STAT 992: Science of Large Language Models

Lecture 9: PCA and factor analysis

Spring 2026
Yiqiao Zhong

Recap: LRH, low-rankness

- **Linear representation hypothesis:** transformers represent concepts as low-dim linear subspaces (esp. vectors) in the hidden states space

- **Low-rankness**: Weight matrix spectra are mostly power law distributed

$$\text{“apple”} = 0.09 \text{ “dessert”} + 0.11 \text{ “organism”} + 0.16 \text{ “fruit”} + 0.22 \text{ “mobile\&IT”} + 0.42 \text{ “others”}.$$

- Caveat: long tails in spectra do matter, they may store rich and diverse knowledge in language

Table edited by Gemini-3

MSA, connection between LRH and low-rankness

- A clean formula for multihead self-attention (MSA): given a hidden state $\boldsymbol{h} \in \mathbb{R}^d$ at a given layer and give position, MSA computes

$$\boldsymbol{h} \longleftarrow \boldsymbol{h} + \sum_j \boldsymbol{W}_j \boldsymbol{\varphi}_j (\tilde{\boldsymbol{W}}_j \boldsymbol{h})$$

- $\boldsymbol{W}_j, \tilde{\boldsymbol{W}}_j \in \mathbb{R}^{d \times d}$ are <u>low-rank</u> matrices, given respectively by value & output weight matrices and key & query weight matrices from an attention head
- $\boldsymbol{\varphi}_j$ is a map that depends on the context history, namely all hidden states from previous positions at the same layer
- A simplified view: ignoring layer normalization, relative positional embedding (RoPE), etc., but interpretability is mostly correct

Interpreting MSA, LRH, and low-rankness

- A clean formula for multihead self-attention (MSA): given a hidden state $\boldsymbol{h} \in \mathbb{R}^d$ at a given layer and give position, MSA computes

$$\boldsymbol{h} \longleftarrow \boldsymbol{h} + \sum_j \boldsymbol{W}_j \boldsymbol{\varphi}_j(\tilde{\boldsymbol{W}}_j \boldsymbol{h})$$

- $\tilde{\boldsymbol{W}}_j$ extracts relevant “concepts” from hidden state (residual stream)
- $\boldsymbol{\varphi}_j$ transforms nonlinearly based on pairwise interaction previous hidden states (representing the context)
- $\boldsymbol{W}_j$ adds high-order “concepts” that interact with other tokens, e.g., previous-token head binds previous token
- Low-rank weight matrices and low-dim “concept” subspaces are connected

Hidden state under factor-analysis view

- A factor-analysis view of hidden states

$$\boldsymbol{h} = a_1\boldsymbol{v}_1 + a_2\boldsymbol{v}_2 + \ldots$$

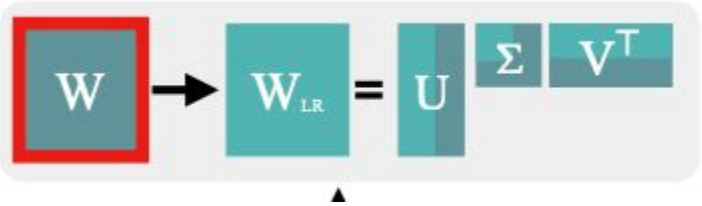
- $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots$ represent concept vectors
- $a_1, a_2, \ldots \in \mathbb{R}$ are activation values of concepts. The activations change as we vary the input sequence
- Concept vectors represent refined semantics and patterns through TF layers
  - Static knowledge is enriched as hidden state is processed by multiple layer (MLP often viewed as main contributors).
  - Mixture-of-experts is motivated as MLP submodules for specialized knowledge
  - MSA forms dynamic concepts by binding selected tokens, forming context-sensitive pattern matching
  - MSA is good at capturing formats and structures from context, e.g., copying, reverse, similar to implementing algorithms with context

Examples of PCA as a tool in LLMs

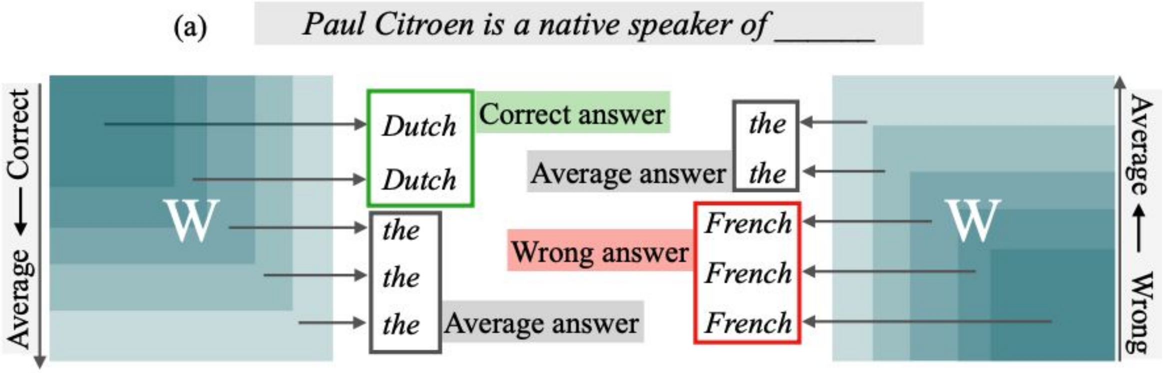
Frequency-related concepts via SVD on weight matrices

- SVD decomposes a weight matrix

$$\boldsymbol{W} = \sum_k \sigma_k \boldsymbol{u}_k \boldsymbol{v}_k^\top$$



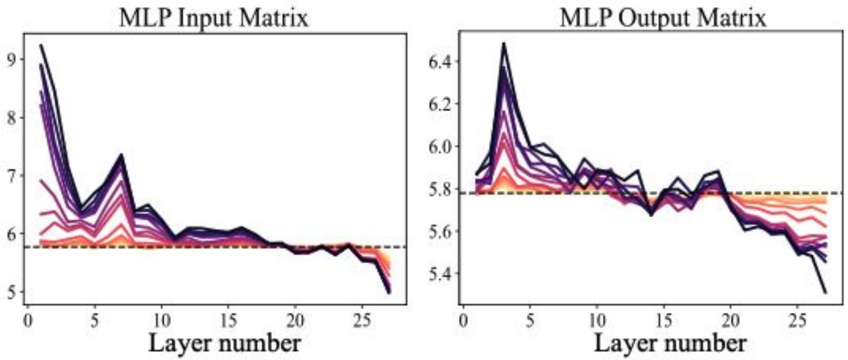
- Top singular value/vector components gives the best low-rank approximation
- Large/small singular value/vector components encode low-frequency / high-frequency words



[The Truth is in There: Improving Reasoning in Language Models with Layer-Selective Rank Reduction](), 2023

Frequency-related concepts via SVD on weight matrices

- Low-rank approximation in later layers sometimes improve downstream accuracy
- A specific case of LoRA

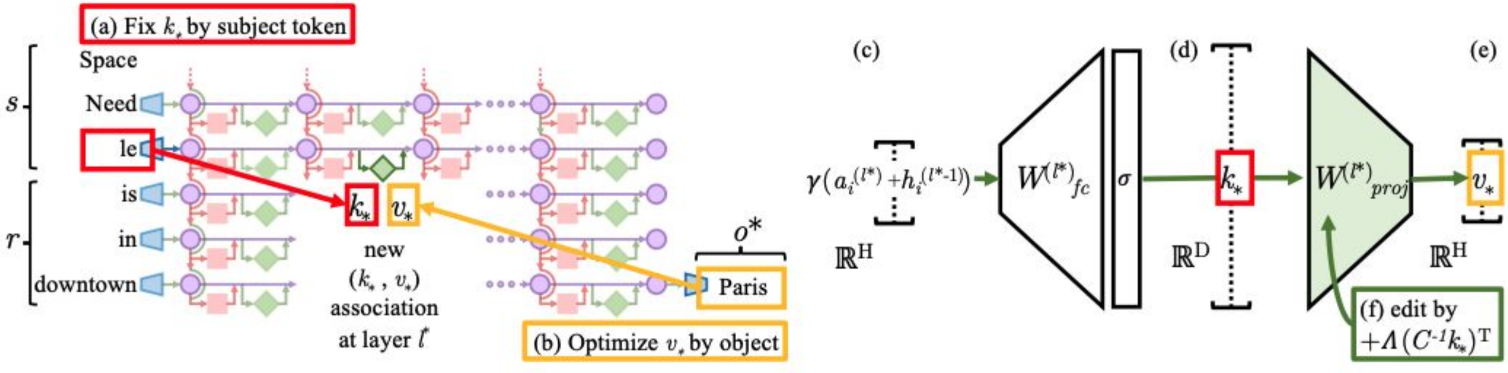


Dataset		Roberta		GPT-J		LLama2	
			LASER		LASER		LASER
CounterFact	Acc	17.3	**19.3**	13.1	**24.0**	35.6	**37.6**
	Loss	5.78	**5.43**	5.78	**5.05**	3.61	**3.49**
HotPotQA	Acc	6.1	**6.7**	**19.6**	19.5	16.5	**17.2**
	Loss	10.99	**10.53**	3.40	**3.39**	3.15	**2.97**
FEVER	Acc	50.0	**52.3**	50.2	**56.2**	59.3	**64.5**
	Loss	2.5	**1.76**	**1.24**	1.27	1.02	**0.91**
Bios Gender	Acc	87.5	**93.7**	70.9	**97.5**	75.5	**88.4**
	Loss	**0.87**	1.13	**3.86**	4.20	3.48	**2.93**
Bios Profession	Acc	64.5	**72.5**	75.6	**82.1**	85.0	**86.7**
	Loss	**4.91**	6.44	**4.64**	4.91	4.19	**4.05**
TruthfulQA	Acc	56.2	56.2	54.9	**55.6**	50.5	**56.2**
	Loss	1.60	**1.42**	1.02	**1.01**	**0.95**	1.04
BigBench-Epistemic Reasoning	Acc	37.1	**41.8**	37.1	**38.3**	44.8	**63.4**
	Loss	9.39	**6.80**	0.74	**0.62**	0.78	**0.73**
BigBench-WikidataQA	Acc	28.0	**30.7**	51.8	**65.9**	59.5	**62.0**
	Loss	9.07	**7.69**	3.52	**2.86**	2.40	**2.31**

The Truth is in There: Improving Reasoning in Language Models with Layer-Selective Rank Reduction, 2023

Editing concepts via one-rank update in weight matrices

- Another form of LoRA: instead of gradient-based fine-tuning, directly editing weights
- Surgery on MLP component: find two concept-encoding vectors, then perform rank-one update to weight matrix
- Caveat: performance on other tasks often degrade with successive edits

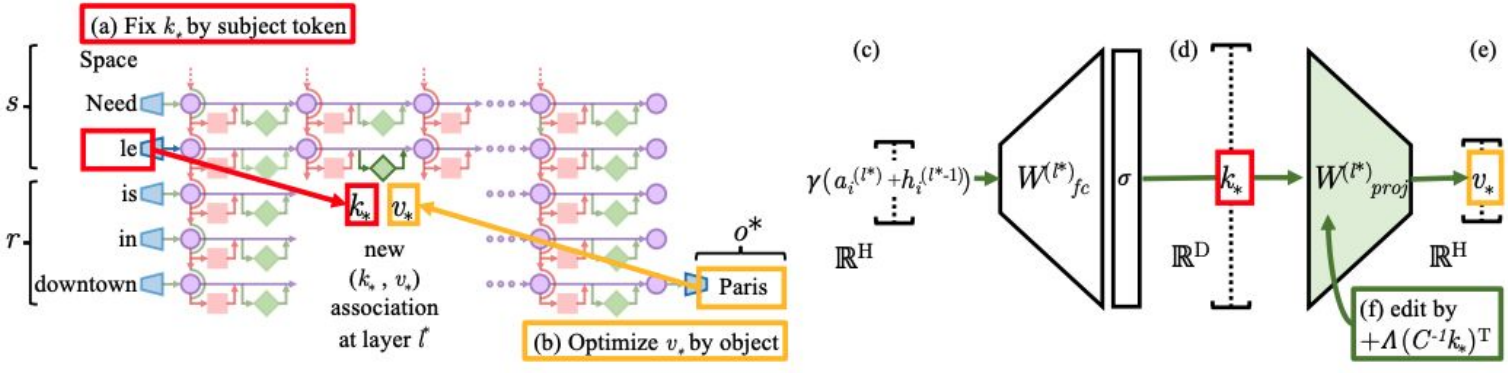


[Locating and Editing Factual Associations in GPT, 2023](#)

Editing concepts via one-rank update in weight matrices

- Associative memory (key–value store) in MLP: key k_* vector retrieves and extracts relevant concepts, value v_* adds concepts to residual stream
- Insert a knowledge with rank-one update to value matrix

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



[Locating and Editing Factual Associations in GPT, 2023]

Removing harmful concepts via weight matrices edits

- Alignment of LLMs: fine-tuning pretrained base model to produce helpful, polite, unharmful chatbots
- Standard methods: RLHF, DPO
- A case study of toxicity: paired toxic/non-toxic sequences
- Factor-analysis view of hidden states:

$$\boldsymbol{x}_i^+ = \underbrace{a_i^+ \boldsymbol{\mu}}_{\text{stopwords}} + \underbrace{\boldsymbol{B} \boldsymbol{f}_i}_{\text{toxic component}} + \underbrace{\tilde{\boldsymbol{B}} \tilde{\boldsymbol{f}}_i}_{\text{context component}} + \underbrace{\boldsymbol{u}_i^+}_{\text{noise}},$$

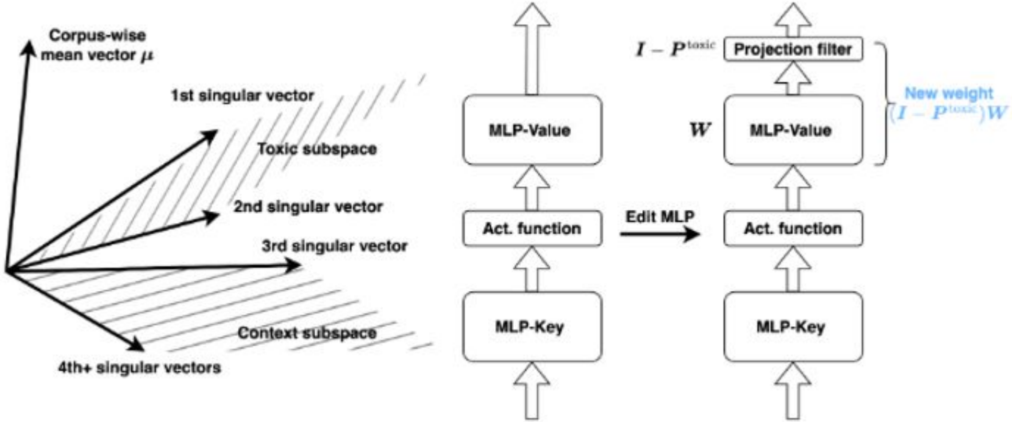
$$\boldsymbol{x}_i^- = a_i^- \boldsymbol{\mu} + \tilde{\boldsymbol{B}} \tilde{\boldsymbol{f}}_i + \boldsymbol{u}_i^-$$

	Top Tokens (Layer 14)	Interpretation
$\boldsymbol{\mu}$	, and the - in (" .	Frequent tokens, stopwords
1st svec	s**t f**k ucker b***h slut F**k holes	Toxic tokens
2nd svec	damn really kinda stupid s**t goddamn	Toxic tokens
3rd svec	disclaimer Opinion LĤ Statement Disclaimer Brief	Context dependent topics
4th svec	nation globalization paradigm continent empire ocracy	Context dependent topics

Model Editing as a Robust and Denoised variant of DPO: A Case Study on Toxicity, 2024

Removing harmful concepts via weight matrices edits

- An observed benefit of editing compared with gradient-based fine-tuning: better sample efficiency



for $\ell \leftarrow L_0$ to L **do**:

Get hidden sentence embeddings at layer l from $\mathcal{D}_{\text{pref}}$: $\boldsymbol{X}_\ell^+, \boldsymbol{X}_\ell^- \in \mathbb{R}^{N \times D}$

Find embedding difference matrix: $\boldsymbol{T}_\ell^0 \leftarrow \left(\boldsymbol{X}_\ell^+ - \boldsymbol{X}_\ell^-\right)$

Remove corpus-wise mean vector: $\boldsymbol{\mu} \leftarrow \text{mean}(\boldsymbol{X}_\ell^-)$ and $\boldsymbol{T}_\ell \leftarrow \boldsymbol{T}_\ell^0 \left(\boldsymbol{I} - \boldsymbol{\mu}\boldsymbol{\mu}^\top / \|\boldsymbol{\mu}\|_2^2\right)$

Find toxic subspace projection matrix by SVD: $\boldsymbol{U\Sigma V}^\top = \boldsymbol{T}_\ell$, $\boldsymbol{P}_\ell^{\text{toxic}} \leftarrow \sum_{i=1}^k \boldsymbol{v}_i \boldsymbol{v}_i^\top$

Edit by projecting away the toxic subspace: $\boldsymbol{W}_\ell^{\text{edited}} \leftarrow \left(\boldsymbol{I} - \boldsymbol{P}_\ell^{\text{toxic}}\right) \mathbf{W}_\ell$

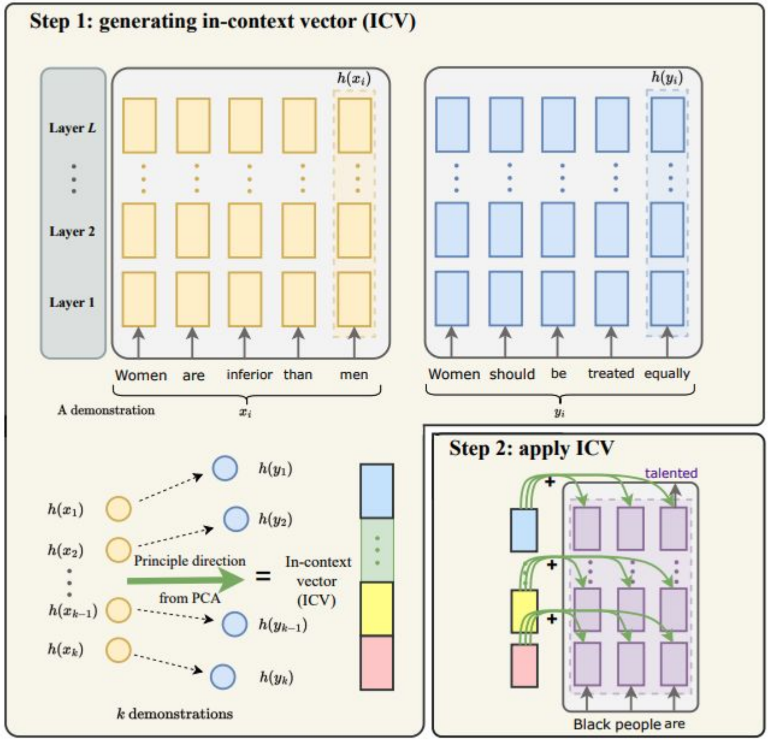
end for

return $\boldsymbol{W}^{\text{edited}}$

[Model Editing as a Robust and Denoised variant of DPO: A Case Study on Toxicity](), 2024

Steering with in-context vector

- Steering: change intermediate hidden states / activations during inference to enhance or suppress certain output
- **In-context vector**: concept-encoding vectors (usually related to a task) calculated from in-context examples



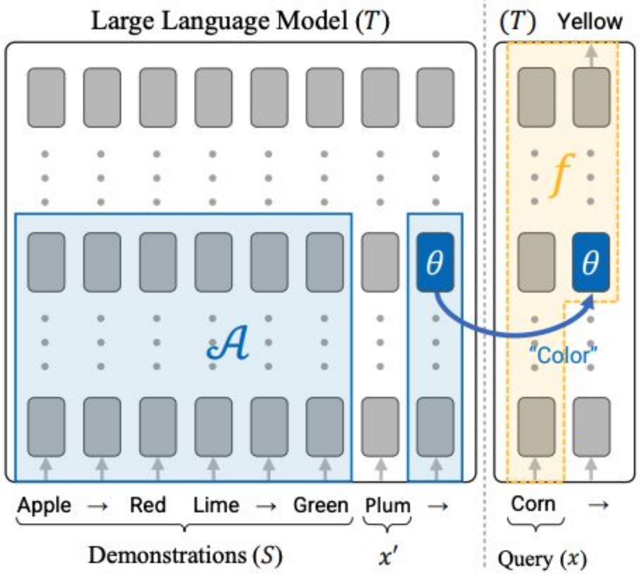
[In-context Vectors: Making In Context Learning More Effective and Controllable Through Latent Space Steering](), 2024

Steering with in-context vector

- Paired examples in context (harmful, unfarmful)

$$\mathcal{X} = \{h(x_1), h(x_2), \ldots, h(x_m)\},$$
$$\mathcal{Y} = \{h(y_1), h(y_2), \ldots, h(y_n)\}.$$

- Apply PCA to find /estimate the task vector



In-Context Learning Creates Task Vectors, 2023

$$\frac{1}{k}\sum_{i=1}^{k}\left(h^\top h(y_i) - h^\top h(x_i)\right)^2. \qquad (2)$$

Lemma 1. *The maximizer of objective Eq.* (2) *subject to* $h^T h = 1$ *is the first principal direction of a set of real-valued data* $\mathcal{D} := \{h(y_1) - h(x_1), h(y_2) - h(x_2), \ldots, h(y_k) - h(x_k)\}$.