Word Embedding Equations

Ki Hyun Kim

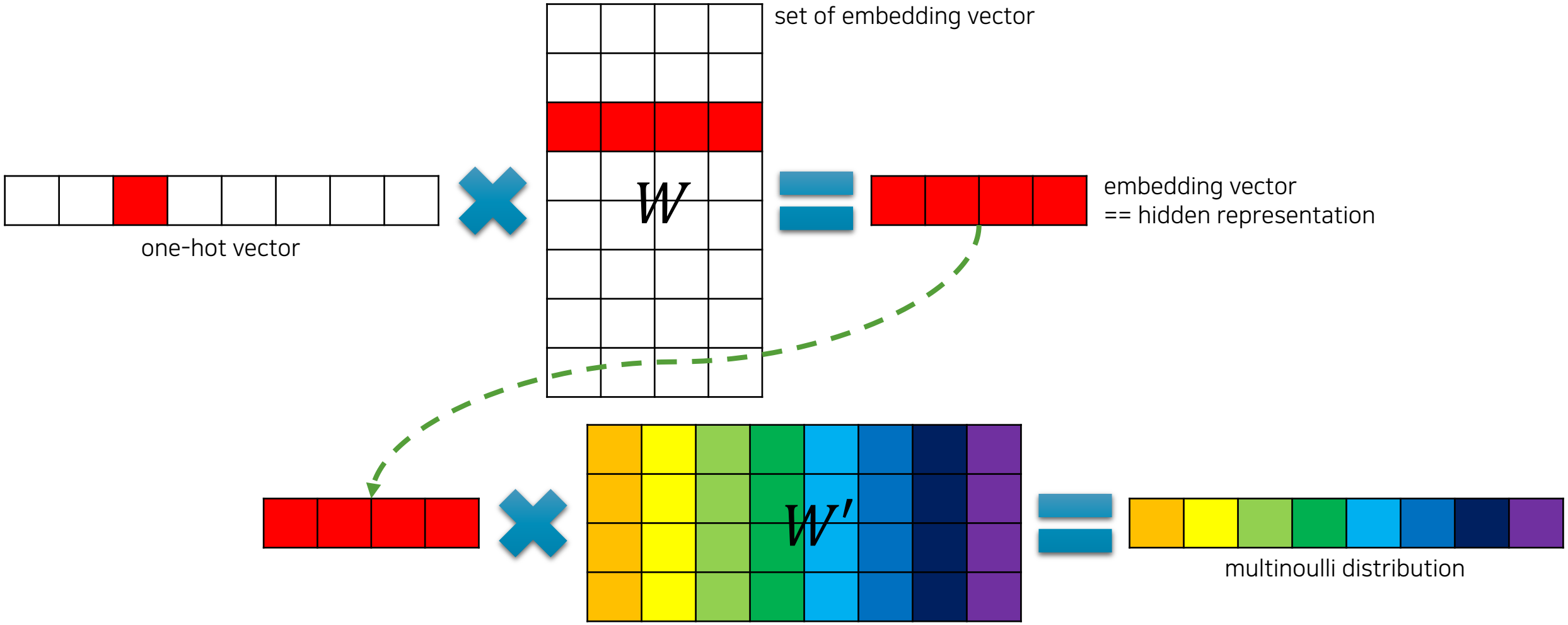
nlp.with.deep.learning@gmail.com

Skip-gram

$$\sum_{t=1}^{T} \sum_{c \in \mathcal{C}_t} \log p(w_c | w_t)$$

$$p(w_c | w_t) = \frac{e^{s(w_t, w_c)}}{\sum_{j=1}^{|V|} e^{s(w_t, w_j)}},$$
$$\text{where } s(w, w') = \mathbf{u}_w^\top \mathbf{v}_{w'}.$$

In Implementation



Negative Sampling

$$\log\left(1 + e^{-s(w_t, w_c)}\right) + \sum_{n \in \mathcal{N}} \log\left(1 + e^{s(w_t, w_n)}\right),$$

where $\mathcal{N}$ is a set of negative examples sampled from the vocabulary.

GloVe

- Turn into regression task from classification task.

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{x \in \mathcal{X}} f(x) \times ||W'Wx - \log C_x||_2^2,$$

$$\text{where } C_x \text{ is a vector of co-occurences with x,}$$

$$W \in \mathbb{R}^{d \times |V|} \text{ and } W \in \mathbb{R}^{|V| \times d}.$$

$$f(x) = \begin{cases} (\text{count(x)} \ / \ \text{thres})^{\alpha} & \text{if count}(x) < \text{thres}, \\ 1 & \text{otherwise.} \end{cases}$$

FastText

- Same as Skip-gram, but it uses sum of subword one-hot vector, instead of using word one-hot vector.

$$\sum_{t=1}^{T} \sum_{c \in \mathcal{C}_t} \log p(w_c|w_t) \qquad p(w_c|w_t) = \frac{e^{s(w_t, w_c)}}{\sum_{j=1}^{|V|} e^{s(w_t, w_j)}},$$

$$s(w, w') = \sum_{g \in \mathcal{G}_w} \mathbf{z}_g^\top \mathbf{v}_{w'},$$

where $\mathcal{G}_w$ is a set of subword n-grams.

e.g. $\mathcal{G}_{w=\text{where}} = \{<\text{wh}, \text{whe}, \text{her}, \text{ere}, \text{re}>\}$.