ROME: Locating and Editing Factual Associations in GPT

Kevin Meng, David Bau, Alex Andonian, Yonatan Belinkov

NeurIPS 2022

The rise of Large Language Models

GPT3, PaLM, ChatGPT, …

… and their lack of interpretability, factuality, hallucinations

We are interested how and where a model stores its factual associations, for two reasons:

1. To understand huge opaque neural networks. The internal computations of large language models are obscure. Clarifying the processing of facts is one step in understanding massive transformer networks.
2. Fixing mistakes. Models are often incorrect, biased, or private, and we would like to develop methods that will enable debugging and fixing of specific factual errors.

Factual Associations

fact tuple: **(s**, r, ***o*)** – **subject**, relation, ***object***

s = Edmund Neupert

r = plays the instrument

o = piano

GPT predictions

Edmund Neupert, performing on the ***piano***

Miles Davis plays the ***trumpet***

Niccolo Paganini is known as a master of the ***violin***

Jimi Hendrix, a virtuoso on the ***guitar***

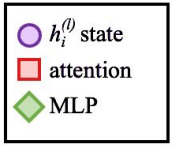
Where and how do the language models store facts?

Hypothesis: MLPs are recalling the factual knowledge

Inspiration in prior work:

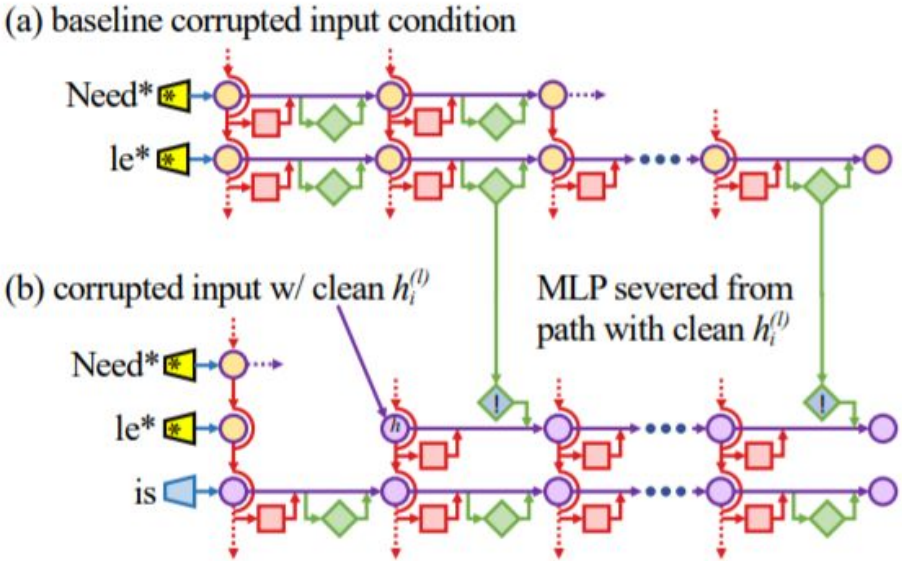
- Softmax Linear Units from Anthropic
- Transformer Feed-Forward Layers Are Key-Value Memories (ACL 2021)

Locating facts by Causal Tracing

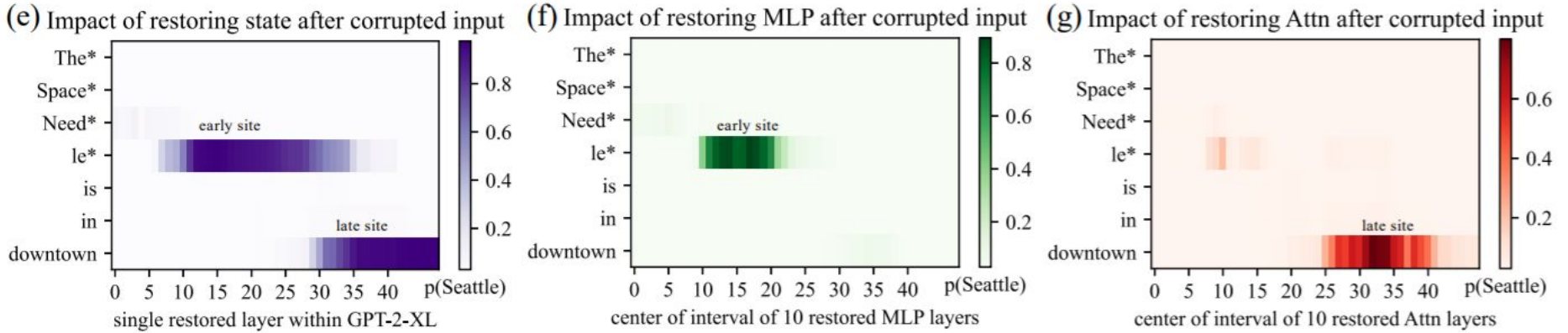


Causal effects of the MLP

If the MLPs are important, what if we just don't let them read their input and show the corrupted input instead?

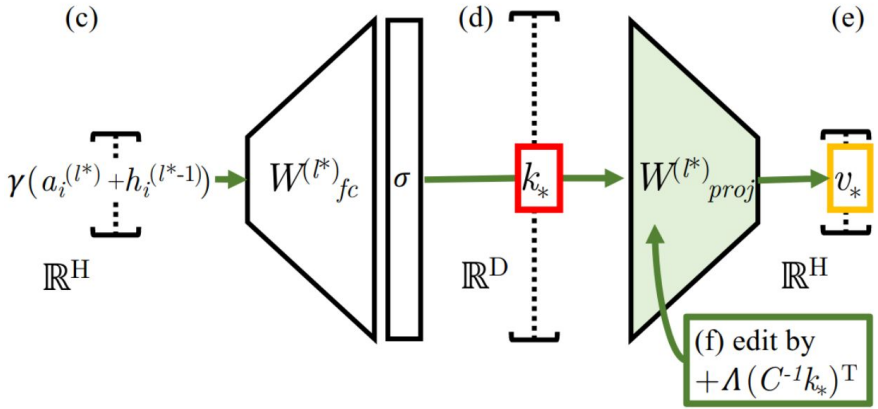


It's all in the MLPs!



But… can we edit these facts?

MLPs as key-value associative memories



Replacing the model's association of an object

We would like to replace an association present in a given MLP: (subject,relation,object) to another one: (subject,relation,object2). This corresponds to replacing the key-value pair induced by the MLP. We do this by modifying the second dense layer.

Steps for replacing key-value pair (k,v) to (k,v*):

1. *Find k that corresponds to the subject:* Choose k to select the subject part of the sentence - look what happens to the subject token (just cache activation in the MLP of interest)
2. *Find v* that corresponds to object2*: "adversarially" (just by normal gradient ascent) modify v, by maximizing probability of object2 instead of object as output of the model
3. Find W' that corresponds to this (k,v*) pair by: (Lambda is calculated from k,v*,W; C is a fuzzy constant)

$$\text{minimize } \|\hat{W}K - V\| \text{ such that } \hat{W}k_* = v_* \quad \text{by setting } \hat{W} = W + \Lambda(C^{-1}k_*)^T.$$

Editing the facts - details

How do we get keys and values??

- Key: show the model the subject and get the input to the MLP
- Value: find the value v that maximizes the probability of the desired object (by backprop)

How to choose the particular MLP?

Causal Tracing returns a range of MLP - they hand pick one layer (17 for GPT2-XL) and show it works surprisingly well

How to Distinguish Knowing a Fact from Saying a Fact?

1. **Specificity** means that when your knowledge of a fact changes, it doesn't change other facts. For example, after learning that the Eiffel Tower is in Rome, you shouldn't also think that every other tourist attraction is also in Rome.
2. **Generalization** means that your knowledge of a fact is robust to changes in wording and context. After learning the Eiffel Tower is in Rome, then you should also know that visiting it will require travel to Rome.

The Space Needle is in Rome.

The Space Needle is located in… (Paraphrase Generalization)

How can I get to the Space Needle ? (Consistency Generalization)

What is there to eat near the Space Needle ? (Consistency Generalization)

Where is the Sears Tower? (Specificity)

CounterFact: benchmarking fact editing

- 22K counterfactual edit examples like “The Space Needle is located in Rome”
- To evaluate generalization of each edit, 3 types of prompts
- The dataset automatically generated using a knowledge graph (no human labeling)

Type	Description	Example(s)	Evaluation Strategy
Counterfactual Statement	A subject-relation-object fact tuple	*The Space Needle is located in Rome.*	Check next-token continuation probs for correct answer
Paraphase Prompts	Direct rephrasings of the same fact	*Where is the Space Needle?* *The Space Needle is in…*	
Neighborh. Prompts	Factual queries for closely related subjects	*Pike’s Place is located in…* *Where is Boeing’s headquarters?*	
Generation Prompts	Prompts that implicitly require knowledge of the counterfactual	*Where are the best places to eat lunch near the Space Needle?* *How can I get there?*	Generate text and compare statistics with text about target

Results on CounterFact

Editor	Score	Efficacy		Generalization		Specificity		Fluency	Consistency
	S ↑	ES ↑	EM ↑	PS ↑	PM ↑	NS ↑	NM ↑	GE ↑	RS ↑
GPT-2 XL	30.5	22.2 (0.9)	-4.8 (0.3)	24.7 (0.8)	-5.0 (0.3)	78.1 (0.6)	5.0 (0.2)	626.6 (0.3)	31.9 (0.2)
FT	65.1	100.0 (0.0)	98.8 (0.1)	87.9 (0.6)	46.6 (0.8)	**40.4 (0.7)**	**-6.2 (0.4)**	607.1 (1.1)	40.5 (0.3)
FT+L	66.9	99.1 (0.2)	91.5 (0.5)	**48.7 (1.0)**	28.9 (0.8)	70.3 (0.7)	3.5 (0.3)	621.4 (1.0)	37.4 (0.3)
KN	**35.6**	**28.7 (1.0)**	**-3.4 (0.3)**	**28.0 (0.9)**	**-3.3 (0.2)**	72.9 (0.7)	3.7 (0.2)	**570.4 (2.3)**	**30.3 (0.3)**
KE	52.2	84.3 (0.8)	33.9 (0.9)	75.4 (0.8)	14.6 (0.6)	**30.9 (0.7)**	**-11.0 (0.5)**	**586.6 (2.1)**	31.2 (0.3)
KE-CF	**18.1**	99.9 (0.1)	97.0 (0.2)	95.8 (0.4)	59.2 (0.8)	**6.9 (0.3)**	**-63.2 (0.7)**	**383.0 (4.1)**	**24.5 (0.4)**
MEND	57.9	99.1 (0.2)	70.9 (0.8)	65.4 (0.9)	12.2 (0.6)	**37.9 (0.7)**	**-11.6 (0.5)**	**624.2 (0.4)**	34.8 (0.3)
MEND-CF	**14.9**	**100.0 (0.0)**	**99.2 (0.1)**	**97.0 (0.3)**	**65.6 (0.7)**	**5.5 (0.3)**	**-69.9 (0.6)**	**570.0 (2.1)**	33.2 (0.3)
ROME	**89.2**	100.0 (0.1)	97.9 (0.2)	96.4 (0.3)	62.7 (0.8)	**75.4 (0.7)**	**4.2 (0.2)**	621.9 (0.5)	**41.9 (0.3)**
GPT-J	23.6	16.3 (1.6)	-7.2 (0.7)	18.6 (1.5)	-7.4 (0.6)	83.0 (1.1)	7.3 (0.5)	621.8 (0.6)	29.8 (0.5)
FT	**25.5**	**100.0 (0.0)**	**99.9 (0.0)**	96.6 (0.6)	71.0 (1.5)	**10.3 (0.8)**	**-50.7 (1.3)**	**387.8 (7.3)**	**24.6 (0.8)**
FT+L	68.7	99.6 (0.3)	95.0 (0.6)	**47.9 (1.9)**	30.4 (1.5)	78.6 (1.2)	**6.8 (0.5)**	**622.8 (0.6)**	35.5 (0.5)
MEND	63.2	97.4 (0.7)	71.5 (1.6)	**53.6 (1.9)**	11.0 (1.3)	53.9 (1.4)	**-6.0 (0.9)**	620.5 (0.7)	32.6 (0.5)
ROME	**91.5**	99.9 (0.1)	99.4 (0.3)	**99.1 (0.3)**	**74.1 (1.3)**	**78.9 (1.2)**	5.2 (0.5)	620.1 (0.9)	**43.0 (0.6)**

Summary of the key findings

1. Factual associations can be localized along three dimensions, to (1) MLP module parameters (2) at a range of middle layers and (3) specifically during processing of the last token of the subject.

2. Individual factual associations can be changed by making small rank-one changes in a single MLP module. We can distinguish between changes in *knowledge* versus superficial changes in language by measuring generalization to other wordings of the same fact.

Thank you for your attention!

Prepared by Szymon Tworkowski & Szymon Antoniak

Sources for the slides:

- [1] https://arxiv.org/abs/2202.05262
- [2] https://www.youtube.com/watch?v=_NMQyOu2HTo - video by Yannic Kilcher
- [3] https://rome.baulab.info/
- [4] https://slideslive.com/38990940

Recent follow-up paper by the same authors: https://arxiv.org/abs/2210.07229