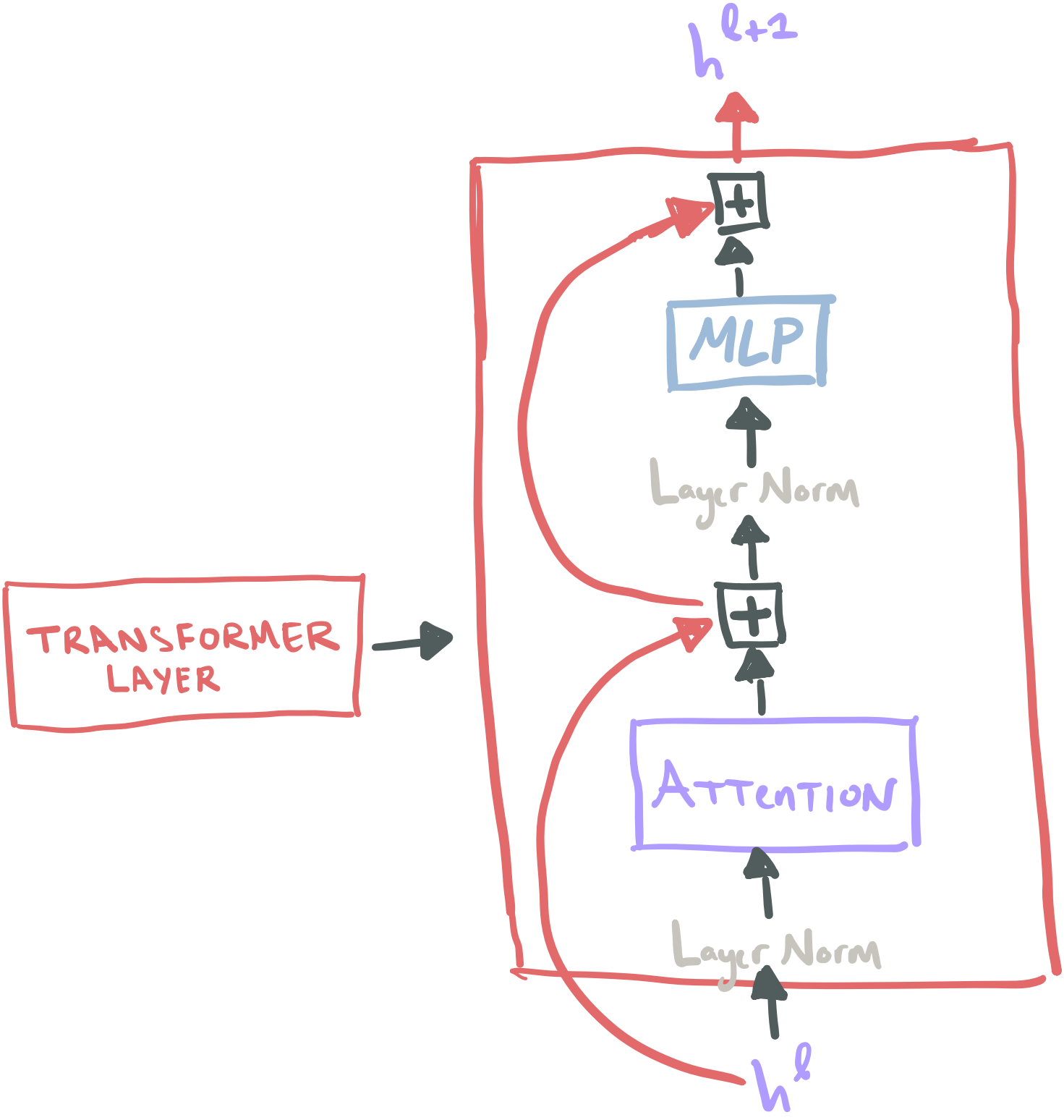
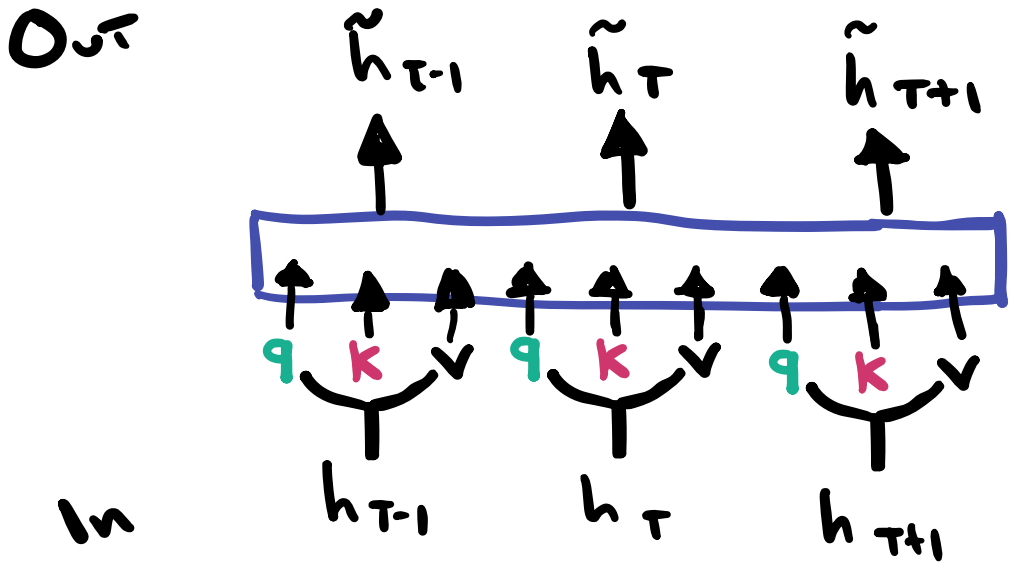
DS 4440 BERT & friends

Last Time: Transformers





In Multi-headed Attention we define H attention heads with independent params W_q^h, W_k^h, W_v^h.

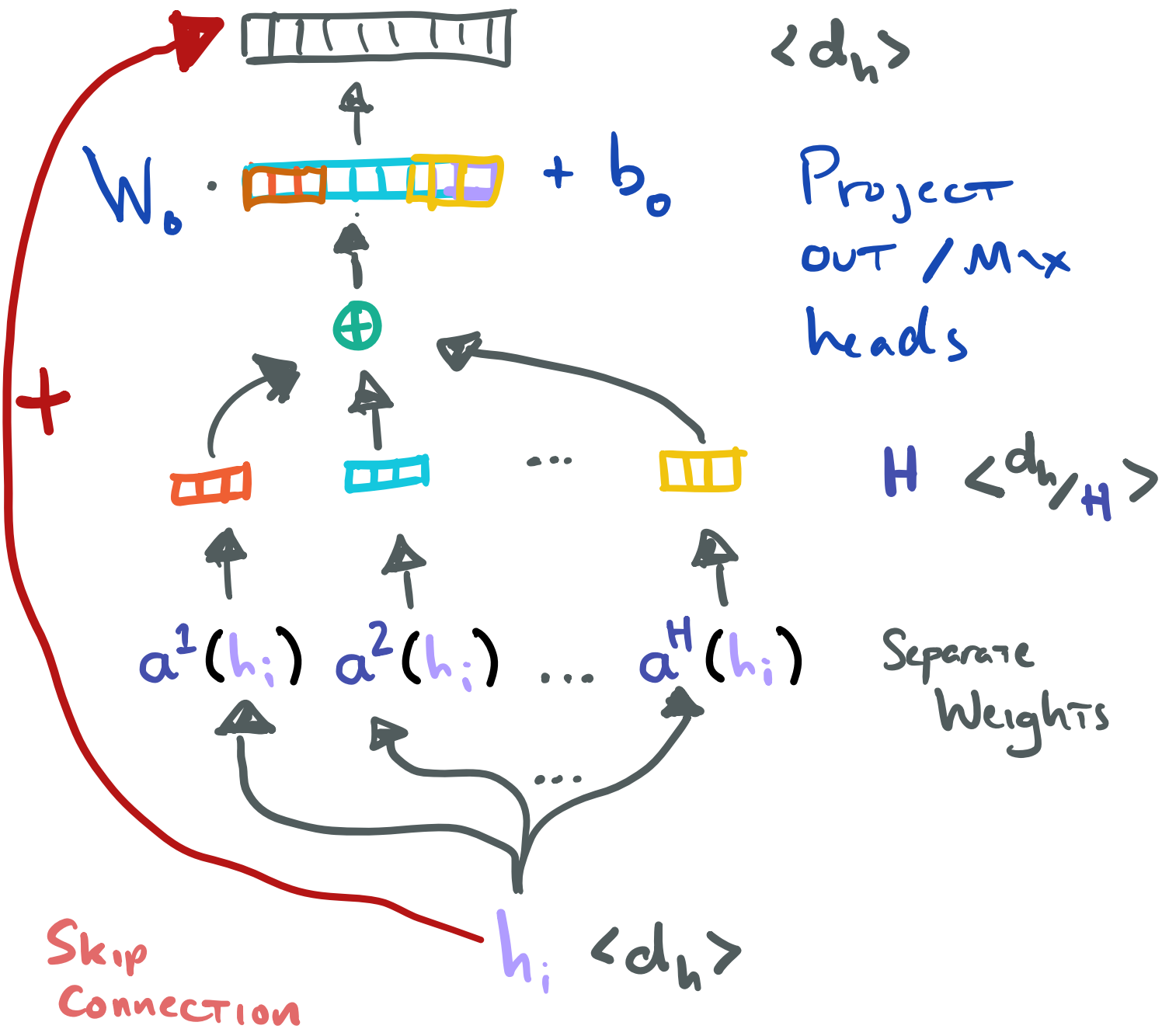
Outputs are combined ($\oplus$).

W_q^h, W_k^h, W_v^h all have shape

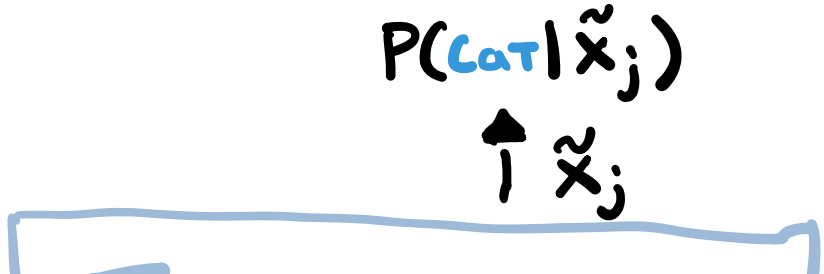
$$\langle d_h \times d_k \rangle$$

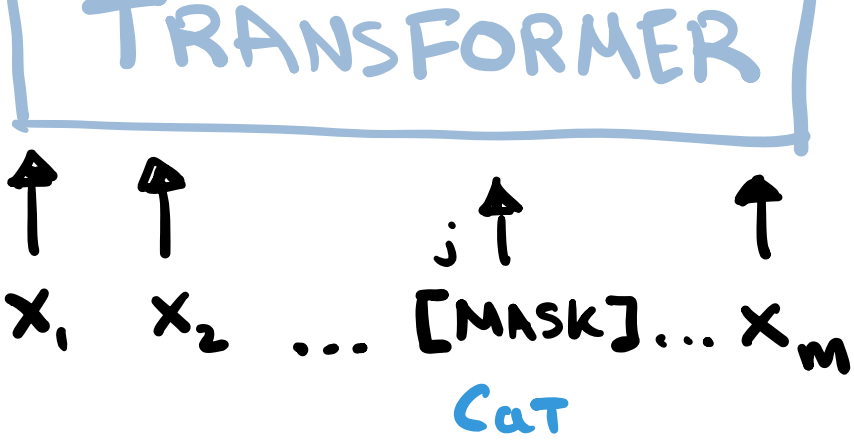
Where $d_k = d_h / H$.

Can view This as H transformations



BERT Bidirectional Encoder Representations from Transformers





"Pre-train" using this Self Supervised objective over lots of text.

How To Mask

"The 1 cat 2 sat 3 on 4 the 5 soft 6 warm 7 mat 8"

[MASK] Purple

Select 15% of positions → corrupt

of these, 80% → [MASK]
10% → Random
10% → Unchanged

} We won't see [MASK] at test time!

$$\hat{y}_i = \underbrace{W_U}_{V \times d_h} g(h_i^{(L)}) + b_U$$

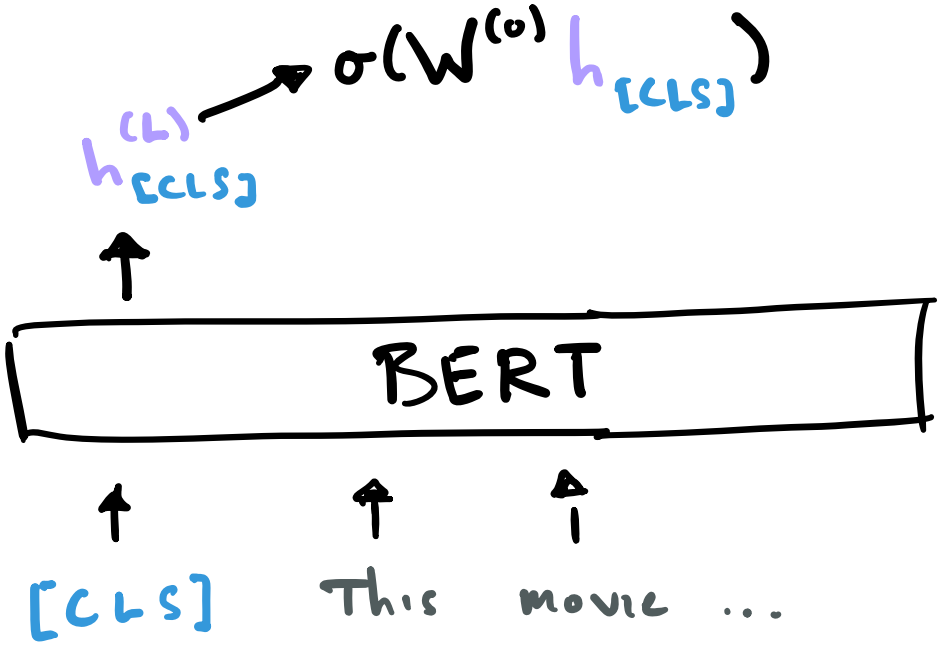
$$\mathcal{L} = -\sum_{i \in S} \log \text{softMax}(\hat{y}_i)_{x_i}$$

Special tokens

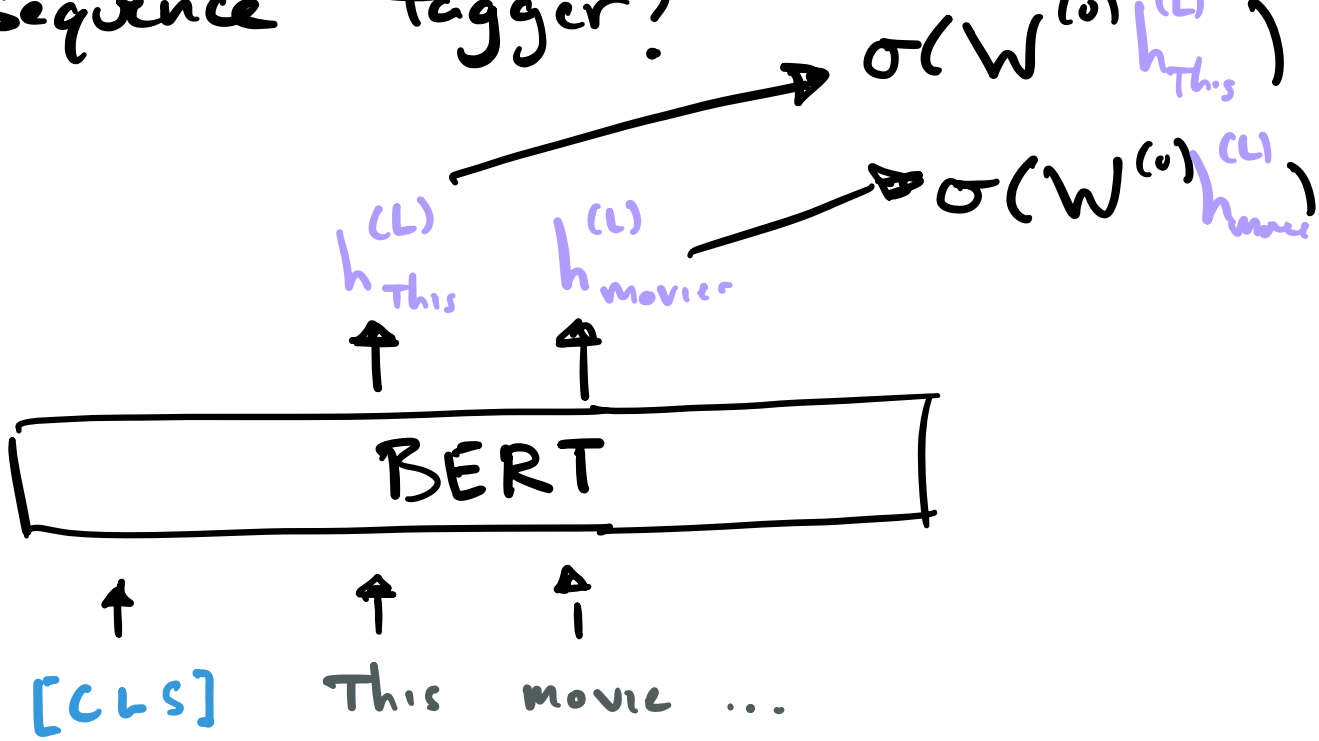
[CLS] prepended to every input

[sep] Separates Segments

How would you train a classifier?



A sequence tagger?



BERT quickly dominated in NLP, often outperforming old methods by wide margins. How?

Enter BERTology or the study of BERT. Sort of the o.g. "mech interp".

How can we study models?

Behaviorally

Create pairs of inputs, observe likelihood Δ's or actual output differences.

The capitol of France is Paris ✓

The capitol of France is Berlin ✗

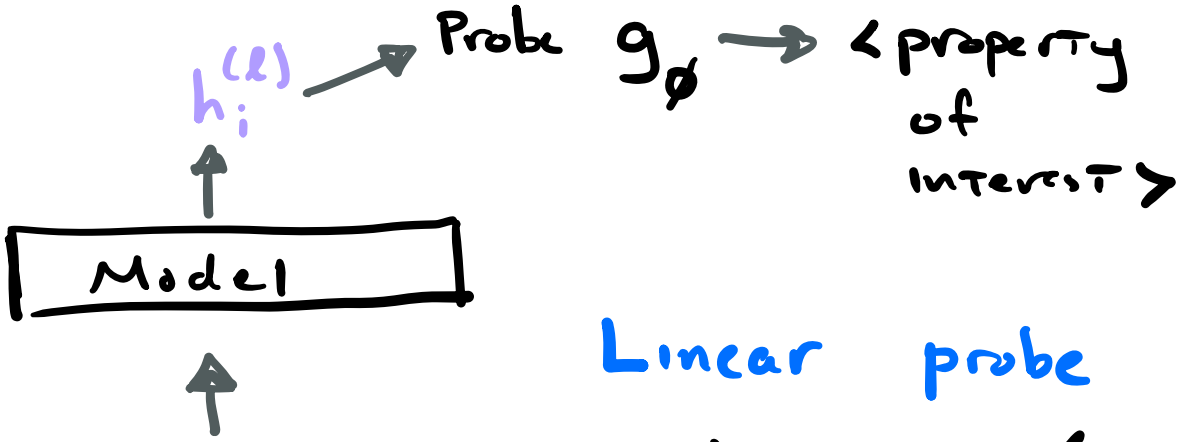
Or for bias: Recent work using MEDQA found

14% change when gender edited!

(But important to have baselines. We found ≈14% flip rate with paraphrases!)

Reoccuring Theme: Interp Methods Need Controls.

Probes

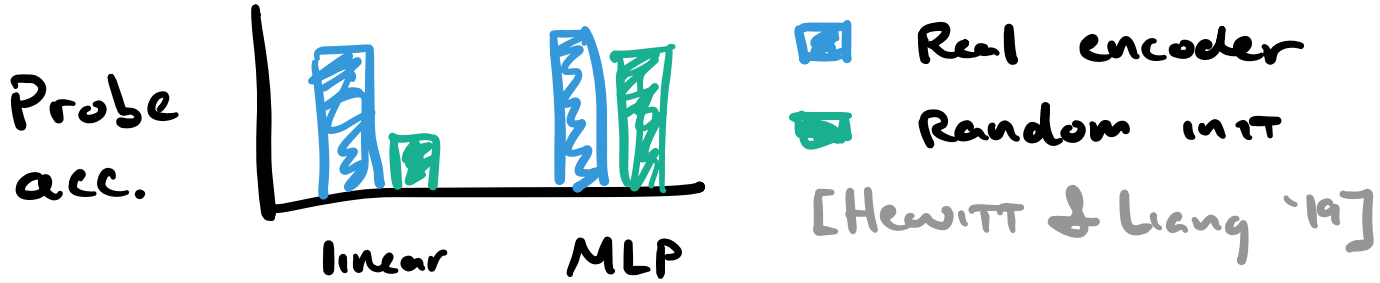


Linear probe

$$g_\phi(h) = \text{SoftMax}(Wh + b)$$

Property is "linearly decodable"

Use a **Control** - especially important for expressive probes like MLPs!

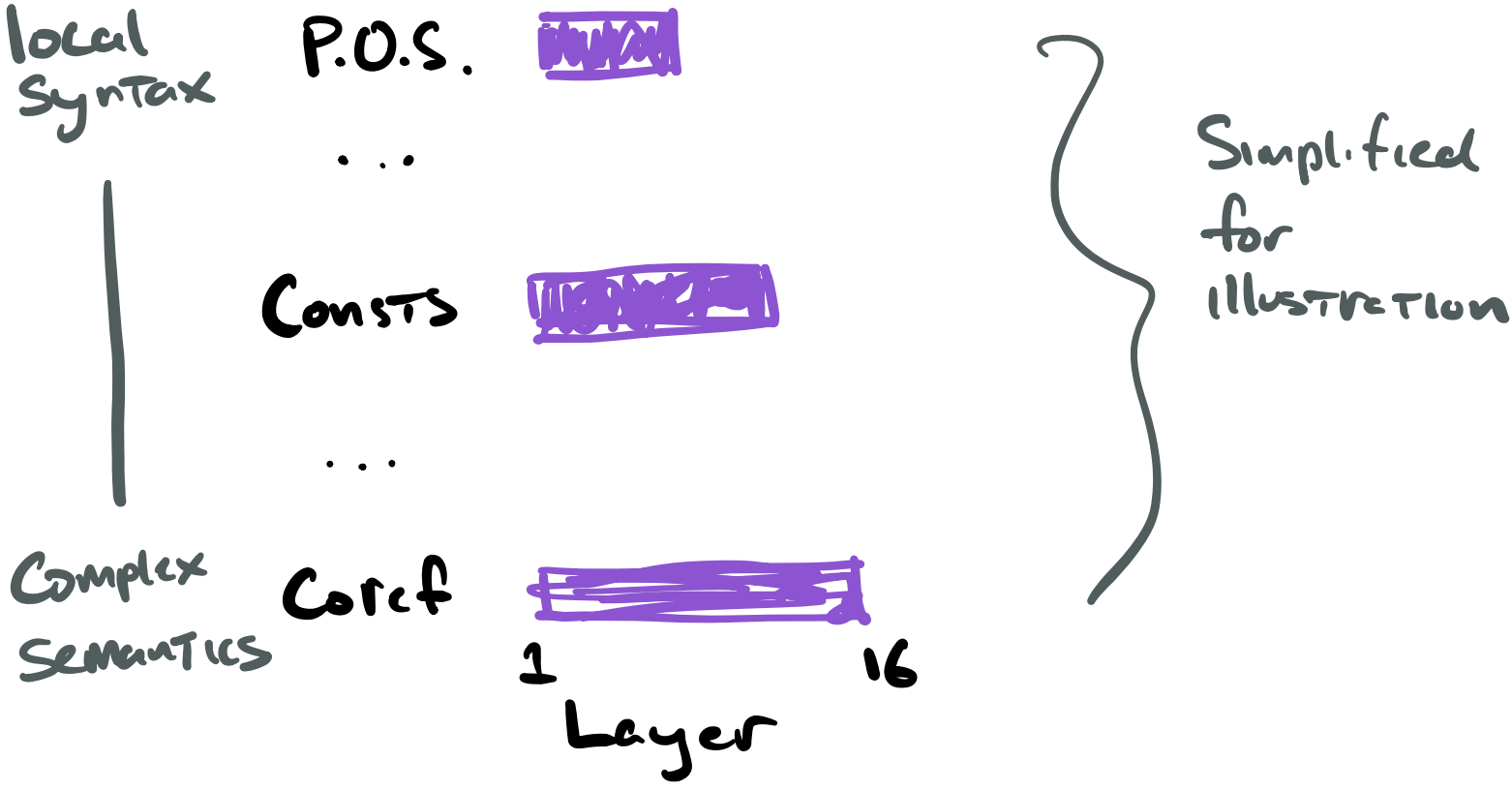


Tenney et al. ('19) used probes to argue that BERT "rediscovers" the classic NLP "pipeline".

$$h_{i,\,Task} = \sum_{l=1}^{L} s_{task}^{(l)} h_i^{(l)}$$

$$s_{task}^{(l)} = Soft\text{-}Max(a_{Task})$$

Learn layer mixture weights per task



Caveat: Probing tells us we can recover information, but not that

it was used by the model.

Attention

People often plot attention weights, but these should be interpreted with some caution. Consider:

Token 1	$\alpha_1 = .9$	$v_1 = [0\ 0]$
Token 2	$\alpha_2 = .05$	$v_2 = [2\ 1]$
Token 3	$\alpha_3 = .05$	$v_3 = [-1\ 3]$

$\bar{h} = [.045\ \ .181]$

Token 1 contributed nothing, but has high α.

More generally, token representations mix info from the context; this makes interp hard.

Ablation

We are often interested in which components are most responsible for an output.

We might e.g. focus on attention heads. We can asses this directly by dropping a head $\text{ATT}_{l,a}$.

But there are $L \cdot H$ heads. For BERT that's 144 — A lot of forward passes. Approximate:

$$\left| \underbrace{\text{ATT}_{l,a}(x)}_{\text{local}} \cdot \underbrace{\frac{\partial \mathcal{L}}{\partial \text{ATT}_{l,a}(x)}}_{\text{incoming}} \right|$$

Can use this to rank heads.

—

BERTology was an early foray into understanding pretrained Transformers.

We will return to these methods (and beyond) later in the semester.