NLP and Word Embeddings

Word2Vec

deeplearning.ai

Skip-grams

I want a glass of orange juice to go along with my cereal.

Context	Target
orange	juice
orange	glass
orange	my

[Mikolov et. al., 2013. Efficient estimation of word representations in vector space.]

Andrew Ng

Model

Vocab size = 10,000k

x → y

Context c ("orange") → Target t ("juice")
6257 4834

O_c → E → e_c → O Softmax → $\hat{y}$

$e_c = E \, O_c$

$e^{\Theta_t^T e_c}$

Θ_t = parameter associated with output t
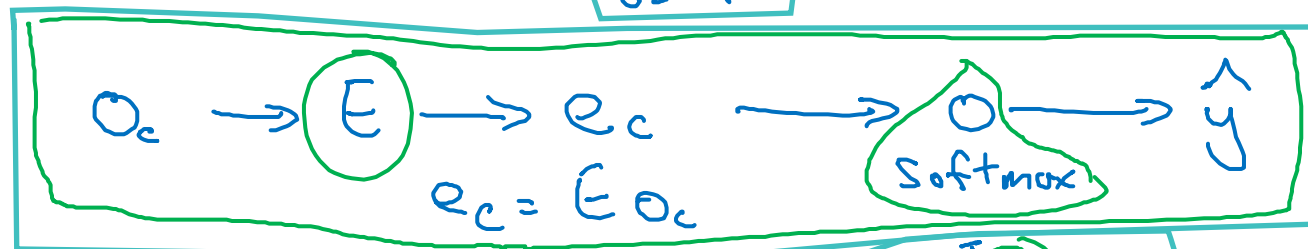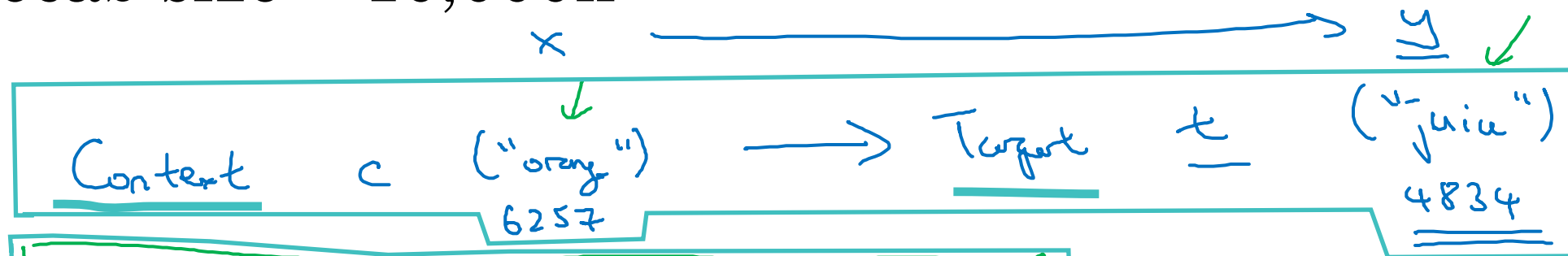
Softmax: $p(t|c) = \dfrac{e^{\Theta_t^T e_c}}{\sum\limits_{j=1}^{10,000} e^{\Theta_j^T e_c}}$

$\mathcal{L}(\hat{y}, y) = -\sum\limits_{i=1}^{10,000} y_i \log \hat{y}_i$

$y = \begin{bmatrix} 0 \\ \vdots \\ i \\ \vdots \\ 0 \end{bmatrix}$ ← 4834

Andrew Ng

Problems with softmax classification

$$p(t|c) = \frac{e^{\theta_t^T e_c}}{\sum_{j=1}^{10,000} e^{\theta_j^T e_c}}$$

Hierachial softmax.

$\log |V|$

This tree puts the frequent words on the top, infrequent words on the bottom.

How to sample the context c?

→ the, of, a, and, to, ...

→ orange, apple, durian

durian

t

$c \to t$

$P(c)$