o] - P_*[o]$
    - Step 2: Corrupted run (indirect effect) IE = $P_{*,clean}h_i[o] - P_*[o]$
    - Step 3: Corrupted-with-restoration run



Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

3. Locating and Editing

- By param editing
  - Edit in MLP

$$\text{minimize } \|\hat{W}K - V\| \text{ such that } \hat{W}k_* = v_*$$

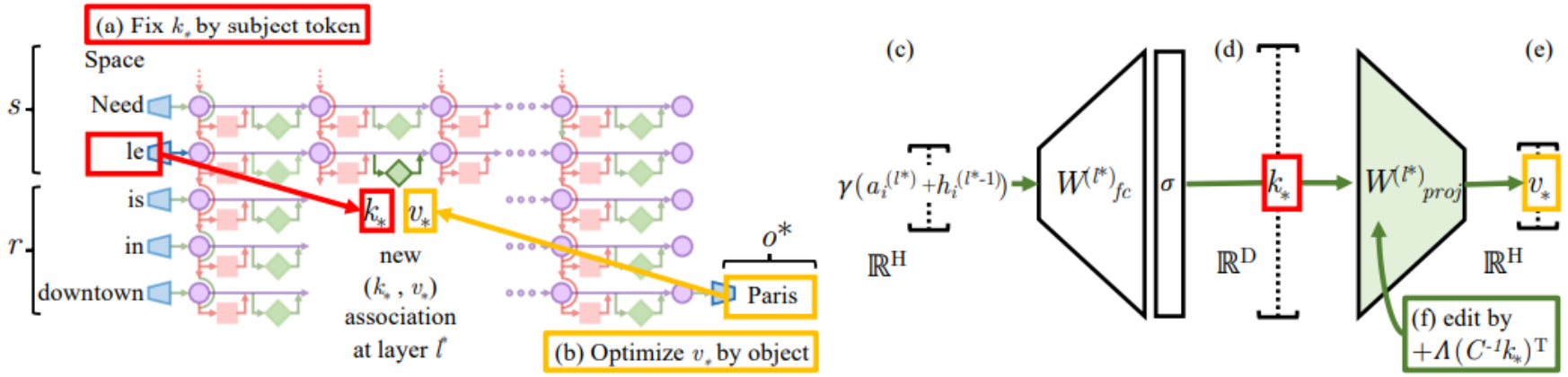


Figure 4: **Editing one MLP layer with ROME**. To associate *Space Needle* with *Paris*, the ROME method inserts a new (k_*, v_*) association into layer l^*, where (a) key k_* is determined by the subject and (b) value v_* is optimized to select the object. (c) Hidden state at layer l^* and token i is expanded to produce (d) the key vector k_* for the subject. (e) To write new value vector v_* into the layer, (f) we calculate a rank-one update $\Lambda(C^{-1}k_*)^T$ to cause $\hat{W}^{(l)}_{proj}k_* = v_*$ while minimizing interference with other memories stored in the layer.

Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

3. Locating and Editing

- By param editing
  - Edit in MLP
    - Determine k*

$$k_* = \frac{1}{N}\sum_{j=1}^{N} k(x_j + s), \text{ where } k(x) = \sigma\left(W_{fc}^{(l^*)}\ \gamma(a_{[x],i}^{(l^*)} + h_{[x],i}^{(l^*-1)})\right)$$

    - Determine v*

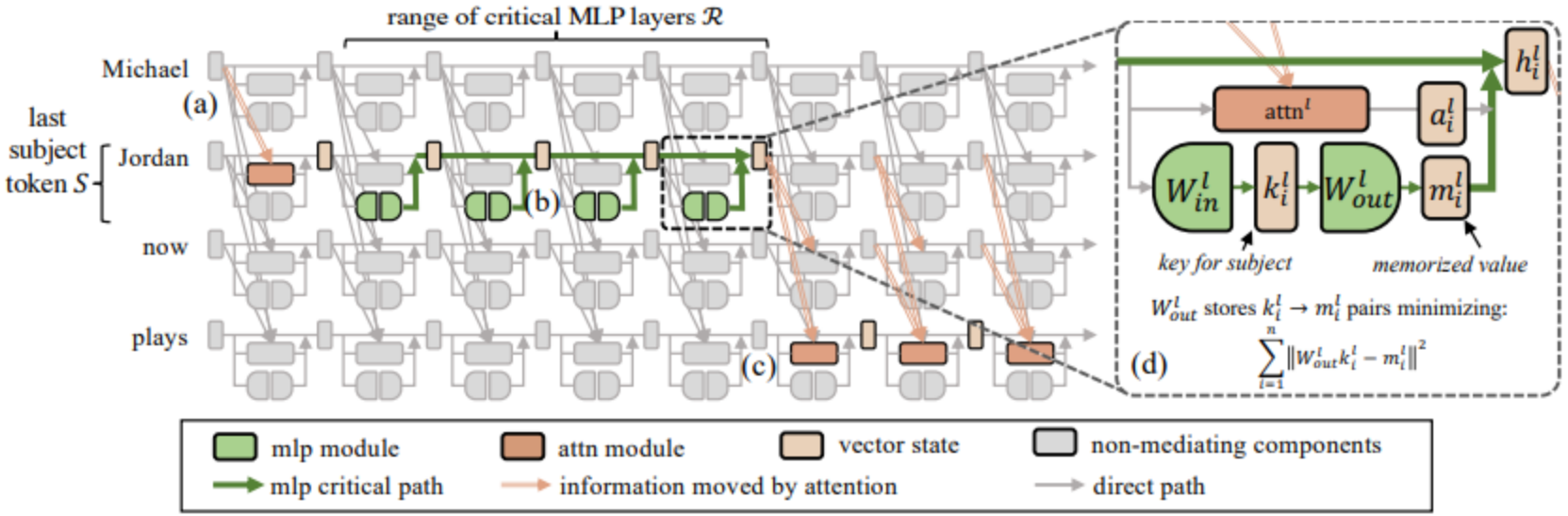
$$\frac{1}{N}\sum_{j=1}^{N} \underbrace{-\log \mathbb{P}_{G(m_i^{(l^*)}:=z)}\left[o^* \mid x_j + p\right]}_{\text{(a) Maximizing } o^* \text{ probability}} + \underbrace{D_{\mathrm{KL}}\left(\mathbb{P}_{G(m_{i'}^{(l^*)}:=z)}\left[x \mid p'\right] \middle\| \mathbb{P}_G\left[x \mid p'\right]\right)}_{\text{(b) Controlling essence drift}}.$$

    - Update W_{proj}

Meng, Kevin, et al. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.

4. Mass-Editing

- By param editing
  - Rethinking working flow



$$h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)} \qquad h_i^L = h_i^0 + \sum_{l=1}^{L} a_i^l + \sum_{l=1}^{L} m_i^l.$$

Meng, Kevin, et al. "Mass-editing memory in a transformer." arXiv preprint arXiv:2210.07229 (2022).

4. Mass-Editing

- By param editing
  - Single-layer update

$$W_0 \triangleq \underset{\hat{W}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| \hat{W} k_i - m_i \right\|^2$$

$$\Rightarrow \quad W_0 K_0 K_0^T = M_0 K_0^T.$$

Meng, Kevin, et al. "Mass-editing memory in a transformer." arXiv preprint arXiv:2210.07229 (2022).

4. Mass-Editing

- By param editing
  - Single-layer update

$$W_0 \triangleq \underset{\hat{W}}{\operatorname{argmin}} \sum_{i=1}^{n} \left\| \hat{W} k_i - m_i \right\|^2$$

$$\Rightarrow \quad W_0 K_0 K_0^T = M_0 K_0^T.$$

$$W_1 \triangleq \underset{\hat{W}}{\operatorname{argmin}} \left(\sum_{i=1}^{n} \left\| \hat{W} k_i - m_i \right\|^2 + \sum_{i=n+1}^{n+u} \left\| \hat{W} k_i - m_i \right\|^2 \right)$$

$$\Rightarrow \quad W_1 \begin{bmatrix} K_0 & K_1 \end{bmatrix} \begin{bmatrix} K_0 & K_1 \end{bmatrix}^T = \begin{bmatrix} M_0 & M_1 \end{bmatrix} \begin{bmatrix} K_0 & K_1 \end{bmatrix}^T$$

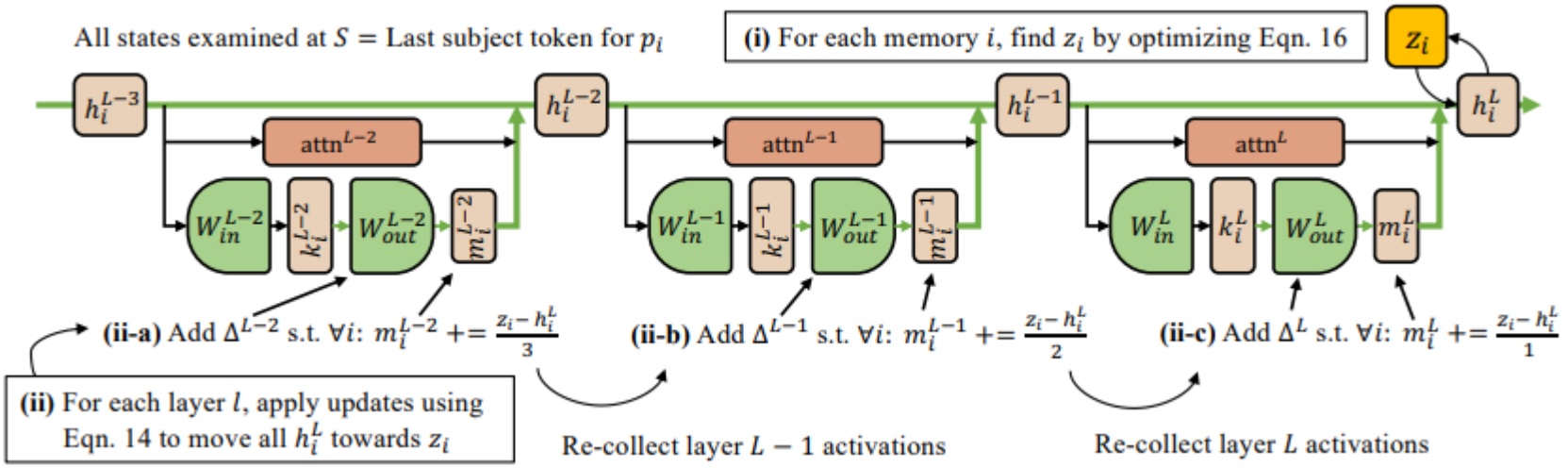
which expands to: $(W_0 + \Delta)(K_0 K_0^T + K_1 K_1^T) = M_0 K_0^T + M_1 K_1^T$

$$\Delta = R K_1^T (C_0 + K_1 K_1^T)^{-1}. \quad C_0 = \lambda \cdot \mathbb{E}_k \left[k k^T \right]$$

subtracting Eqn. 8 from Eqn. 12 : $\Delta(K_0 K_0^T + K_1 K_1^T) = M_1 K_1^T - W_0 K_1 K_1^T.$

Meng, Kevin, et al. "Mass-editing memory in a transformer." arXiv preprint arXiv:2210.07229 (2022).

4. Mass-Editing



$$\text{setting } \hat{W}_{out}^l := W_{out}^l + \Delta^l \text{ for all } l \in \mathcal{R} \text{ optimizes } \min_{\{\Delta^l\}} \sum_i \left\| z_i - \hat{h}_i^L \right\|^2$$

$$k_i^l = \frac{1}{P} \sum_{j=1}^{P} k(x_j + s_i), \text{ where } k(x) = \sigma\left(W_{in}^l \, \gamma\left(h_i^{l-1}(x)\right)\right)$$

$$m_i^l = W_{out} k_i^l + r_i^l \text{ where } r_i^l \text{ is the residual given by } \frac{z_i - h_i^L}{L - l + 1}$$

Meng, Kevin, et al. "Mass-editing memory in a transformer." arXiv preprint arXiv:2210.07229 (2022).

4. Mass-Editing

Algorithm 1: The MEMIT Algorithm

Data: Requested edits $\mathcal{E} = \{(s_i, r_i, o_i)\}$, generator G, layers to edit $\mathcal{S}$, covariances C^l
Result: Modified generator containing edits from $\mathcal{E}$

for $s_i, r_i, o_i \in \mathcal{E}$ **do** // Compute target z_i vectors for every memory i
  hook $G(h_i^L \mathrel{+}= \delta_i)$
  optimize $\operatorname{argmin}_{\delta_i} \frac{1}{P} \sum_{j=1}^{P} -\log \mathbb{P}_{G(h_i^L \mathrel{+}= \delta_i)} [o_i \mid x_j \oplus p(s_i, r_i)]$ (Eqn. 16)
  $z_i \leftarrow h_i^L + \delta_i$
end
for $l \in \mathcal{R}$ **do** // Perform update: spread changes over layers
  $h_i^l \leftarrow h_i^{l-1} + a_i^l + m_i^l$ (Eqn. 2) // Run layer l with updated weights
  for $s_i, r_i, o_i \in \mathcal{E}$ **do**
    $k_i^l \leftarrow k_i^l = \frac{1}{P} \sum_{j=1}^{P} k(x_j + s_i)$ (Eqn. 19)
    $r_i^l \leftarrow \frac{z_i - h_i^L}{L - l + 1}$ (Eqn. 20) // Distribute residual over remaining layers
  end
  $K^l \leftarrow [k_i^{l_1}, ..., k_i^L]$
  $R^l \leftarrow [r_i^{l_1}, ..., r_i^L]$
  $\Delta^l \leftarrow R^l {K^l}^T (C^l + K^l {K^l}^T)^{-1}$ (Eqn. 14)
  $W^l \leftarrow W^l + \Delta^l$ // Update layer l MLP weights in model
end

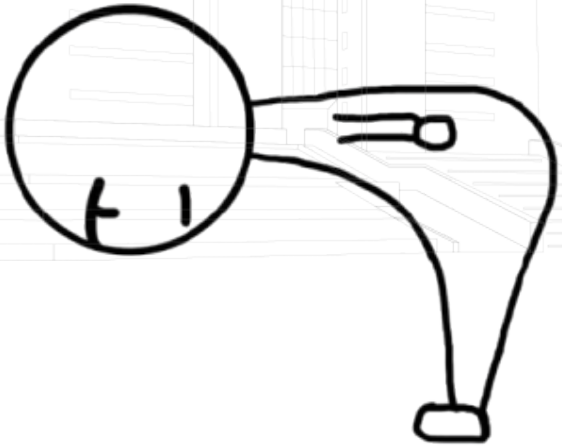
Meng, Kevin, et al. "Mass-editing memory in a transformer." arXiv preprint arXiv:2210.07229 (2022).

5 Discussion

- Summary
  - Factual associations in GPT
  - Knowledge in Feed-Forward
  - Locate with perturbation
  - Edit with data / fine-tune (including editing)
  - Avoid forgetting
  - Algorithm complexity

5 Discussion

- Definition of knowledge
- Form of knowledge
- Location of knowledge
- Knowledge transfer and injection



Thanks

Data Mining Lab, Big Data Research Center, USETC

Wei Han, wei.hb.han@gmail.com