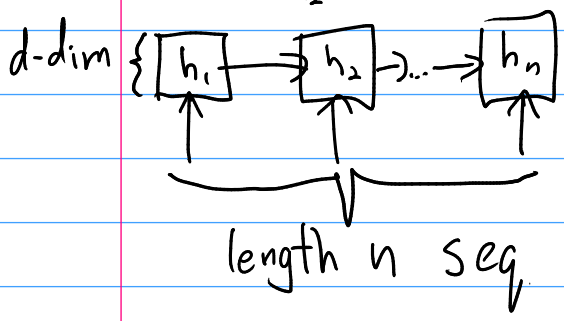
Attention

Before: RNNs for sequence model

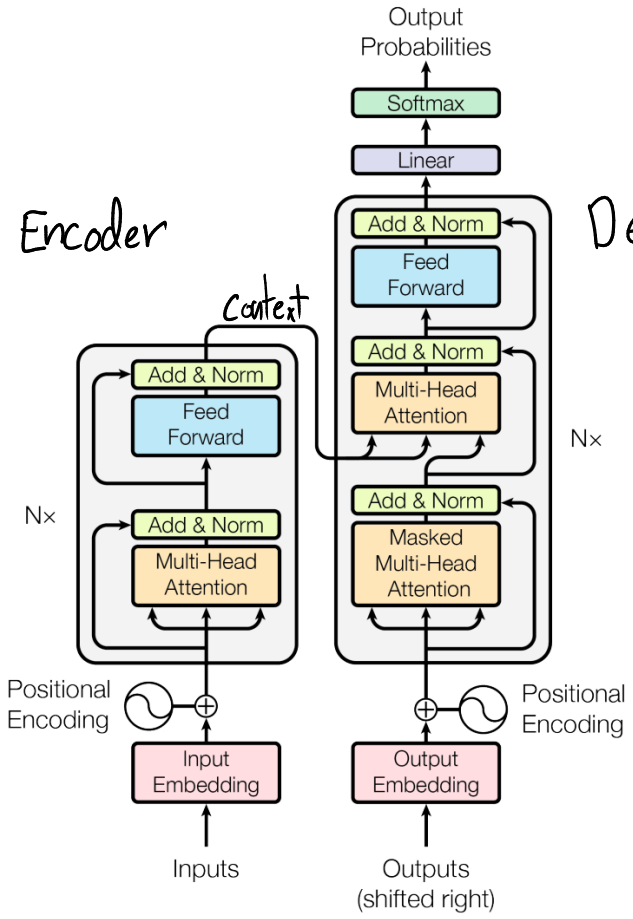
$h_2 = \sigma(Wh_1 + Zx_2)$



Computation: $O(d^2)$ per update
Total: $O(nd^2)$ computation

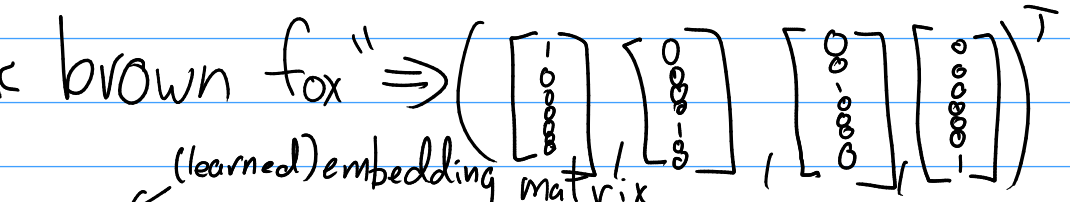
Not easily parallelizable
Longest path: $O(n)$ length
Some challenges in optimization

Attention mechanism
Transformer Architecture (Vaswani et al '17)



Input/Output Encoding

$X = (X_1, \ldots, X_n)^T \in \mathbb{R}^{n \times p}$, $X_i \in \mathbb{R}^p$, one hot

"The quick brown fox" $\Rightarrow$ 

Embedding: $XW_e \leftarrow \mathbb{R}^{n \times d}$ (learned embedding matrix)

$\mathbb{R}^{n \times p}$ $\mathbb{R}^{p \times d}$ $d \ll p$

512 10k+

Positional encoding

The dog is behind the cat.
The cat is behind the dog.

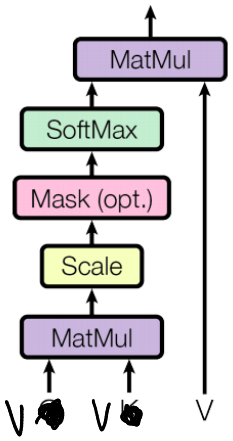
Construct $V \in \mathbb{R}^{n \times d}$:

a) Use fixed $V_{t,2i} = \sin(t/10000^{2i/d})$ $i = 0$ to $\frac{d}{2} - 1$
$V_{t,2i+1} = \cos(t/10000^{2i/d})$

b) Learn a V

Add to embedding: $XW_e + V \in \mathbb{R}^{n \times d}$

~~Scaled Dot-Product~~ Attention
Self



Self-Attention

Given $V \in \mathbb{R}^{n \times d}$, compute

$$V \leftarrow \text{softmax}(VV^T/\lambda)V$$

(λ: scaling factor $= \sqrt{d}$; $\lambda = 1$ for discussion)

First: VV^T

$$\begin{bmatrix} \text{word}_1 \\ \text{word}_2 \\ \vdots \\ \text{word}_n \end{bmatrix} \cdot \begin{bmatrix} \text{word}_1 & \cdots & \text{word}_n \end{bmatrix} \in \mathbb{R}^{n \times n}$$

entry (i,j): dot product of word i and j vecs

If $i \neq j$ "similar", large val

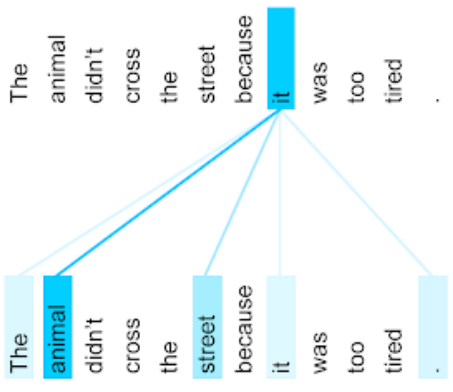
different, small val

Softmax: make each row a distribution

$$\underbrace{\text{Softmax}(VV^T)}_{\mathbb{R}^{n \times n}} \times \underbrace{V}_{\mathbb{R}^{n \times d}} = \mathbb{R}^{n \times d}$$

Replace each row of V by weight sum of rows

$$\begin{bmatrix} \frac{2}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & \frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{4} & \frac{3}{4} & \frac{3}{4} \end{bmatrix}$$



Computation: $O(n^2 d)$ $VV^T \in O(n^2 d)$

$\text{Softmax}(VV^T)V \in O(n^3 d)$

vs RNNs $O(nd^2)$

Shorter max path $O(1)$

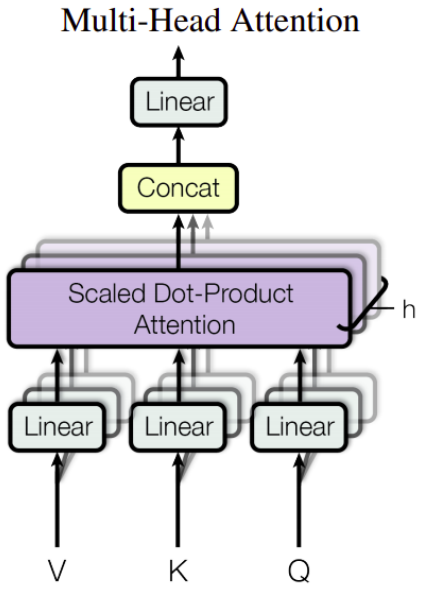
Multi-headed attention

For $i = 1$ to h: $\mathbb{R}^{d \times \frac{d}{h}}$

$V_i = V W^{V_i}$

$V_i = \text{softmax}(V_i V_i^T \lambda) V_i$

$V \in \text{Concat}(V_1, \ldots, V_h) W$ $\mathbb{R}^{d \times d}$

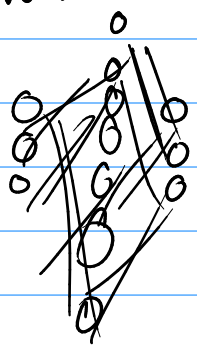


Other details

- Residuals, layer norm
- FF network

$FFN(v) = W_2 \text{ReLU}(W_1 v)$

Shared across position



Apply to each row of matrix

Decoder

Masked self attention

Given $Q \in \mathbb{R}^{n \times d}$, compute

$$Q \leftarrow \text{Softmax}(\text{mask}(QQ^T/\lambda))Q$$

$$\text{mask}(M)_{ij} = \begin{cases} -\infty \text{ if } i < j \\ M_{ij} \text{ else} \end{cases}$$

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad \text{mask}(M) = \begin{bmatrix} 1 & -\infty & -\infty \\ 4 & 5 & -\infty \\ 7 & 8 & 9 \end{bmatrix}$$

Attention w/ context

Given context V, matrix Q, compute

$$Q \leftarrow \text{Softmax}(Q^T V/\lambda)V$$

Final Softmax

Convert vectors from $\mathbb{R}^d$ back to $\mathbb{R}^p$

gives pred seq $\hat{y}$

Optimize:

$$\min_{\text{(all params in model)}} \sum_{\text{all example}} \sum_{j=1}^{l} -y_j \log \hat{y}_j$$

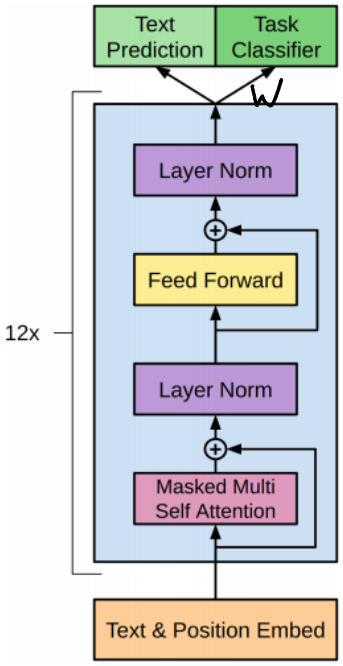
$\hat{y}_j$ ← logit corresponding to true output

y_j ↑ true jth output

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	**41.29**	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1	**$3.3 \cdot 10^{18}$**	
Transformer (big)	**28.4**	**41.8**	$2.3 \cdot 10^{19}$	

GPT-1

Two stages of training.

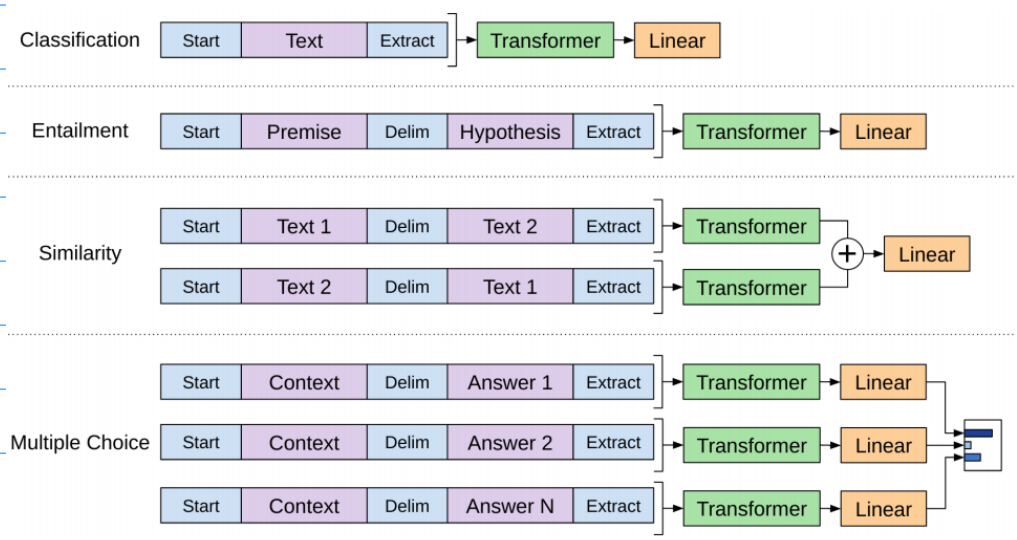


1. Unsupervised Pre-training
   - Lots of text used
   - Objective: next word prediction

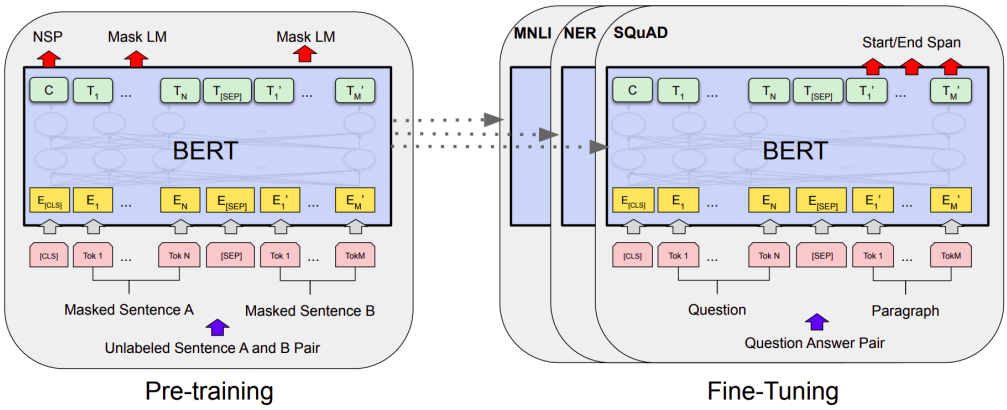
$$\min_{\theta} \sum_{\substack{\text{all} \\ \text{examples}}} \sum_{i=1}^{n} -\log P_\theta(x_i | x_1, \ldots, x_{i-1})$$

2. Supervised Fine-tuning

$$\min_{\theta, W} \sum_{\substack{\text{all} \\ \text{examples}}} \left(-\log P_{\theta, W}(y|x) + \lambda \sum_{i=1}^{n} -\log P_\theta(x_i | x_1, \ldots, x_{i-1}) \right)$$



BERT



How to train:

a) Input: I took my [mask] for a walk.

⇒ I took my dog for a walk.

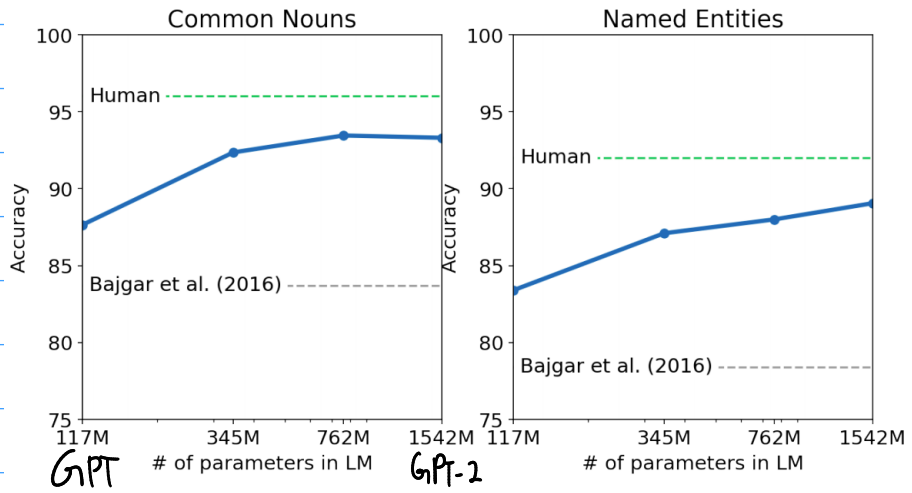
b) Feed in sentences A and B

- Either B follows A, or is random
- Learn to predict which is the case

Sum two losses for pretraining

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

GPT-2 10x bigger than GPT



GPT-3

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
cheese =>                           ← prompt
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← example
cheese =>                           ← prompt
```

Few-shot

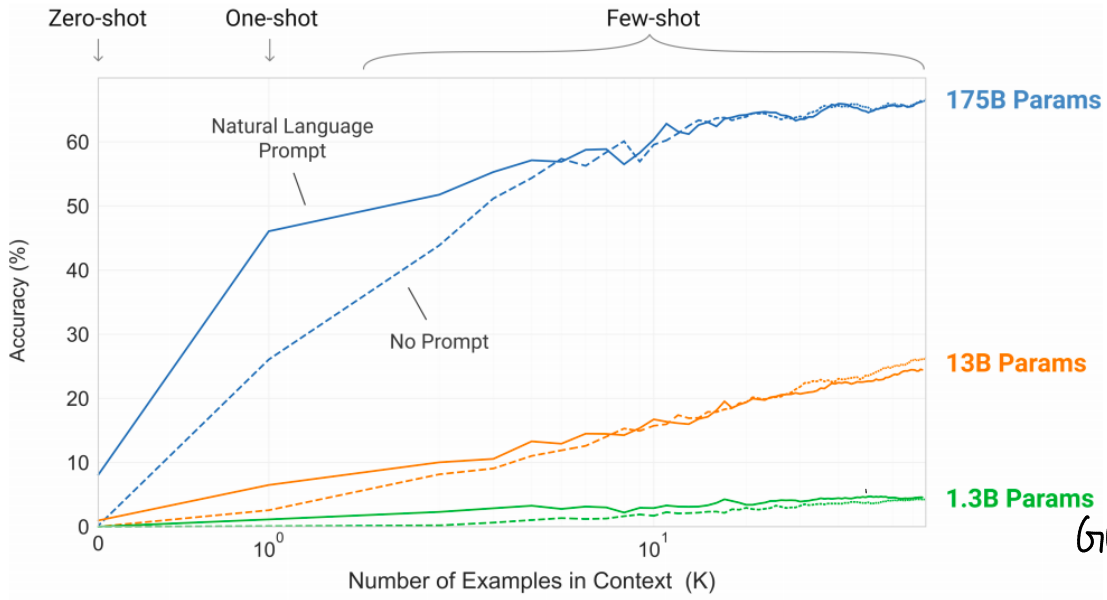
In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
Translate English to French:        ← task description
sea otter => loutre de mer          ← examples
peppermint => menthe poivrée
plush girafe => girafe peluche
cheese =>                           ← prompt
```

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

```
sea otter => loutre de mer          ← example #1
        ↓
   gradient update
        ↓
peppermint => menthe poivrée        ← example #2
        ↓
   gradient update
        ↓
        ...
        ↓
plush giraffe => girafe peluche     ← example #N
        ↓
   gradient update
        ↓
cheese =>                           ← prompt
```



GPT-3

GPT2

Context →	The City BY C. P. CAVAFY TRANSLATED BY EDMUND KEELEY [Poem text omitted] SOME TREES John Ashbery [Poem text omitted] Shadows on the Way Wallace Stevens

-------- Generated Poem 1 --------

I must have shadows on the way
If I am to walk I must have
Each step taken slowly and alone
To have it ready made

And I must think in lines of grey
To have dim thoughts to be my guide
Must look on blue and green
And never let my eye forget
That color is my friend
And purple must surround me too

The yellow of the sun is no more
Intrusive than the bluish snow
That falls on all of us. I must have
Grey thoughts and blue thoughts walk with me
If I am to go away at all.

-------- Generated Poem 2 --------

There is a way in the world.
That way a man may live in a small town
His lifetime and see no man come,
Never the stir of men about a market,
Never the bobbing of covered carts,
Never the creak passed on the stair
Closing some other life; an old man
Can prosper in that town, can prosper
When his store is seen to close, his wife
Fades in his house, and the world shouts in vain
That no one comes to hear him
Where his small bell tolls.
In that small town
Everything is known.
The ways of the young men are known,
And the ways of the young women
As they walk to the Spring;
The rumors are all known,
The quarrels, the hurt, the envy
Of the men and the women.
Till a day comes
That way out of the world
That leads to a wood.
Young, the young man
Walks, a lover now,
Knows the streets as they are and is of them.

Comes to the wood.

There, in the wood, among trees,
He sees shadows on the way, hears voices,
hears the wind and the rustling of leaves;
Through an open glade
He sees a shape and the shape hears:
It waits as he waits,
Waits as the shadows wait,
As the voices wait;
Shadows on the way, voices in the wind.

-------- Generated Poem 3 --------

The sun was all we had. Now, in the shade
All is changed. The mind must dwell on those
White fields, that to its eyes were always old;
Those ancient gleams, convoluting
The way they lay among the huge roots,
The great dark tomes of reverie,
The plumed labyrinth of the sea.

When our own breath gonne faint and fade
Up the long stairway of the dead
We shall remember well
The blindness of the light, the whiteness
Of that white land. The footsteps, and
The grasses in that light, and in the shade
The menacing life.

We played, too, once, in the warmest rooms.
To one content, there is one content
For all mankind. The forest glades
Are all the more for being in fable,
And wood and town in legend shadowed-
Once more to return. The one content
Has in it every happiness, more brave
For being purposeless; and so
Those island gardens, deep in the mind,
Make of remoteness familiarity
And of the unspoken name a friend.

-------- Generated Poem 4 --------

Nobody will come to this place. It is a road that leads nowhere.
The solitude is deep. The mountains are high.
But they are desolate, and they turn the traveler's face
Towards the North. All the sounds of the world are far away.
When the wind rises above the trees,
The boughs bow to the ground.
Even the birds that inhabit the tangle of weeds
That is the roadside cover, are silent. One listens,
But hears no roar of the forest. One is alone.
One will be taken.
One will be taken.
There is no utterance, there is no conversation,
But one is uneasy all the same....
There is a thin blue mist,
A darkness rising like smoke,
And within that darkness
A possession of the heart.
One will be taken.... It was here, and it will be here again-
Here, under this sky empty and full of light.