Language Embeddings

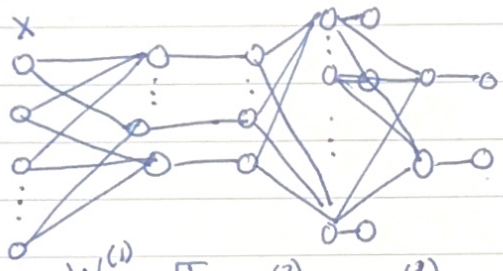
Plan

Review
Autoencoders
Contrastive Learning
PCA

Logistics

Games tonight!
Check in form!
Scribed notes!
Zoom!

Review



1. Model $f: \mathbb{R}^q \to \mathbb{R}^K$
2. Loss $\mathcal{L}(w)$
3. Optimizer

$$w^{(t+1)} = w^{(t)} - \alpha \nabla \mathcal{L}(w^{(t)})$$

momentum $v^{(t+1)} = (1-\beta)\, v^{(t)} + \beta \nabla_w \mathcal{L}(w^{(t)})$

adaptivity $s^{(t+1)} = (1-\beta)\, g^{(t)} + \beta \left[\nabla_w \mathcal{L}(w^{(t)})\right]^2$

Backprop:

Forward:

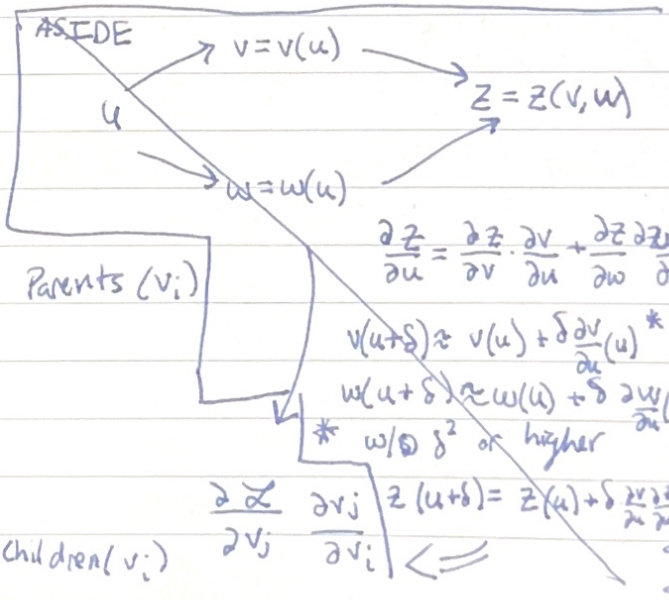
for i in $\{1, \dots, N\}$:

compute v_i from Parents(v_i)

Backward:

for i in $\{N, \dots, 1\}$:

compute $\frac{\partial \mathcal{L}}{\partial v_i} = \sum_{j \in \text{Children}(v_i)} \frac{\partial \mathcal{L}}{\partial v_j} \frac{\partial v_j}{\partial v_i}$



$$\frac{\partial z}{\partial u} = \frac{\partial z}{\partial v}\frac{\partial v}{\partial u} + \frac{\partial z}{\partial w}\frac{\partial w}{\partial u}$$

$v(u+\delta) \approx v(u) + \delta \frac{\partial v}{\partial u}(u)$ **

$w(u+\delta) \approx w(u) + \delta \frac{\partial w}{\partial u}(u)$

* w/o δ^2 or higher

$z(u+\delta) = z(u) + \delta \frac{\partial z}{\partial v}\frac{\partial v}{\partial u} + \delta \frac{\partial z}{\partial w}\frac{\partial w}{\partial u}$

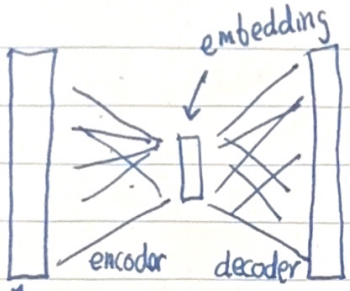
Language Embeddings

Our models take vectors... how can we convert words to vectors?

Approach #1: One-hot encodings

- ⊖ not meaningful
- ⊖ large

Approach #2: Autoencoders



one hot encoding
vocab size x1

Why train decoder?

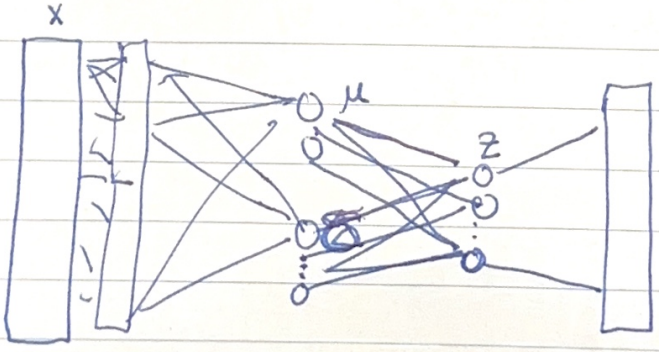
Questions for Activity:

- What loss?
- How to make similar words close?

Approach #3: Variational Autoencoders

What if we want embeddings to be nicely distributed in latent space?

$z \sim N(0, I)$ $z \in \mathbb{R}^r$ $z_i = \mu_i + \sigma_i \epsilon_i$ for $N(0,1)$



$\mathcal{L}(w) = \| f(x) - x \|_2^2 + \lambda \, KL(N(0,I), N(\mu,\sigma))$

↑ distance between distributions

$KL(p\|q) = H(p,q) - H(p)$

Contrastive Learning

We're working on unsupervised task ie., no labels

positive (word, next_word) hopefully close

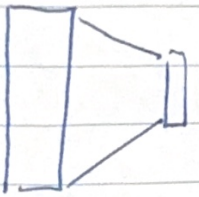
↳ (x, x^+) $f(x)^T f(x^+)$ large

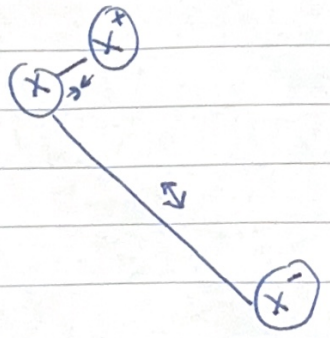
negative (word, unrelated_word) probably far

↳ (x, x^-) $|f(x)^T f(x^-)|$ small

$$\mathcal{L}(w) = -\sum_{x, x^+} f(x)^T f(x^+) + \sum_{x, x^-} \left[f(x)^T f(x^-)\right]^2$$

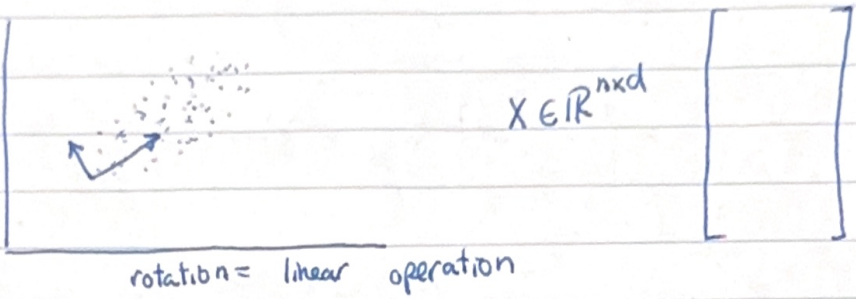
$f: \mathbb{R}^d \to \mathbb{R}^n$





Principal Component Analysis

Motivation: visualize points in $\mathbb{R}^d$ meaningfully



Find $v \in \mathbb{R}^d$: Xv is meaningful ie., captures variation of data!

$(n \times d)(d \times 1)$

$\iff$ $\|Xv\|_2^2$ is large $\iff$ $v^T X^T X v$ is large

$\max_{v: \|v\|_2^2 = 1} v^T X^T X v \iff$ largest eigenvalue of $X^T X$

$\max_{v: \|v\|_2^2 = 1, v \perp v^{(1)}} v^T X^T X v \iff$ 2nd largest eigenvalue

$$X^T X = \sum_{i=1}^{r} \lambda_i v^{(i)} v^{(i)T}$$ (r ← rank)

$(d \times 1)(1 \times d)$

where $\|v^{(i)}\|_2^2 = 1$ and $v^{(i)} \cdot v^{(j)} = 0$

maximize eigenvalue corresponding to largest λ_i

sanity check: $X^T X v^{(j)} = \sum_{i=1}^{r} \lambda_i v^{(i)} v^{(i)T} v^{(j)} = \lambda_j v^{(j)} v^{(j)} \cdot v^{(j)} = \lambda_j v_j$