

# A Neural Probabilistic Language Model

## 1 Main Idea

This is the seminal paper on neural language modeling that first proposed learning distributed representations of words. There is an obvious distinction made for predictions in a discrete vocabulary space vs. predictions in a continuous space, and the solution proposed is to have real-valued word feature vectors that are learnt along with the joint probability function of their occurrence in sequences in the corpus.

## 2 Method

- The probability function is expressed as the product of the conditional probabilities of the next word given the current word.
- In contrast to future works like Word2Vec, this paper considers the next word's probability distribution to be conditioned only on a window of  $n$  words that precede it.
- Neural architecture comprises of an embedding layer, a hidden *tanh* layer and an output *softmax* layer.
- Experiments benchmarked this system on the Brown and AP News corpora and it was compared to n-gram models, with the performance metric being perplexity (lower = better).

## 3 Observations

- The neural language model had test perplexity gains of 24% for Brown and 8 on AP news compared to the state-of-the-art n-gram models (smoothed trigram).

## References