Log-Bilinear Language Model

- Has much less parameters and non-linear activations
- Measures similarity between the word and the context:

$$p(w_i|w_{i-n+1},\dots w_{i-1}) = \frac{\exp(\hat{r}^T r_{w_i} + b_{w_i})}{\sum_{w \in V} \exp(\hat{r}^T r_w + b_w)}$$

Representation of word:

$$r_{w_i} = C(w_i)^T$$

Representation of context:

$$\hat{r} = \sum_{k=1}^{n-1} W_k C(w_{i-k})^T$$

Andriy Mnih and Geoffrey Hinton. 2007. Three new graphical models for statistical language modelling. In *Proceedings of the 24th international conference on Machine learning* (ICML '07)