Incorporação de passagens de texto a partir da edição de associações factuais em LLMs

membro(s):

Eduardo Seiti de Oliveira, RA 940011

Descrição do Projeto

- Incorporação em um LLM pré-treinado das informações contidas em uma passagem de texto arbitrária — não contida no conjunto de pré-treino — utilizando a técnica de edição de associações factuais descritas em [1] e [2].

- Comparação do resultado (performance e tempo) com RAG.

Técnica: Identificação dos *Causal Traces*

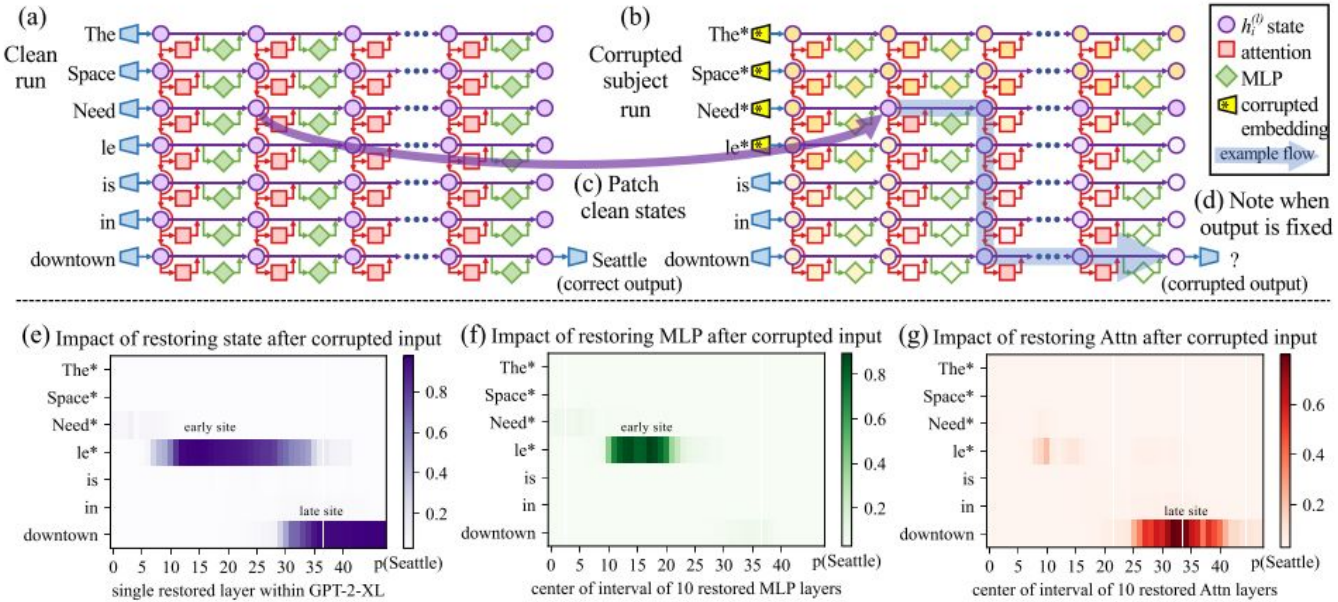

Figure 1: **Causal Traces** compute the causal effect of neuron activations by running the network twice: (a) once normally, and (b) once where we corrupt the subject token and then (c) restore selected internal activations to their clean value. (d) Some sets of activations cause the output to return to the original prediction; the light blue path shows an example of information flow. The causal impact on output probability is mapped for the effect of (e) each hidden state on the prediction, (f) only MLP activations, and (g) only attention activations.

Reproduzido de [1]

Técnica: Edição das associações factuais

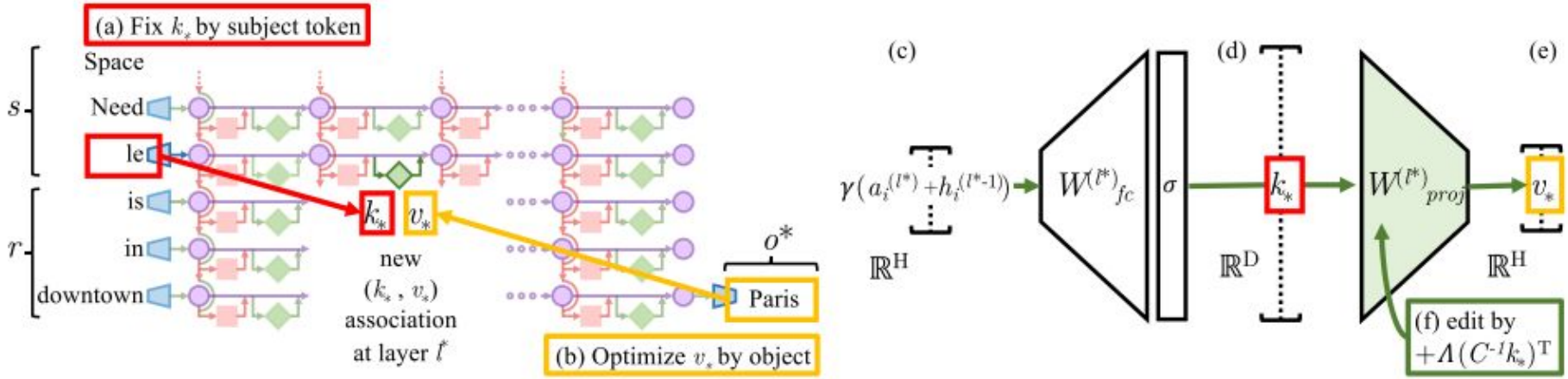

Figure 4: **Editing one MLP layer with ROME**. To associate *Space Needle* with *Paris*, the ROME method inserts a new (k_*, v_*) association into layer l^*, where (a) key k_* is determined by the subject and (b) value v_* is optimized to select the object. (c) Hidden state at layer l^* and token i is 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.

Reproduzido de [1]

Metodologia

1. Reprodução da técnica proposta em um novo LLM — ex. MS Phi-1.5.
2. Avaliação da técnica na incorporação de fatos compostos.
   - Na tripla (*sujeito*, *relação*, *objeto*) que representa uma associação factual, ter *objeto* com um termo composto; os trabalhos originais avaliam objetos de apenas 1 palavra.
3. Avaliação da técnica na geração de texto a partir de diversos fatos inseridos para um mesmo sujeito.
   - Verificar se as associações inseridas podem ser utilizadas livremente pelo LLM no processo de geração de texto.
4. Avaliação da técnica na incorporação do conhecimento contido em uma passagem de texto.
   - Utilização de LLM (Groq LlaMA 80B) para decomposição de passagem arbitrária de texto em sequência de associações factuais a serem inseridas.
   - Comparação da performance com RAG.

Datasets

Embora a técnica proposta não inclua retreinamento do LLM, será necessário um (pequeno) conjunto de dados para a realização dos testes. Esse conjunto pode utilizar como base o dataset *Biographies*, proposto em [3].

Com o suporte de LLM (Groq LlaMA 80B) serão criadas perguntas focadas para cada uma das avaliações propostas.

Métricas

- RAGAS, onde:
  - Associações factuais serão os Contextos;
  - Perguntas e Respostas terão sido geradas pelo LLM.

Resultados

Resultados **esperados** se for a primeira entrega:

- Método de edição de associações factuais configurado com sucesso para novo LLM.
- Compreensão aprofundada do método aplicado.

Resultados **preliminares** se forem entregas intermediárias:

- Avaliação da edição de associações factuais para objetos compostos e múltiplos fatos por sujeito — expectativa de que não ocorra perda, e o modelo editado seja capaz de responder as perguntas corretamente.

Resultados **finais** se for entrega final:

- Avaliação da edição de associações factuais para incorporar conteúdo de uma passagem completa — expectativa de que não ocorra perda, e da definição de relação de custo (tempo de execução) entre edição x RAG.

Referências

1. Meng, Kevin, David Bau, Alex Andonian, and Yonatan Belinkov. "Locating and editing factual associations in GPT." Advances in Neural Information Processing Systems 35 (2022): 17359-17372.
2. Meng, Kevin, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. "Mass-editing memory in a transformer." arXiv preprint arXiv:2210.07229 (2022).
3. Du, Yilun, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. "Improving factuality and reasoning in language models through multiagent debate." arXiv preprint arXiv:2305.14325 (2023).

Cronograma

Lista de atividades a serem feitas antes de cada entrega:

- 06 de junho - entrega I — Plano de Trabalho;
- 13 de junho - entrega II — Técnica aplicada a novo LLM;
- 20 de junho - entrega III — Avaliações fatos compostos e múltiplos fatos;
- 27 de junho - entrega final — Avaliação da incorporação de passagem.