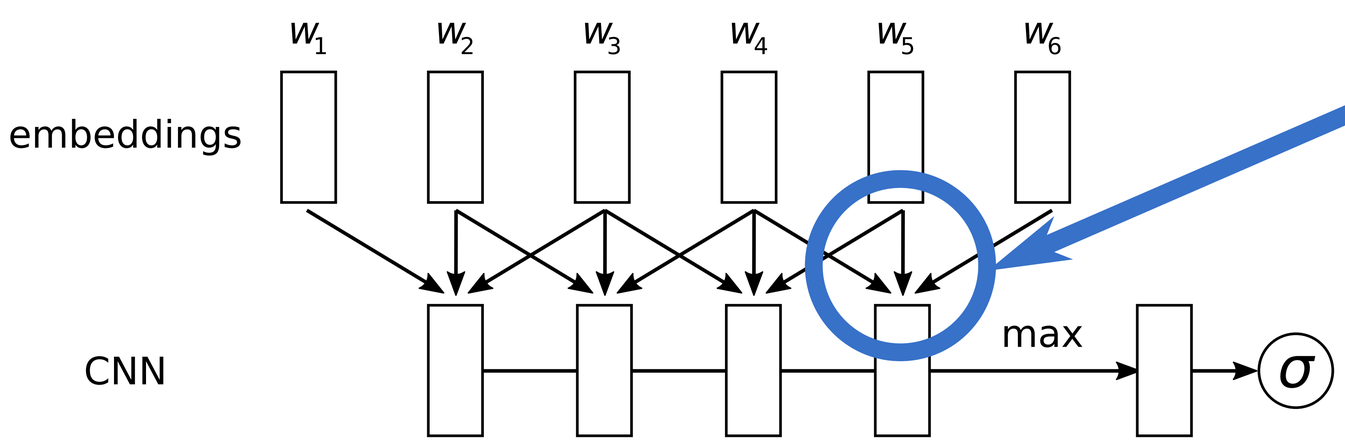
Neural Networks as Explict Word-Base Rules

Jindřich Libovický libovicky@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University

CNN for sentiment classification of IMDB movie reviews



What are the words that corespond the most to the filters in the trained CNN?

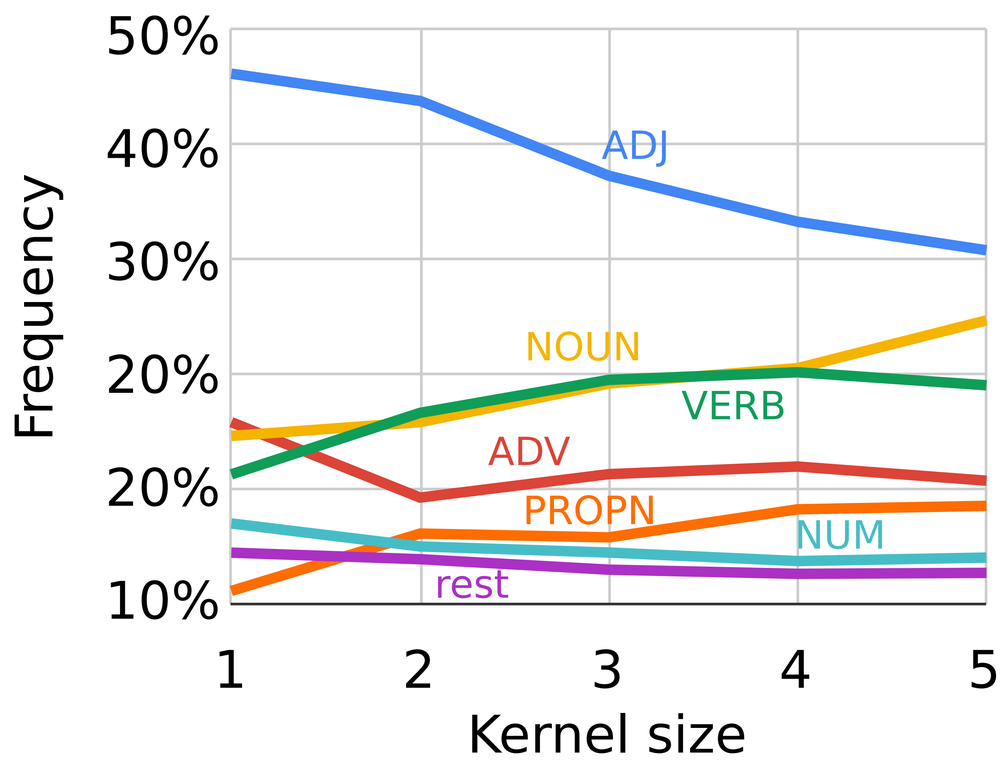
yawn, incoherent, pointless, delight, perfect, brilliant, innocent, disappointing, ...

Simple CNN for sentence classification can be **interpreted** using **word rules**.

Recovering the original models

kernel size	original CNN	rules n-grams	rules BOW
1	83.4	83.0	83.0
2	85.9	82.4	81.5
3	87.2	81.8	81.4
4	87.7	81.9	81.0
5	87.8	82.0	81.6

POS tags in the filters



Sentiment lexicon

- 60% extracted words in Opinion Lexicon by Hu and Liu (2004)
- 99% accuracy w.r.t. the lexicon

This work has been supported by the grant 18-02196S of the Czech Science Foundation.

The poster was presented on August 1, 2019 at the BlackboxNLP workshop in Florence, Italy.