STAT 992: Science of Large Language Models

Lecture 7: Layerwise structures of embeddings

Spring 2026

Yiqiao Zhong

Deep neural net is a composition of multiple layers

- Input $\boldsymbol{h}^{(0)} = \boldsymbol{x}$
  - Image: $H \times W \times C$
  - Text: $T \times d$
- Left: pre-ResNet model (–2015)

$$\boldsymbol{h}^{(\ell+1)} = \boldsymbol{f}_\ell(\boldsymbol{h}^{(\ell)})$$

- Right: post-ResNet model (2015–)

$$\boldsymbol{h}^{(\ell+1)} = \boldsymbol{h}^{(\ell)} + \boldsymbol{f}_\ell(\boldsymbol{h}^{(\ell)})$$

- Each layer “processes” the representation in the composition

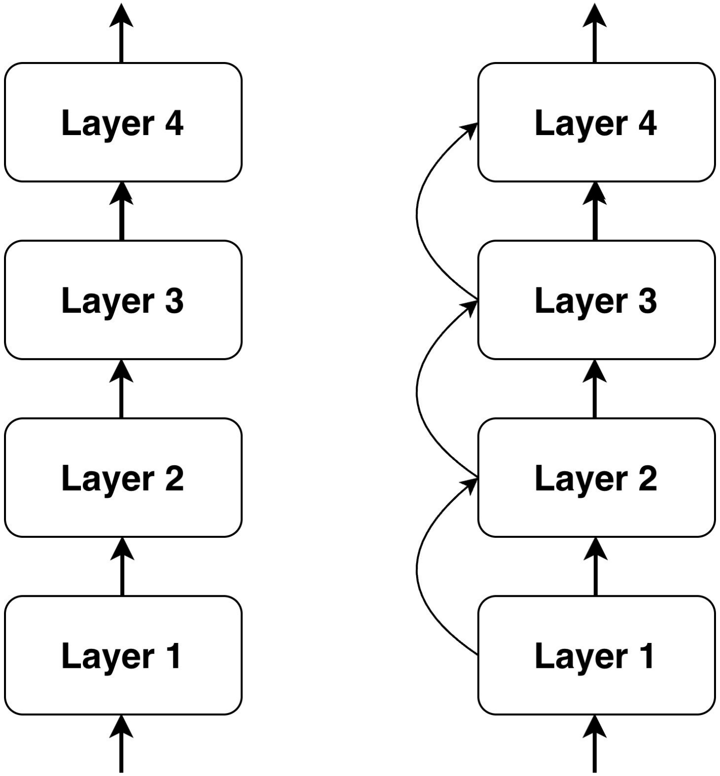
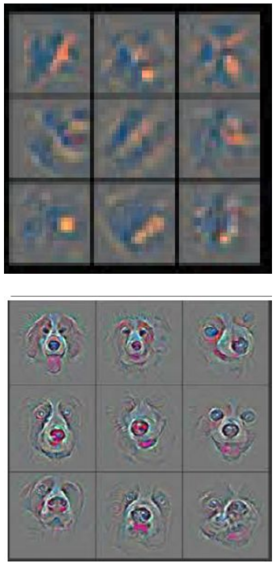


Figure NOT edited by genAI

DNNs generally represent hierarchical features through layer composition

CNN: layers extract <u>hierarchical</u> features

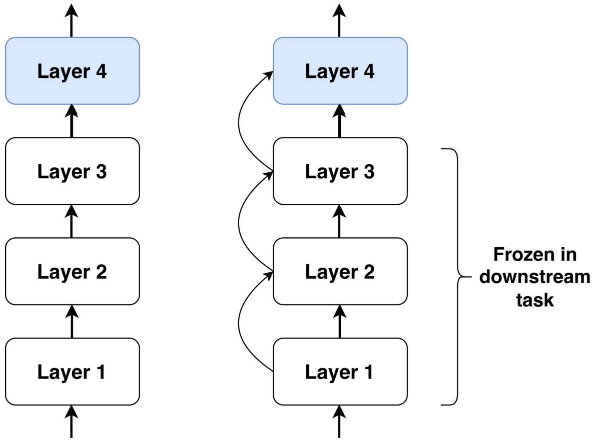
- Soon after AlexNet in 2012, various studies confirmed hierarchical feature representations
  - DeconvNets (see figure): pseudo-inverse map
  - Activation maximization: what input maximizes a feature
  - Grad-CAM: gradient-based input sensitivity
- Lower layers encode wavelet / Gabor filters
- Higher layers encode more abstract concept

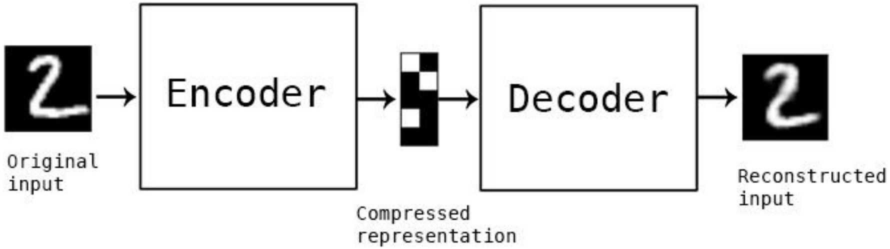



Visualizing 9 randomly selected feature maps in CNN
Layer 1–5
[Visualizing and Understanding Convolutional Networks](), 2013

Hierarchical features throughout DL development

- Transfer learning & fine-tuning
    - Only top layers are optimized in downstream tasks
- Autoencoders and GANs
    - Extracting High-level concept in latent space
- The “magic” of feature learning
    - DL models are not explicitly told to find meaningful features (most trained by minimizing a simple loss
    - They find meaningful features anyway





Source: [link]

Pre-DL methods are poor at hierarchical features

- Manual construction of nonlinear map
  - For example, use rule-based heuristics to compute features given an image or a sentence
- Kernel method
  - Popular in 2000s, choice of kernel determines nonlinear map
- Not scale well with data and dimension
  - Not adaptive to data distribution
  - Curse of dimensionality

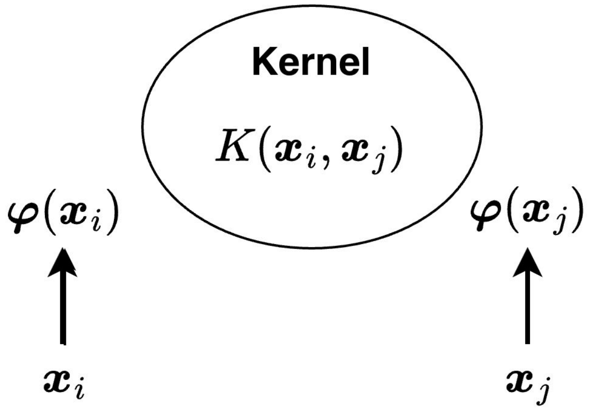
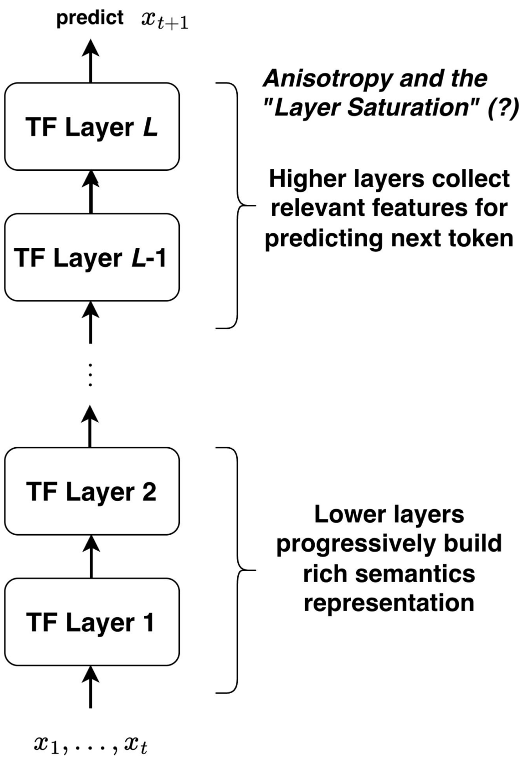


Figure NOT edited by genAI

Layerwise functionality of transformers

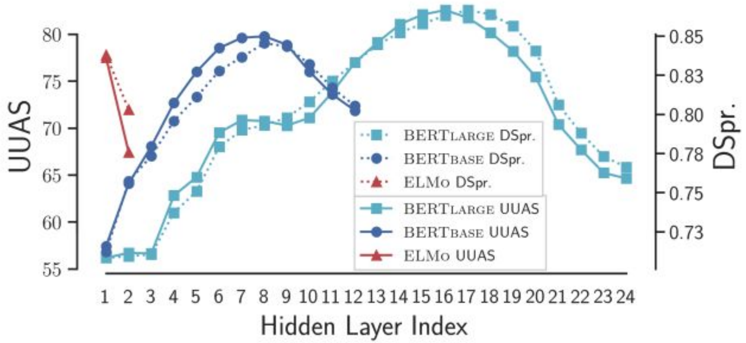
- **Autoregressive training**: minimizing cross-entropy loss reduces mismatch between prediction and actual next token
- **The "Mid-Layer Bottleneck"**: Models build semantic rich features in earlier layers, target next-token prediction in later layers
- **Anisotropy in last few layers**: representations often become highly anisotropic, likely training artifact



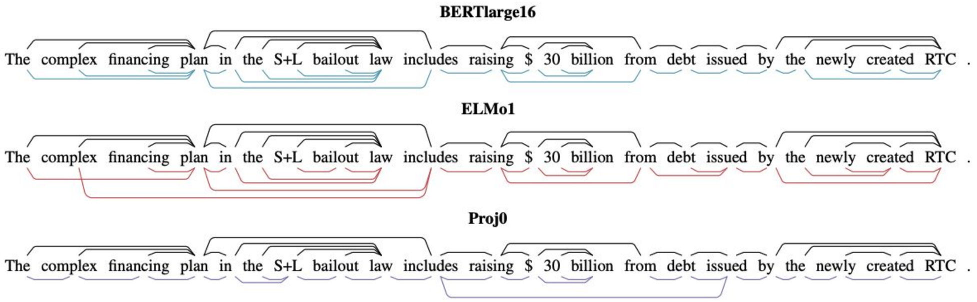
Layerwise analysis of embeddings in transformers

How do transformer embeddings encode synthetics

- Classical NLP has dedicated dataset with annotated syntax tree
- Use trained embeddings (hidden states) distance to construct syntax tree as model's representation of syntax
- Multiple layers help models to find more accurate syntax, peaking at a mid layer



A Structural Probe for Finding Syntax in Word Representations, 2019

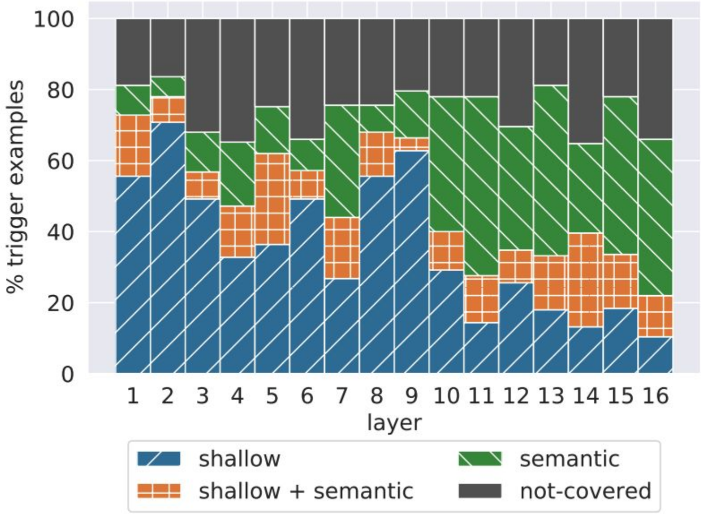


Syntax tree constructed from language models

How do transformers FFNs encode knowledge

- Recall the key-value memory paper
  - Use logit lens (projection to vocab using unembedding matrix) to interpret value vectors
- Early layers tend to trigger shallow concepts, later layers complex concepts

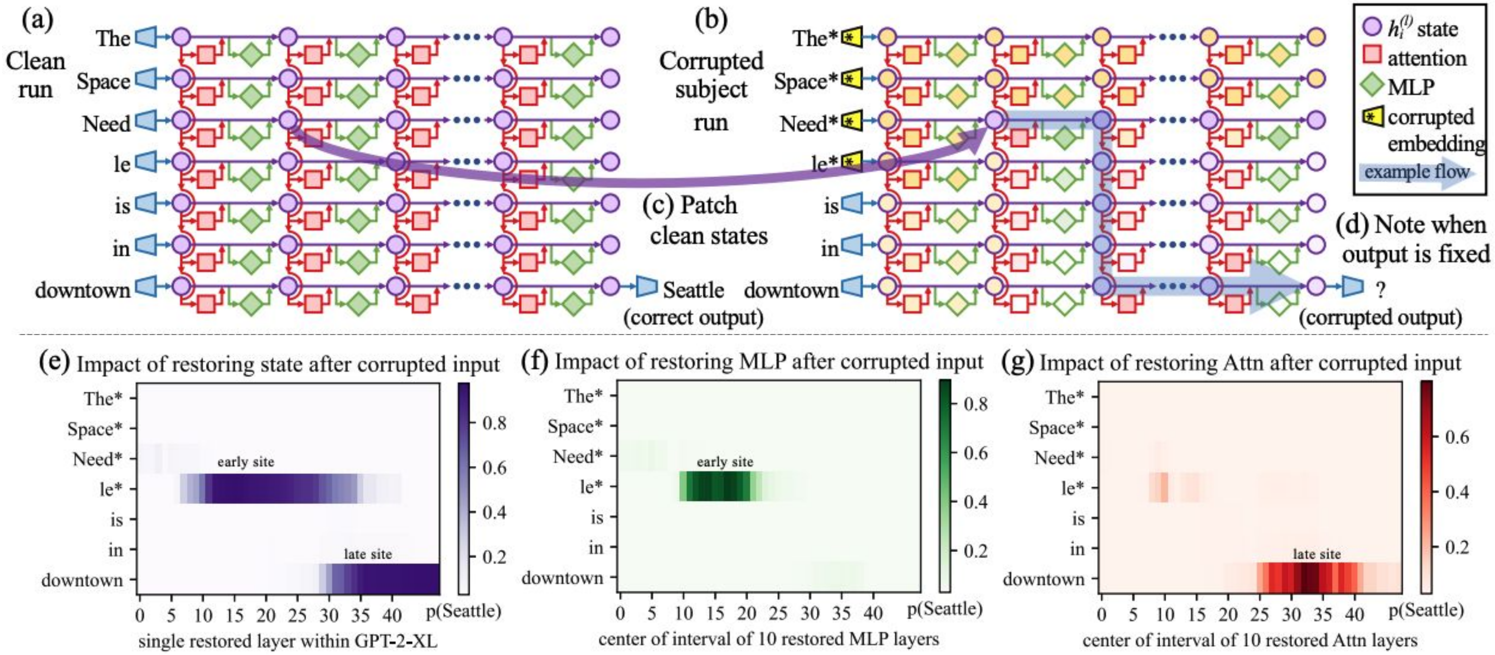
Key	Pattern	Example trigger prefixes
$\mathbf{k}^{1}_{449}$	Ends with "*substitutes*" (shallow)	*At the meeting, Elton said that "for artistic reasons there could be no substitutes* *In German service, they were used as substitutes* *Two weeks later, he came off the substitutes*
$\mathbf{k}^{6}_{2546}$	Military, ends with "*base*"/"*bases*" (shallow + semantic)	*On 1 April the SRSG authorised the SADF to leave their bases* *Aircraft from all four carriers attacked the Australian base* *Bombers flying missions to Rabaul and other Japanese bases*
$\mathbf{k}^{10}_{2997}$	a "part of" relation (semantic)	*In June 2012 she was named as one of the team that competed* *He was also a part of the Indian delegation* *Toy Story is also among the top ten in the BFI list of the 50 films you should*
$\mathbf{k}^{13}_{2989}$	Ends with a time range (semantic)	*Worldwide, most tornadoes occur in the late afternoon, between 3 pm and 7* *Weekend tolls are in effect from 7:00 pm Friday until* *The building is open to the public seven days a week, from 11:00 am to*
$\mathbf{k}^{16}_{1935}$	TV shows (semantic)	*Time shifting viewing added 57 percent to the episode's* *The first season set that the episode was included in was as part of the* *From the original NBC daytime version , archived*



[Transformer Feed-Forward Layers Are Key-Value Memories](), 2021

An intervention approach to interpreting embeddings

- Distinct causal effects between early layers vs later layers, SA vs MLP
- More in future lectures



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

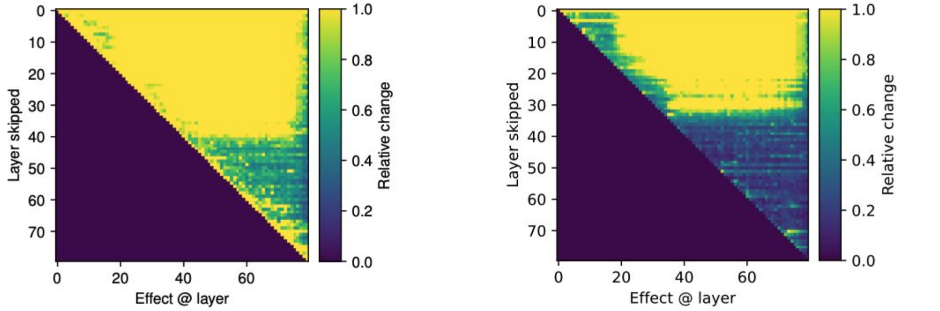
Later layers tend to be additive

- Later layers compute a map $\boldsymbol{h} \mapsto (\varphi_L + \mathbf{id}) \circ \cdots (\varphi_\ell + \mathbf{id}) \circ \boldsymbol{h}$
- Additivity means

$$(\varphi_L + \mathbf{id}) \circ \cdots (\varphi_\ell + \mathbf{id}) \approx \sum_{k=\ell}^{L} \varphi_L + \mathbf{id}$$

- It is possible when $\varphi_k = \boldsymbol{U}_k \circ \varphi'_k \circ \boldsymbol{V}_k^\top$ where $(\boldsymbol{U}_k)$ are orthogonal and $(\boldsymbol{V}_k)$ are orthogonal, i.e., “reading” and “writing” use orthogonal subspaces

- Effects of early layers and later layers tend to be decoupled
  - Redundancy, pruning possible
  - No complex high-order compositions in later laters
  - Refining embeddings for prediction



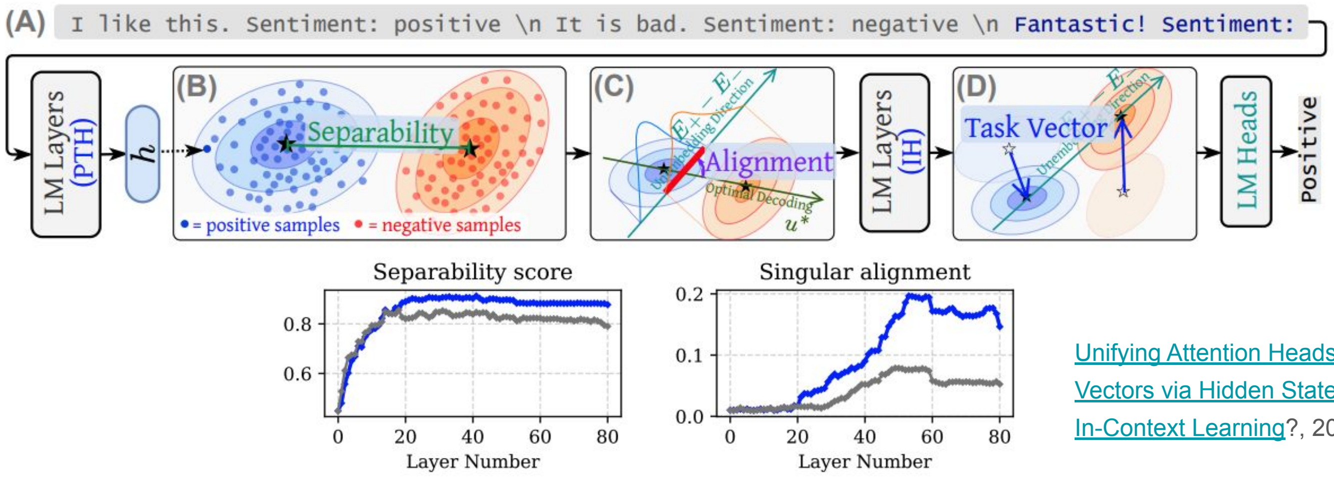
(a) Effect of skipping a layer on later layers’ contributions in the *all* timesteps.

(b) Effect of skipping a layer on later layers’ contributions in *future* timesteps.

Do Language Models Use Their Depth Efficiently?, 2025

Geometric view of layerwise effects

- Early layers: promote <u>separability</u> of concepts (not ready for prediction yet)
- Later layers: increase <u>alignment</u> with unembeddings, gradual angular refinement of embedding
- Analogy: early layers extract high-order interaction like tensors, later layers run a logistic regression on top of sophisticated features



[Unifying Attention Heads and Task Vectors via Hidden State Geometry in In-Context Learning](?), 2025