IN5550: Neural Methods in Natural Language Processing
Sub-lecture 5.2
Deep learning and language models

Andrey Kutuzov, Shenbin Qian

University of Oslo

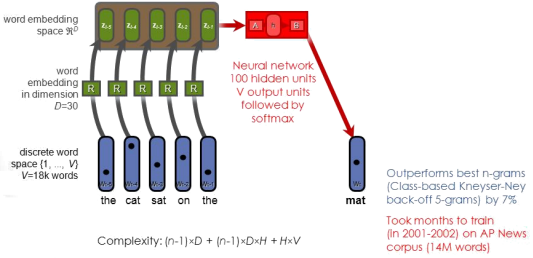


Contents

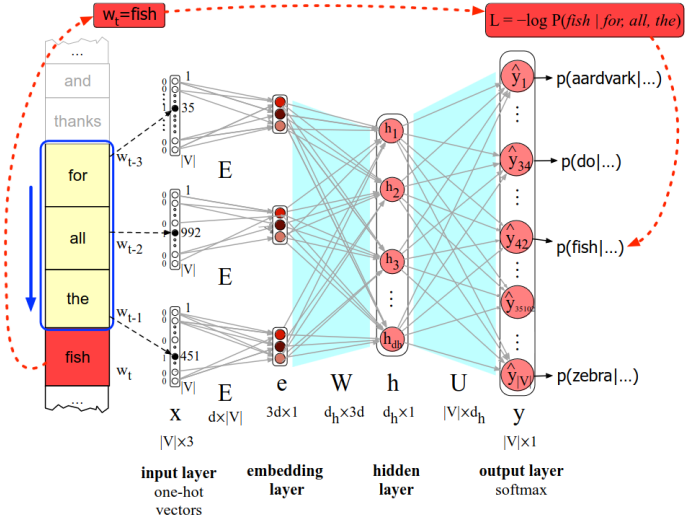
1. New way: neural language modeling
   - Feedforward neural LM
   - Modern LMs
2. Benefits and shortcomings

New way: neural language modeling

- Neural LM model proposed in [Bengio et al., 2003]:
- concatenate learned embeddings of the previous k words;
- this concatenation is fed into a feed-forward neural network...
- ...with hidden layers and non-linearities;
- cross-entropy loss, the next words as the gold predictions.
- Output probability distribution over possible next words across the vocabulary V, using softmax and the second embedding matrix (remember word2vec?).
- Input and output vocabularies can be different.



Feedforward neural LM



Feedforward neural LM moving through the text of 'The Hitchhiker's Guide to the Galaxy'

(from Jurafsky and Martin, 2026)

New way: neural language modeling

What about word embeddings?

- The neural LM learns representations for words as a byproduct of the training process.
- Language models are trained on raw texts, no manual annotation needed.
- And we have lots of raw texts.
- Language modeling can be a tool to provide good embeddings for other tasks
  - in addition to being useful in itself (to predict next words or generate text)

Modern LMs

The world is changing fast

- Modern state-of-the-art neural language models are mostly based on Transformer architectures.
- Of course you are aware of ChatGPT (`https://gpt.uio.no/`) or DeepSeek (`https://chat.deepseek.com`) models
- NorMistral: web chat interface for large Norwegian language models: `https://chat.llm.sigma2.no/` (pilot service, but fully functional).
- More on that in the next lectures.

Contents

1. New way: neural language modeling
   - Feedforward neural LM
   - Modern LMs

2. Benefits and shortcomings

Benefits and shortcomings

Benefits

- Neural models outperform non-neural LMs as measured by perplexity.
- Scale well: higher k leads to linear increase in the parameters number...
- ...in traditional LMs it was exponential.
- Words in different positions share statistical strength.
- Generalizations to unseen data: similar words get similar representations:
  - '*fox eats*': seen 1000 times; '*dog eats*': seen 1000 times; '*wolf eats*': seen 0 times; $\hat{P}([wolf, eats]) \gg 0$, because '*wolf*' is similar to '*fox*' and '*dog*'.
- Can easily add more hidden layers.

Shortcomings

- Expensive softmax over V in the output layer.
- Increasing the output $|V|$ can significantly slow down the network.
- There are ways to deal with this (again, recall word2vec).

References I


Bengio, Y., Ducharme, R., and Vincent, P. (2003).
A neural probabilistic language model.
Journal of Machine Learning Research, 3:1137–1155.