Class

Recap

Reminders

word2vec

GloVe

ELMo, BERT, GPT

Recap

$(x_1, \ldots, x_n)$ input sequence

$(y_1, \ldots, y_n)$ output sequence

- associate inputs with outputs

  I love you a lot

  Te amo mucho

- variable length

- long range dependency

$$q_i = W_q x_i$$

$$k_i = W_k x_i \qquad \text{for } i \in [n]$$

$$v_i = W_v x_i \qquad = \{1, 2, \ldots, n\}$$

$$w_{ij} = q_i^T k_j$$

$$\text{Softmax}(w_{i1}, w_{i2}, w_{i3}, \ldots, w_{in})$$

$$\frac{\exp(w_{ij})}{\sum_{j'} \exp(w_{ij'})}$$

$$y_i = \sum_{j=1}^{n} w_{ij} v_j$$

$y_i \in \mathbb{R}^d$

$v_j \in \mathbb{R}^d$

$q_i \in \mathbb{R}^{q\text{-dim}}$ $\quad x_i \in \mathbb{R}^{in\text{-dim}}$

W_q, W_k

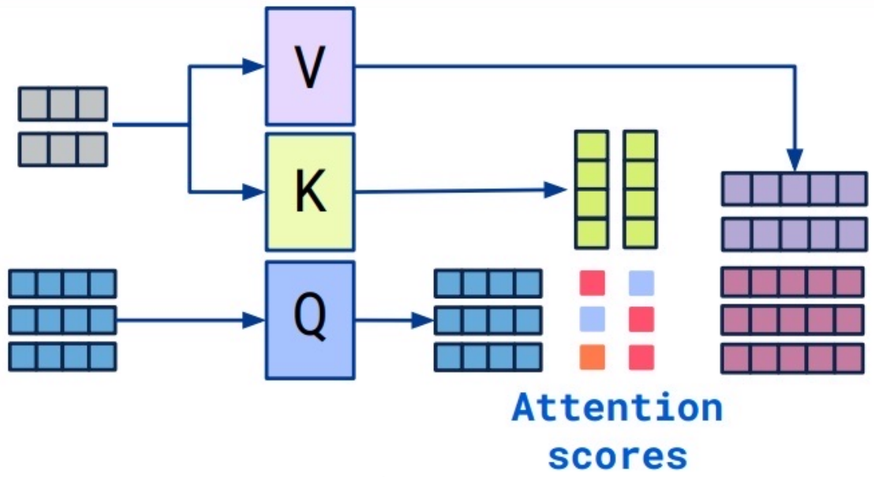
$q\text{-dim} \times in\text{-dim}$

W_v

$d \times in\text{-dim}$

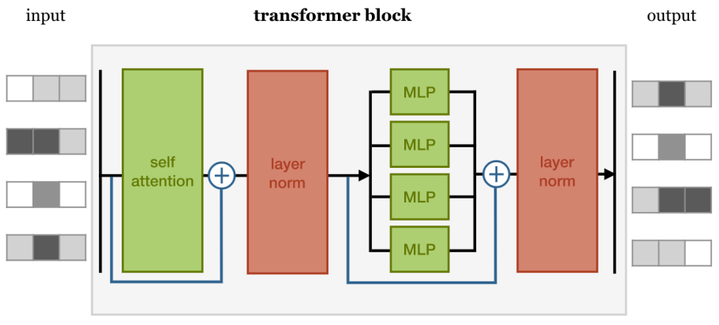
$(x_1, \ldots, x_n)$

$(z_1, \ldots, z_d)$



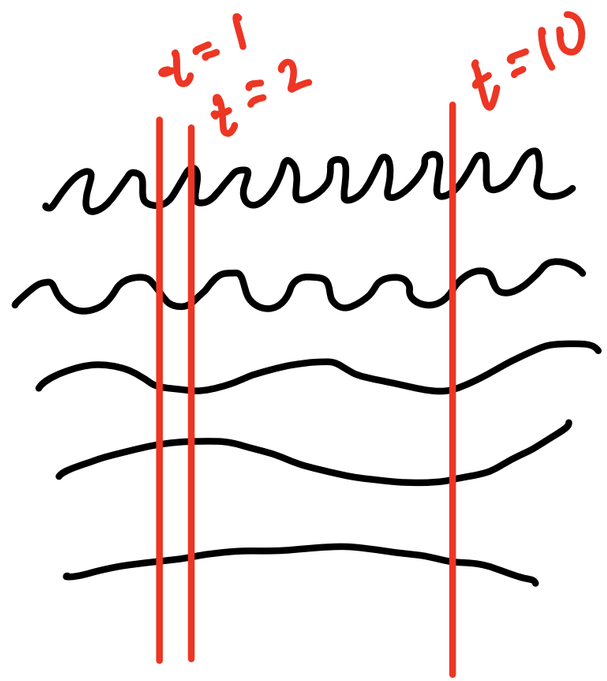
Cross attention

Transformer



$X_i =$ embedding + positional encoding

$$\begin{bmatrix} \sin(f_1 t) \\ \sin(f_2 t) \\ \vdots \end{bmatrix}$$



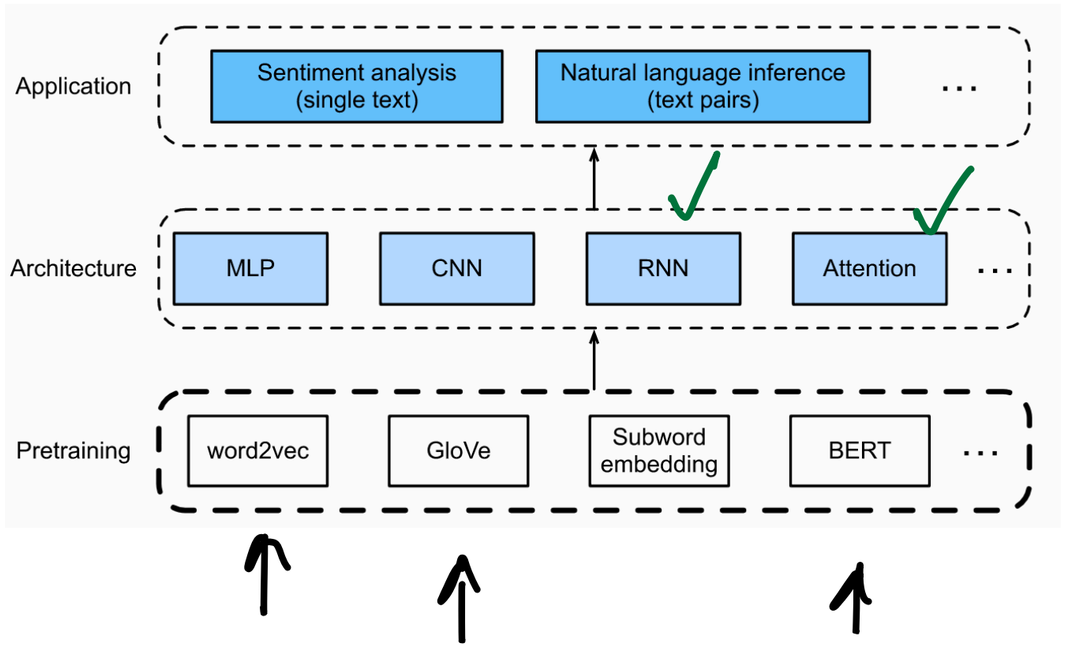
$$\|a\|_2 = \beta \iff \|a\|_2^2 = \beta^2$$

$$\|a\|_2^2 = a^T a = \beta^2$$

Reminders

- Form 23/26
- Homework due 5pm Friday
  - ↳ self grade due Monday
- Proposal due 5pm Monday

Overview

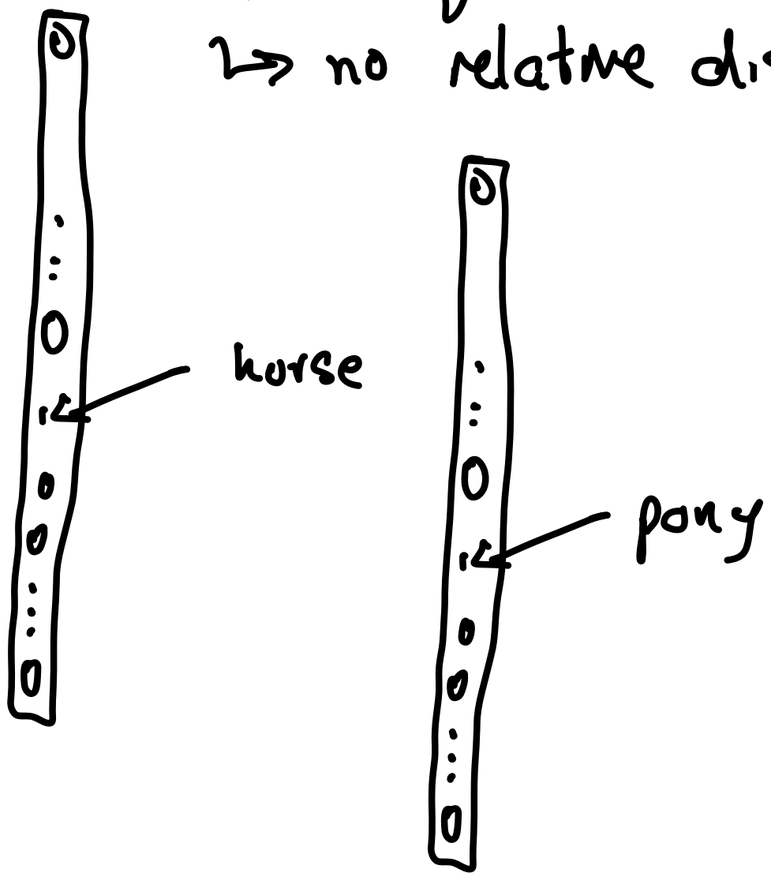


word → vector (meaningful)

One Hot Encoding

↳ many dimensions

↳ no relative distance



word2vec

2-gram

> Gatsby believed in the green light, the orgastic future that year by year recedes before us. It eluded us then, but that's no matter—tomorrow we will run faster, stretch out our arms farther. . . . And one fine morning——
>
> So we beat on, boats against the current, borne back ceaselessly into the past.

x, y

(Gatsby, believed)

(believed, in)

(in, the)

(the, green)

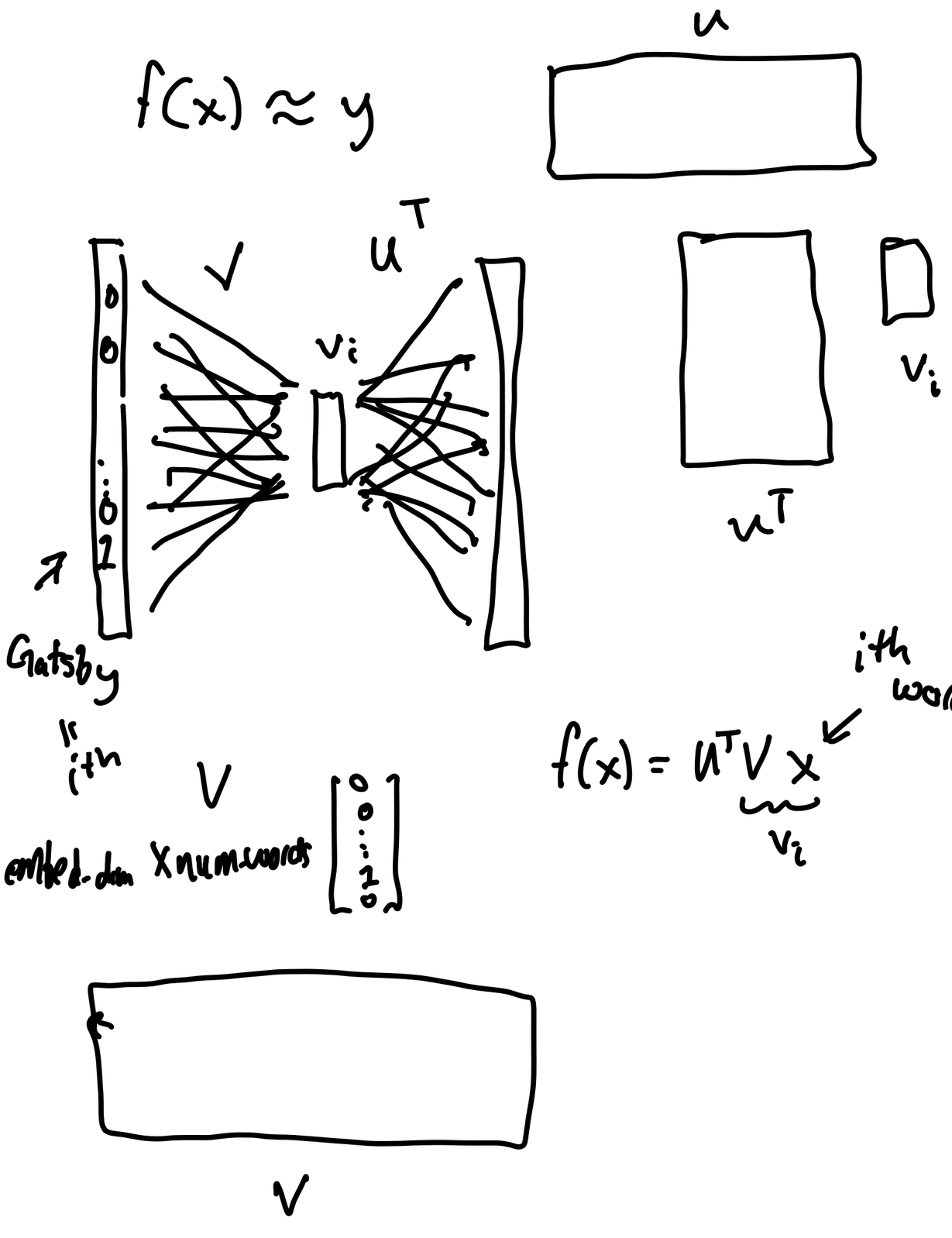
⋮

problems

↳ too sparse

↳ throw away half the model

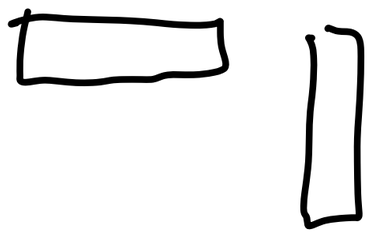
(Gatsby, in)



$f(x)$ ← embedding of x

(x, y) $\quad f(x) \approx f(y)$

$f(x)^T f(y)$ large



if x is close in meaning to y

$f(x) = Vx$

$$\mathcal{L}(V) = -\sum_{x,y} f(x)^T f(y) + \sum_{x,z} \left(f(x)^T f(z)\right)^2$$

↗ positive pair

↗ negative pair

Extension: (x, y)

also include neighbors of neighbors

n-gram vs z-gram

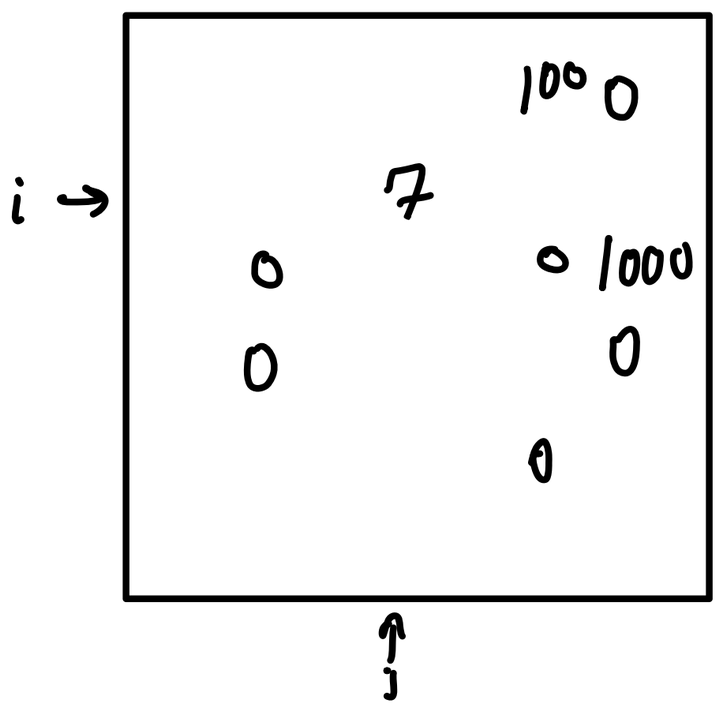
GloVe

Global Vectors

X = co-occurrence matrix

n = # unique words

$X \in \mathbb{R}^{n \times n}$



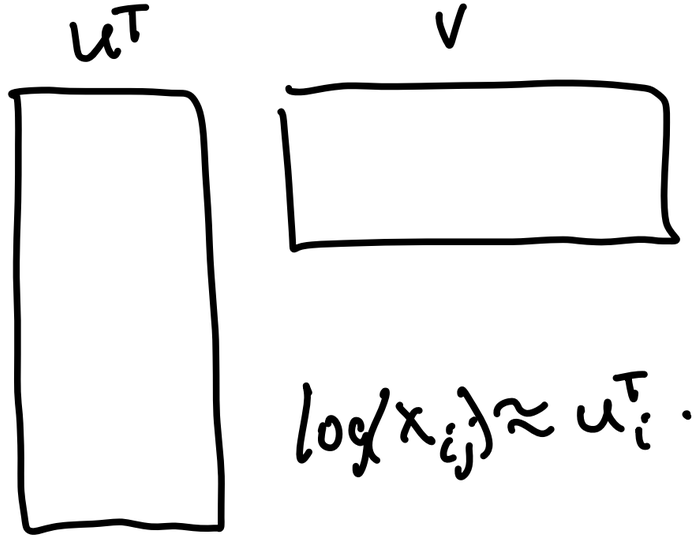
$\log(X) \approx U^T V$

$n \times h \quad h \times n$

trainable parameters

$U, V \in \mathbb{R}^{h \times n}$

$c, b \in \mathbb{R}^n$



$\log(X_{ij}) \approx u_i^T \cdot v_j$

embedding v_i for word i

$= V \times$ ← one hot encoded vector for word i

$$\mathcal{L} = \frac{1}{2} \sum_{i,j} \left(u_i^T v_j + b_i + c_j - \log(X_{ij}) \right)^2 f(X_{ij})$$

$\sum_{i,j}$: really want to only look at non zero

$u_i^T v_j$: embedding

$f(X_{ij})$: close to 1 for large values, close to 0 for small values

Embedding with context

↳ crane

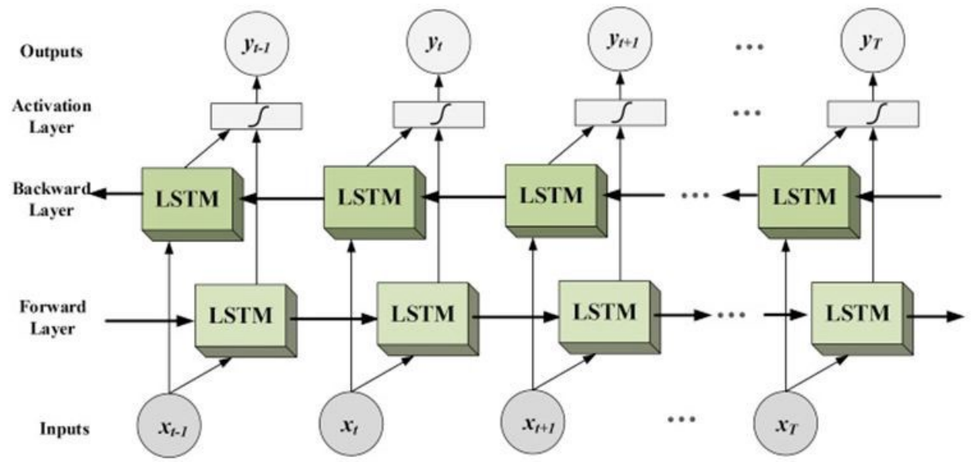
↳ date

↳ dough

ELMo

Embeddings from Language Models

Bidirectional LSTM



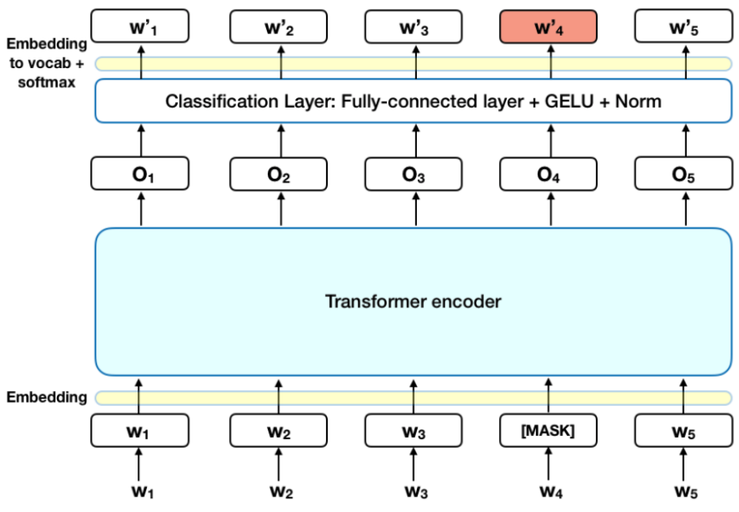
Idea: Replace linear layer in word2vec with fancier architecture

BERT

Bidirectional

Encoder
Representations from
Transformers

Idea: Replace with
transformer



↳ still for next sentence

↳ drop out random words

↳ piece tokenization

walking → walk, ing

GPT

↳ masked tokens

↳ huge data set and
lots of parameters