Deep Learning for Natural Language Processing

Introduction to transfer learning and pre-trained embeddings

UNIVERSITY OF GOTHENBURG

CHALMERS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Richard Johansson

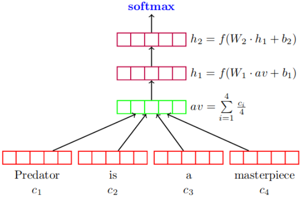
richard.johansson@gu.se

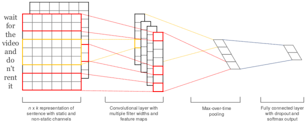
recap: embeddings

- in a neural network, an **embedding layer** represents a symbol as a continuous vector

2739 ("cucumber") ⟶ [0.7, -1.2, ..., -0.1]

- we've seen how **word embeddings** are used as the first layer in NLP systems such as categorizers



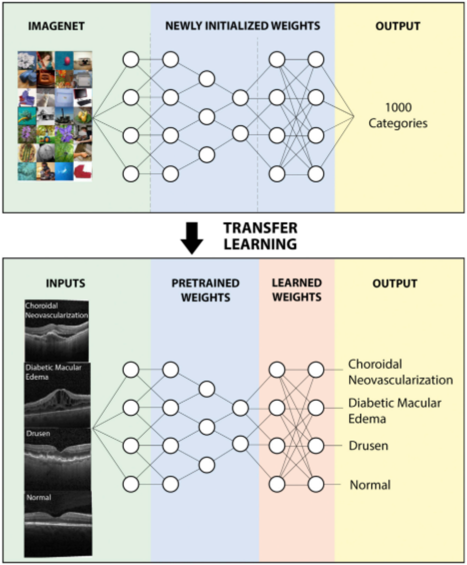


- so far, we trained the word embeddings **from scratch**

transfer learning: idea and motivation

- in **transfer learning**, we try to exploit previously learned knowledge when solving new tasks
- in practice: after training, we **reuse some part** of the model
- why? because it can **reduce the need for training data** for the target task
- commonly used when training ML models for vision tasks

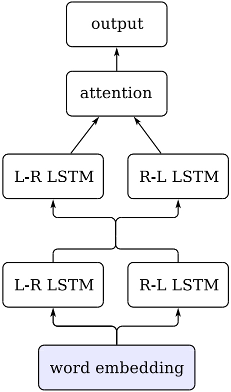
transfer learning in vision



[source]

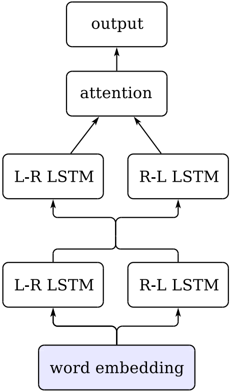
transfer learning in NLP

this lecture:

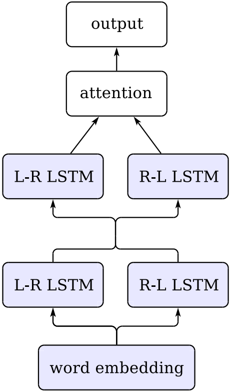


transfer learning in NLP

this lecture:



later:



key challenges for transfer learning

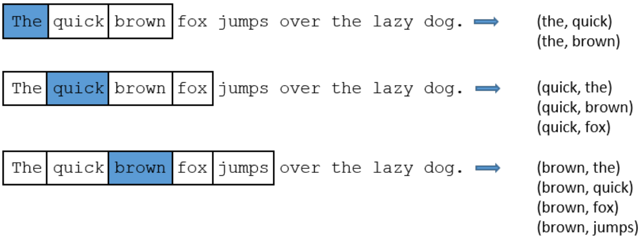
- learning **generally useful** representations
  - so we need fairly general training tasks
- finding **training data**
  - ideally, an unlimited supply!

key challenges for transfer learning

- learning **generally useful** representations
  - so we need fairly general training tasks
- finding **training data**
  - ideally, an unlimited supply!
  - in NLP, we prefer to use **raw text** (unannotated) for pre-training representations

predicting contexts

- all pre-training methods for word embeddings are based on predicting what kind of **context** a word appears in
  - for instance, the surrounding words
- easy to generate large amount of training data



justification in terms of linguistic theory

- *"you shall know a word by the company it keeps"* (Firth, 1957)
- two words probably have a similar "meaning" if they tend to appear in similar **contexts**
- the **distributional hypothesis** (Harris, 1954): the distribution of contexts in which a word appears is a good proxy for the "meaning" of that word

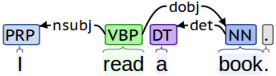
example: most frequent verbs near *cake* and *pizza*

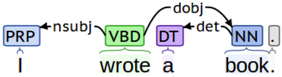
- ***cake***: eat, bake, throw, cut, buy, get, decorate, garnish, make, serve, order

- ***pizza***: eat, bake, order, munch, buy, serve, garnish, name, get, make, heat

so what kinds of "contexts" can we use?

- surrounding words: rest of today's talk
- alternatives:
  - documents (Landauer and Dumais, 1997)
  - syntax (Padó and Lapata, 2007)

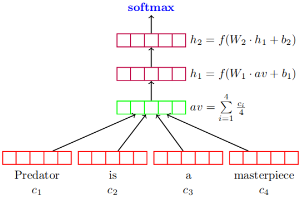


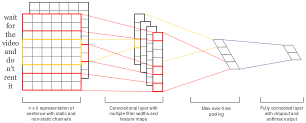


  - images (Lazaridou et al., 2015)

using word embeddings in NLP applications

- the pre-trained word embeddings can then be "plugged" into NLP applications





- how? two alternatives:
  - let the word embeddings be fixed
  - **fine-tune** the embeddings for the application

next lecture clips

- the SGNS (`word2vec`) training algorithm
- evaluation and interpretation
- more training methods
- research outlook

references

J. Firth. 1957. *Papers in Linguistics 1934–1951*. OUP.

Z. Harris. 1954. Distributional structure. *Word* 10(23):146–162.

T. K. Landauer and S. T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review* 104:211–240.

A. Lazaridou, N. T. Pham, and M. Baroni. 2015. Combining language and vision with a multimodal skipgram model. In *NAACL*.

S. Padó and M. Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33(2):161–199.