Lecture 8/8: Self Attention, Transformer NN

- demo for HW6 pb 3,4
- PROJECTS due Sunday 8/10
- HW6 due Fri 8/15 (incl. demo)
- last class Wed 8/13 (word-to-vec NLP)

Transformer

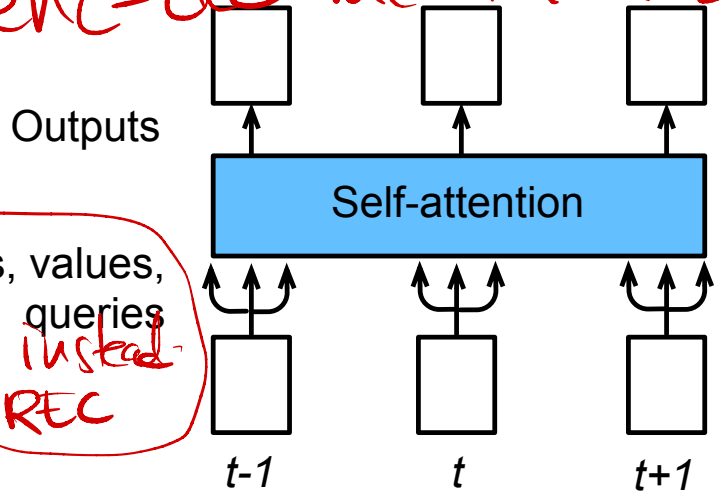
movie concept $\overset{?}{\Rightarrow}$ transformer NN

- NO CLASS FRI 8/15
  use time for PROJ, HW6
- TRACE evals
- grades due Mon 8/18

courses.d2l.ai/berkeley-stat-157/index.html

Self-attention

- To generate *n* outputs with *n* inputs, we can copy each input into a key, a value and a query
- No sequential information is preserved
- Run in parallel



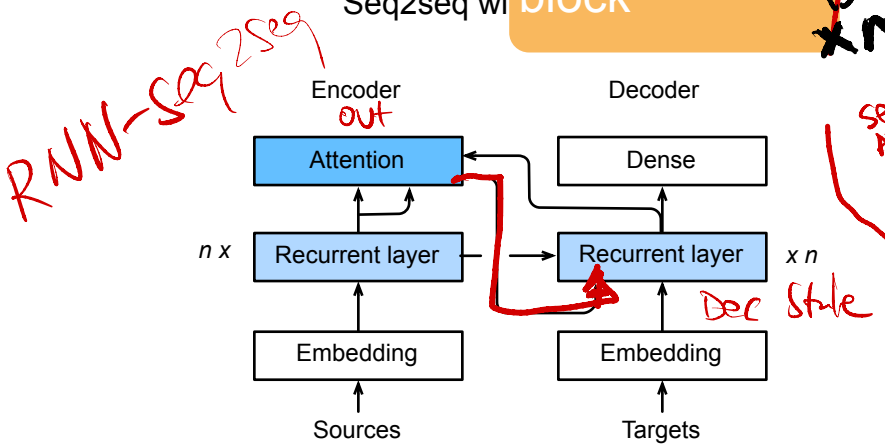
Transformer Architecture

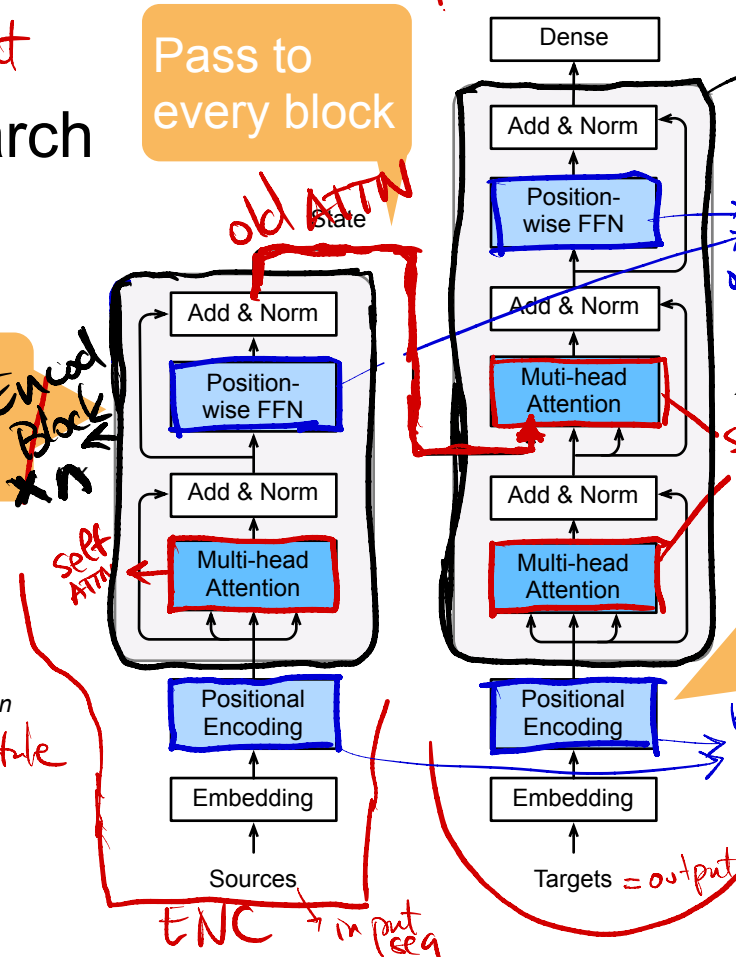
- It's an encoder-decoder arch
- Differ to seq2seq with attention in 3 places

Transformer block

Pass to every block

Sequential information

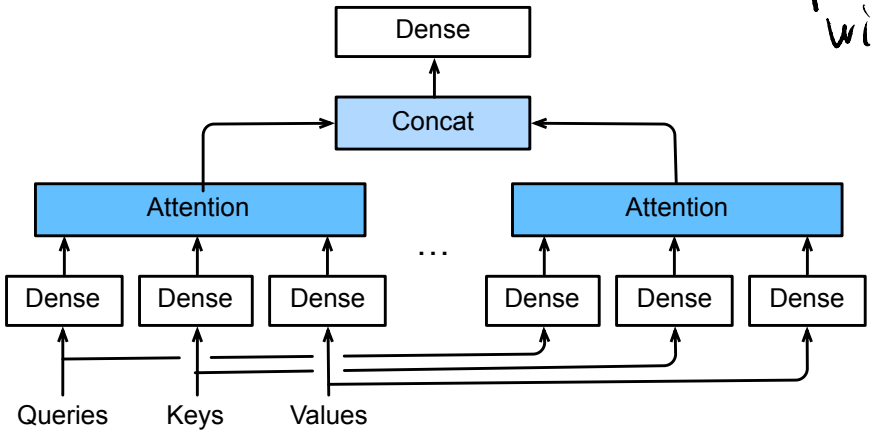


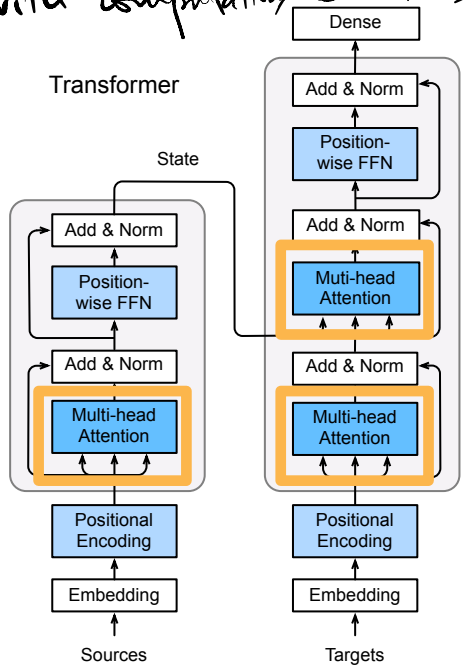


courses.d2l.ai/berkeley-stat-157/index.html

Multi-head Attention

Add & Norm → layer normalization (after comp block)
- normalizes output at that layer
- prevent problems with computation activations, with gradients





$$\mathbf{W}_q^{(i)} \in \mathbb{R}^{p_q \times d_q}, \mathbf{W}_k^{(i)} \in \mathbb{R}^{p_k \times d_k}, \text{ and } \mathbf{W}_v^{(i)} \in \mathbb{R}^{p_v \times d_v}$$

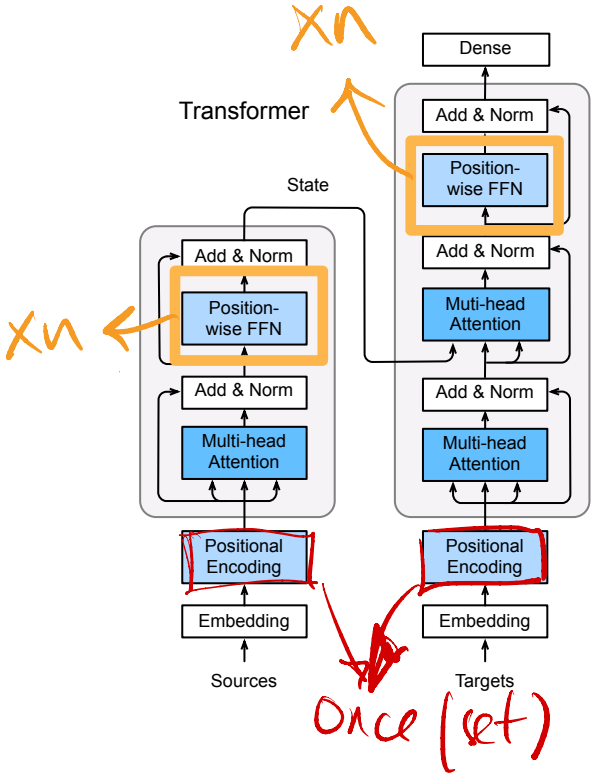
$$\mathbf{o}^{(i)} = \text{attention}(\mathbf{W}_q^{(i)}\mathbf{q}, \mathbf{W}_k^{(i)}\mathbf{k}, \mathbf{W}_v^{(i)}\mathbf{v})$$

for $i = 1, \ldots, h$

$$\mathbf{o} = \mathbf{W}_o \begin{bmatrix} \mathbf{o}^{(1)} \\ \vdots \\ \mathbf{o}^{(h)} \end{bmatrix}$$

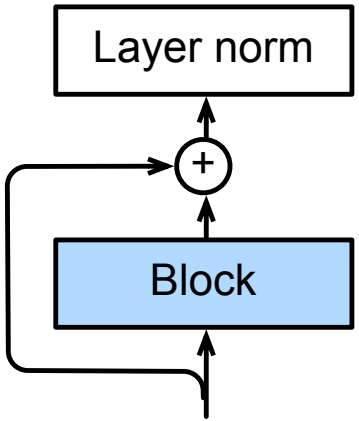
Position-wise Feed-Forward Networks

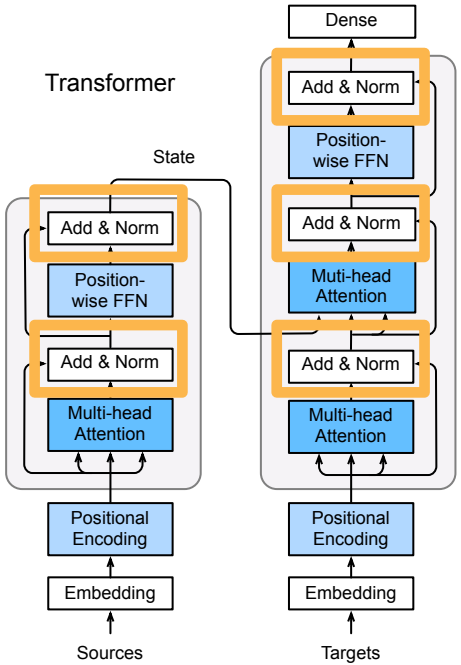
- Reshape input (batch, seq len, fea size) into (batch * seq len, fea size)
- Apply a two layer MLP
- Reshape back into 3-D
- Equals to apply two (1,1) conv layers



Add and Norm

- Layer norm is similar to batch norm
- But the mean and variances are calculated along the last dimension
- `X.mean(axis=-1)` instead of the first batch dimension in batch norm `X.mean(axis=0)`



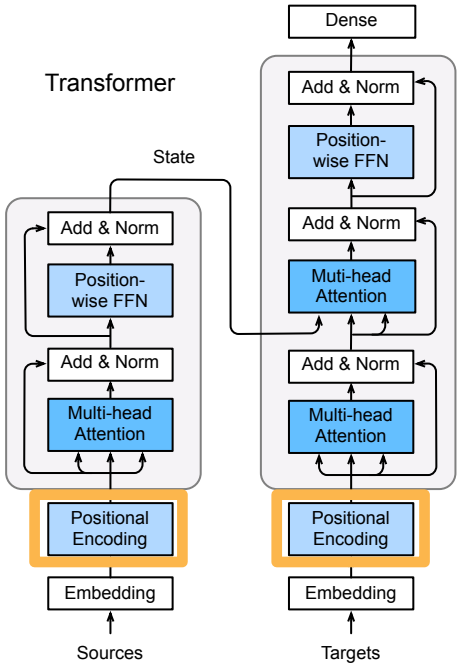


Positional Encoding

- Assume embedding output $X \in \mathbb{R}^{l \times d}$ with shape (seq len, embed dim)
- Create $P \in \mathbb{R}^{l \times d}$ with

$$P_{i,2j} = \sin(i/10000^{2j/d})$$
$$P_{i,2j+1} = \cos(i/10000^{2jd})$$

- Output $X + P$

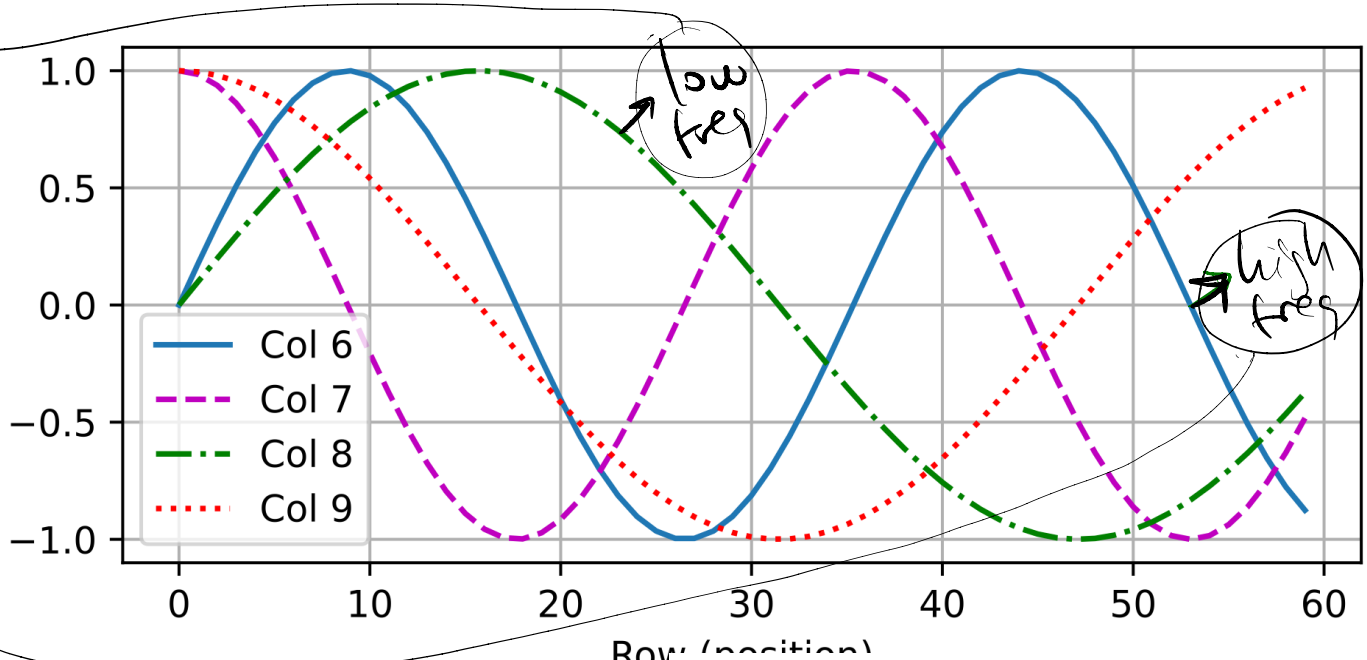


0	0	0	0	
1	0	0	1	
2	0	1	0	
3	0	1	1	
4	1	0	0	
5	1	0	1	
6	1	1	0	← last col
7	1	1	1	

discrete ⟹ to continuous

pattern by col

4(0) 2(0) 0
4(1) 2(1) 1
4(2) 2(2) 0
4(3) 2(3) 1



trigonometry

Relative positioning (already done) by δ

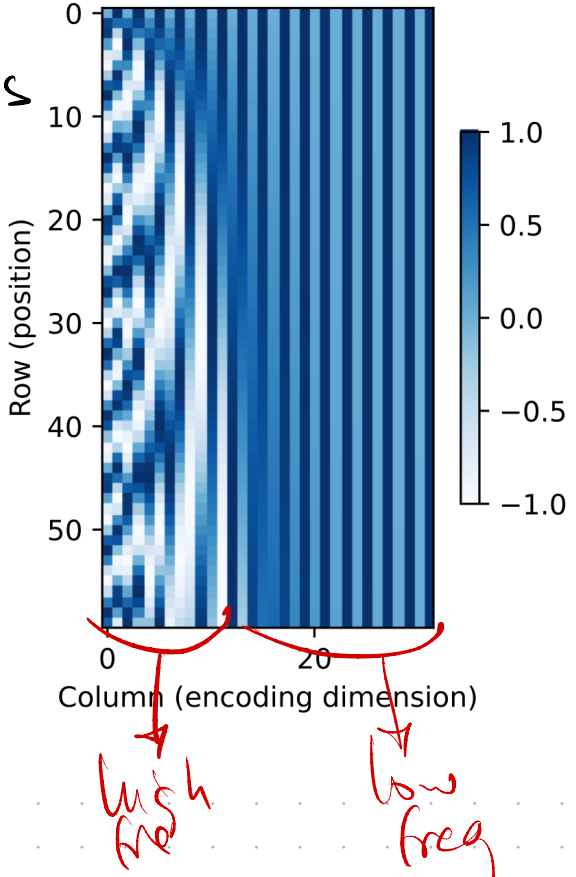
steps

$$\omega_j = 1/10000^{2j/d}$$

$$\begin{bmatrix} P_{i+\delta,2j} \\ P_{i+\delta,2j+1} \end{bmatrix} = \begin{bmatrix} \cos(\delta\omega_j) & \sin(\delta\omega_j) \\ -\sin(\delta\omega_j) & \cos(\delta\omega_j) \end{bmatrix} \begin{bmatrix} P_{i,2j} \\ P_{i,2j+1} \end{bmatrix}$$

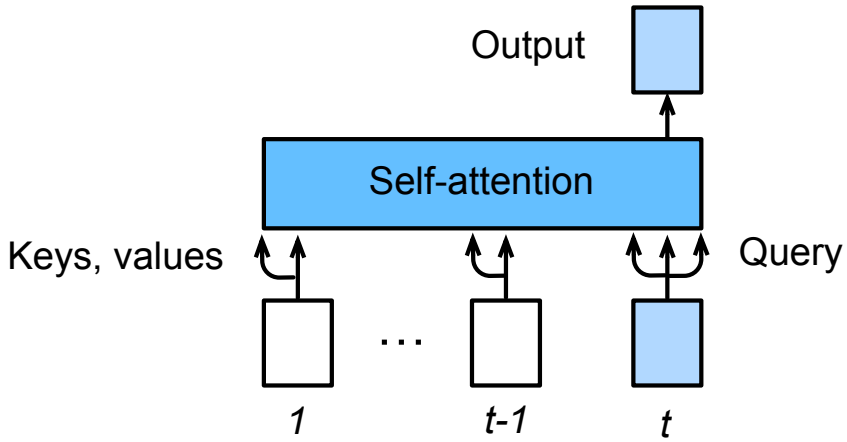
pos at $i+\delta$

pos at i



Predicting

- Predict at time *t*:
  - Inputs of previous times as keys and values
  - Input at time *t* as query, as well as key and value, to predict output



Code…

courses.d2l.ai/berkeley-stat-157/index.html

BERT

courses.d2l.ai/berkeley-stat-157/index.html

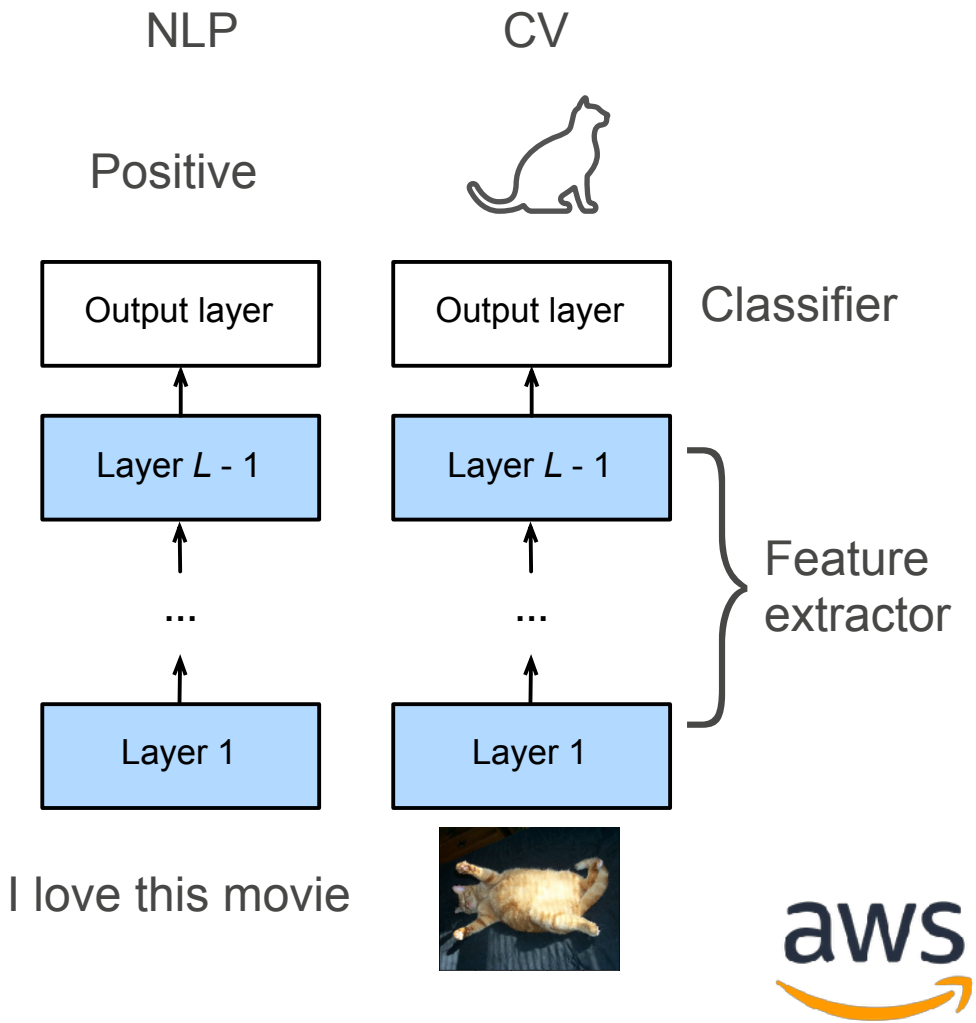
aws

Transfer Learning in NLP

- Use pre-trained models to extract word/sentence features for the new task
  - E.g. word2vec or language model
- Often don’t update the pre-trained models
- Need to construct a new model to capture the information needed for the new task
  - Word2vec ignores sequential information, language model only looks in a single direction

Motivation of BERT

- A fine-tuning based approach for NLP
- The pre-trained model capture sufficient data information
- Only need to add a simple output layer for a new task

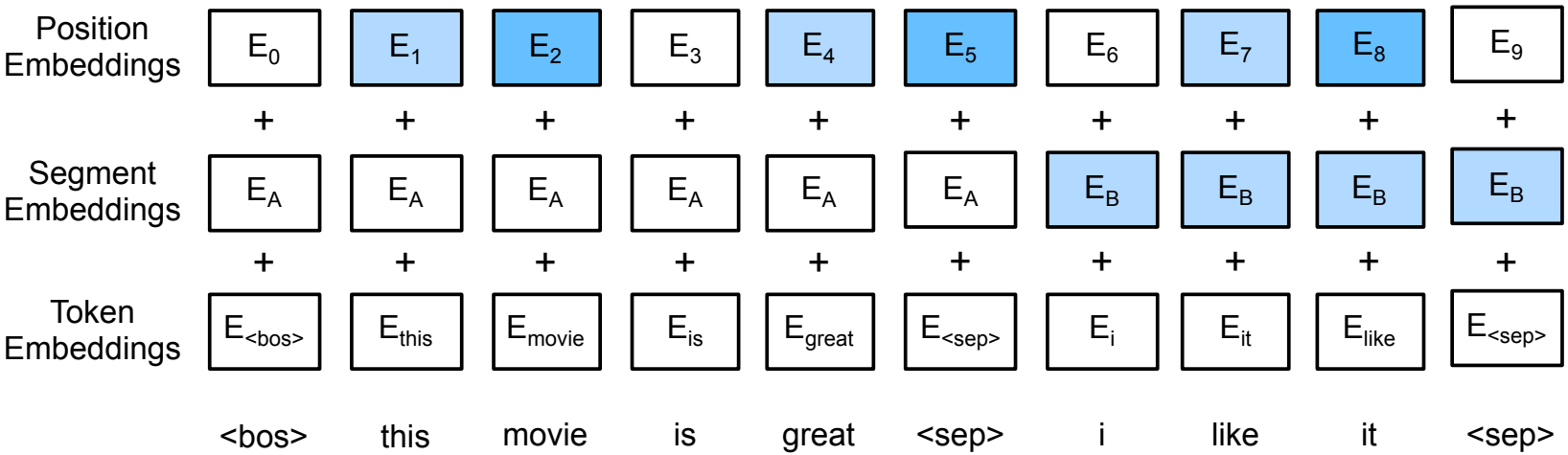


BERT Architecture

- A (big) Transformer encoder (without the decoder)
- Two variants:
  - Base: #blocks = 12, hidden size = 768, #heads = 12, #parameters = 110M
  - Large: #blocks = 24, hidden size = 1024, #heads = 16, #parameter = 340M
- Train on large-scale corpus (books and wikipedia) with > 3B words

Modification of inputs

- Each example is a pair of sentences
- Add an additional segment embedding



Pre-training Task 1: Masked Language Model

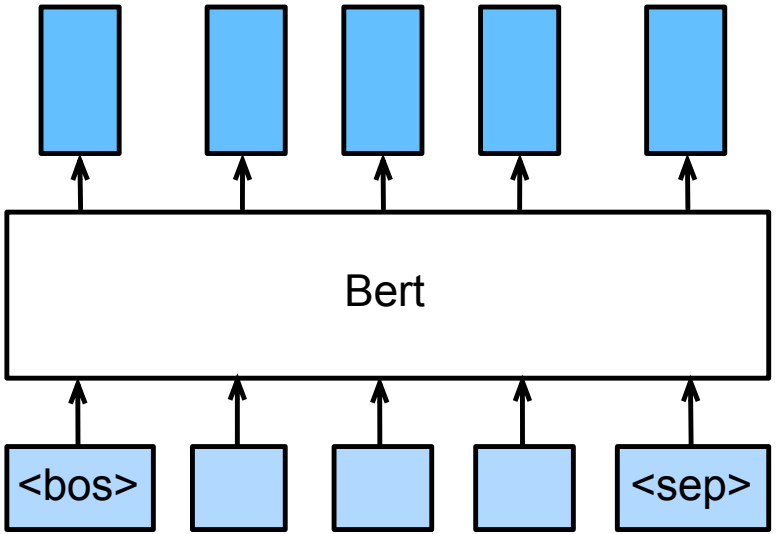
- Randomly mask (e.g. 15%) tokens in each sentence, predict these masked tokens
  - Transformer is bidirectional, which breaks the unidirectional limit of standard LM
- No mask token (<mask>) in fine turning tasks
  - 80% of the time, replace selected tokens with <mask>
  - 10% of the time, replace with randomly picked tokens
  - 10% of the time, keep the original tokens

Pre-training Task 2: Next Sentence Prediction

- 50% of time, choose a sequential sentence pair
  - <bos> this movie is great <sep> i like it <sep>
- 50% of time, choose a random sentence pair
  - <bos> this movie is great <sep> hello world <sep>
- Feed the Transformer output of <bos> into a dense layer to predict if it is a sequential pair

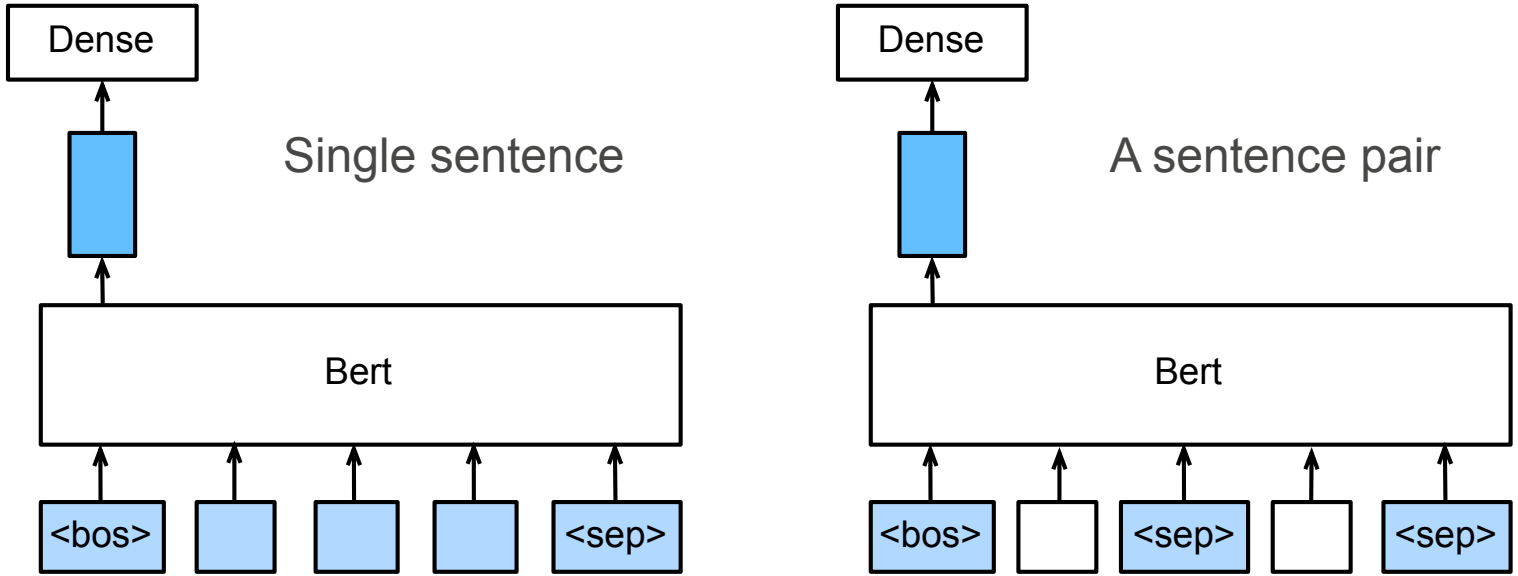
Bert for Fine Tuning

- Bert returns a feature vector for each token that captures the context information
- Different fine-tuning tasks use a different set of vectors



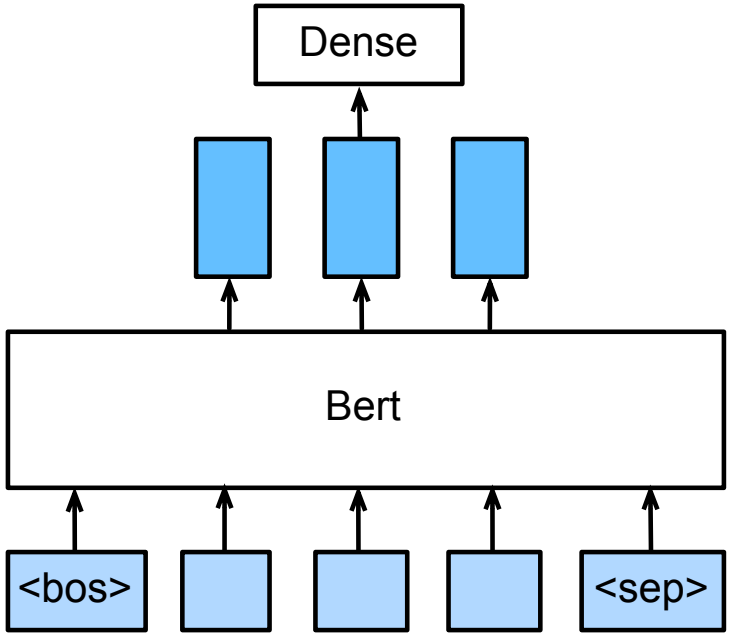
Sentences Classification

- Feed the <bos> token vector into a dense output layer



Named Entity Recognition

- Identify if a token is a named entity such as person, org, and locations…
- Feed each non-special token vector into a dense output layer

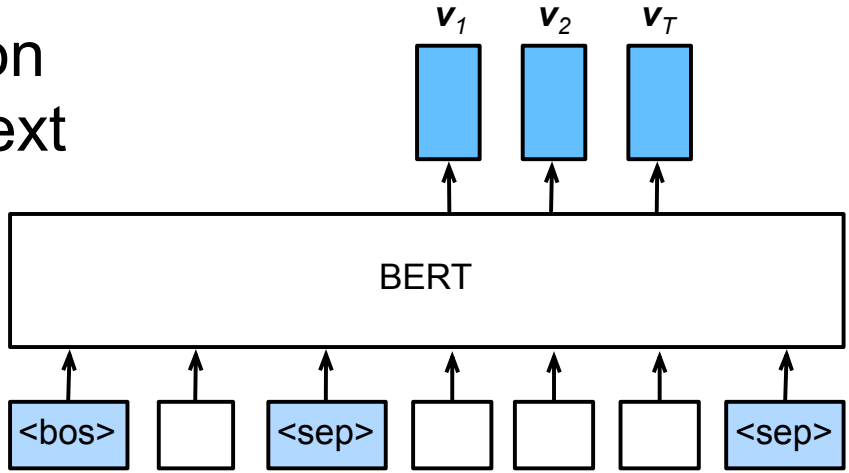


Question Answering

- Given a question and a description text, find the answer, which is a text segment in the description
- Given $\mathbf{p}_i$ the *i*-th token in the desperation, learn $\mathbf{s}$ so that

$$p_1, \ldots, p_T = \text{softmax}(\langle \mathbf{s}, \mathbf{v}_1 \rangle, \ldots, \langle \mathbf{s}, \mathbf{v}_T \rangle)$$

p_i is the probability *i*-th token is the segment start. Same for the end



Check GluonNLP for code…

courses.d2l.ai/berkeley-stat-157/index.html