Deep Learning for Natural Language Processing

Transfer learning using language models

UNIVERSITY OF
GOTHENBURG

CHALMERS

WASP | WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Richard Johansson

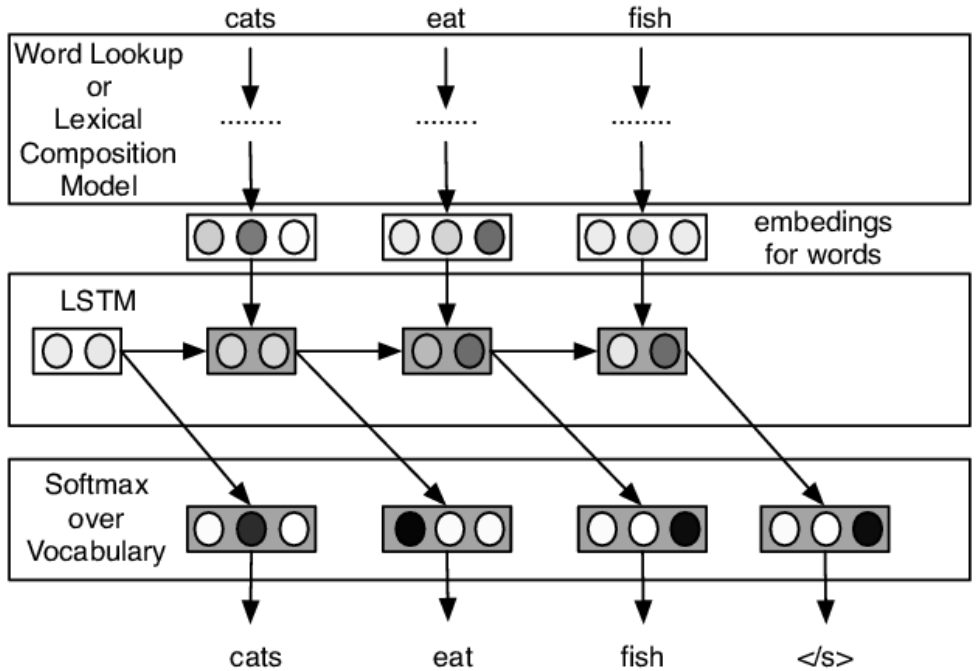
richard.johansson@gu.se

language models

- **language models** compute a probability for a given text
  - how probable is this sentence?
  - the user typed some words; what is the most likely next word?
  - which sequence is more probable?
    - *precedent Smith* or *president Smith*?
    - *strong tea* or *powerful tea*?

neural language models

- neural LMs were introduced by Bengio et al. (2003): first just using embeddings and a feedforward model
- RNNs for LMs introduced by Mikolov et al. (2010)



- modern representative: (Jozefowicz et al., 2016)

[source]

is language modeling a useful task for transfer learning?

I was sad because my football team had ___

is language modeling a useful task for transfer learning?

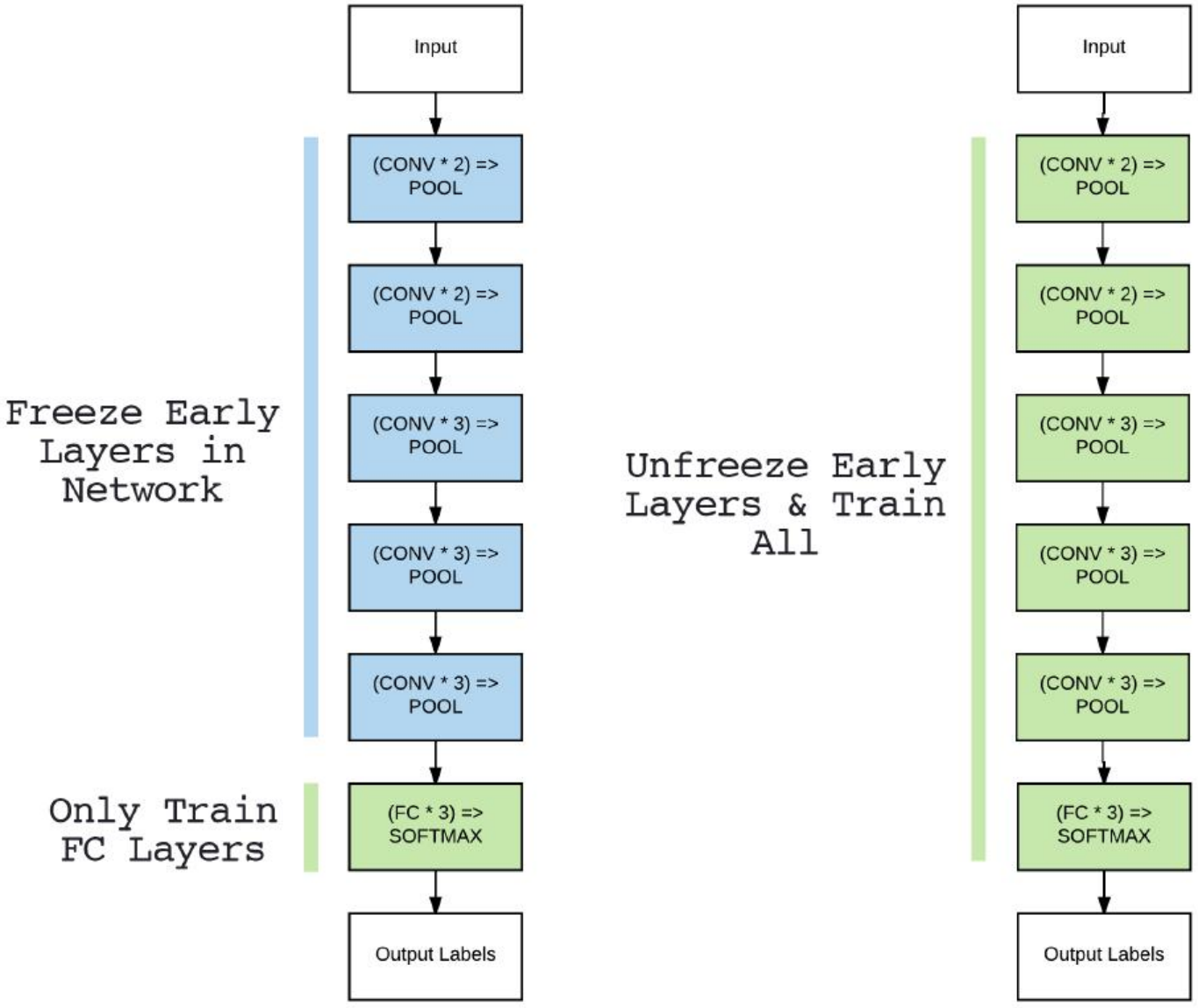
I was sad because my football team had ___

- to predict the correct word (probably ***lost***), the model needs to handle several linguistic levels:
  - **semantics and world knowledge**: what can football teams do? why am I sad?
  - **syntax and morphology**: we expect a verb in the past participle in this position
- also, **training data** is easy to access: no annotation needed

using language models for transfer learning

- the idea of using LMs for transfer learning had been floating around for some time, e.g. Dai and Le (2015)
- but the idea really caught on in 2018 with the publication of ELMo (Peters et al., 2018)

two high-level approaches to transfer learning in NNs



[source]

tradeoffs between freezing and fine-tuning

- if we **freeze** the pre-trained model, training is fast but there is a risk that the pre-trained part is not optimal for our task
- if we **fine-tune**, we are more flexible but risk forgetting what we learned previously: **catastrophic forgetting** (McCloskey and Cohen, 1989)

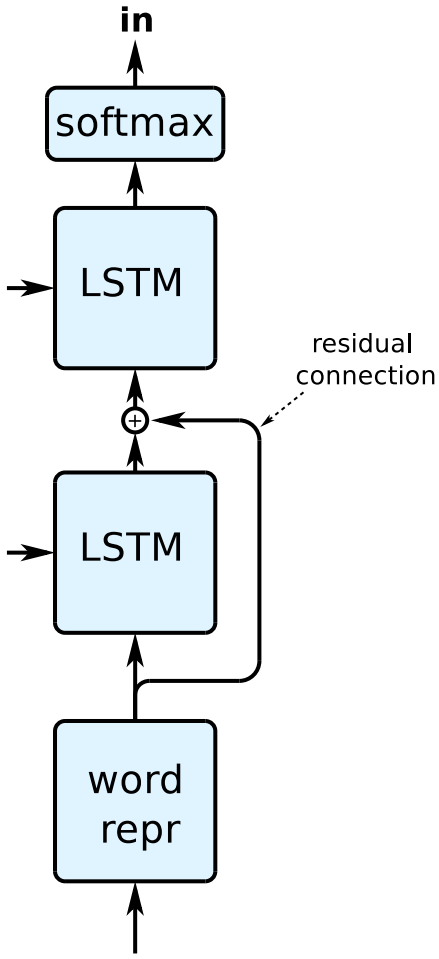
ELMo: details

- **ELMo** (**E**mbeddings from **L**anguage **Mo**dels) sparked the transfer learning craze (Peters et al., 2018)

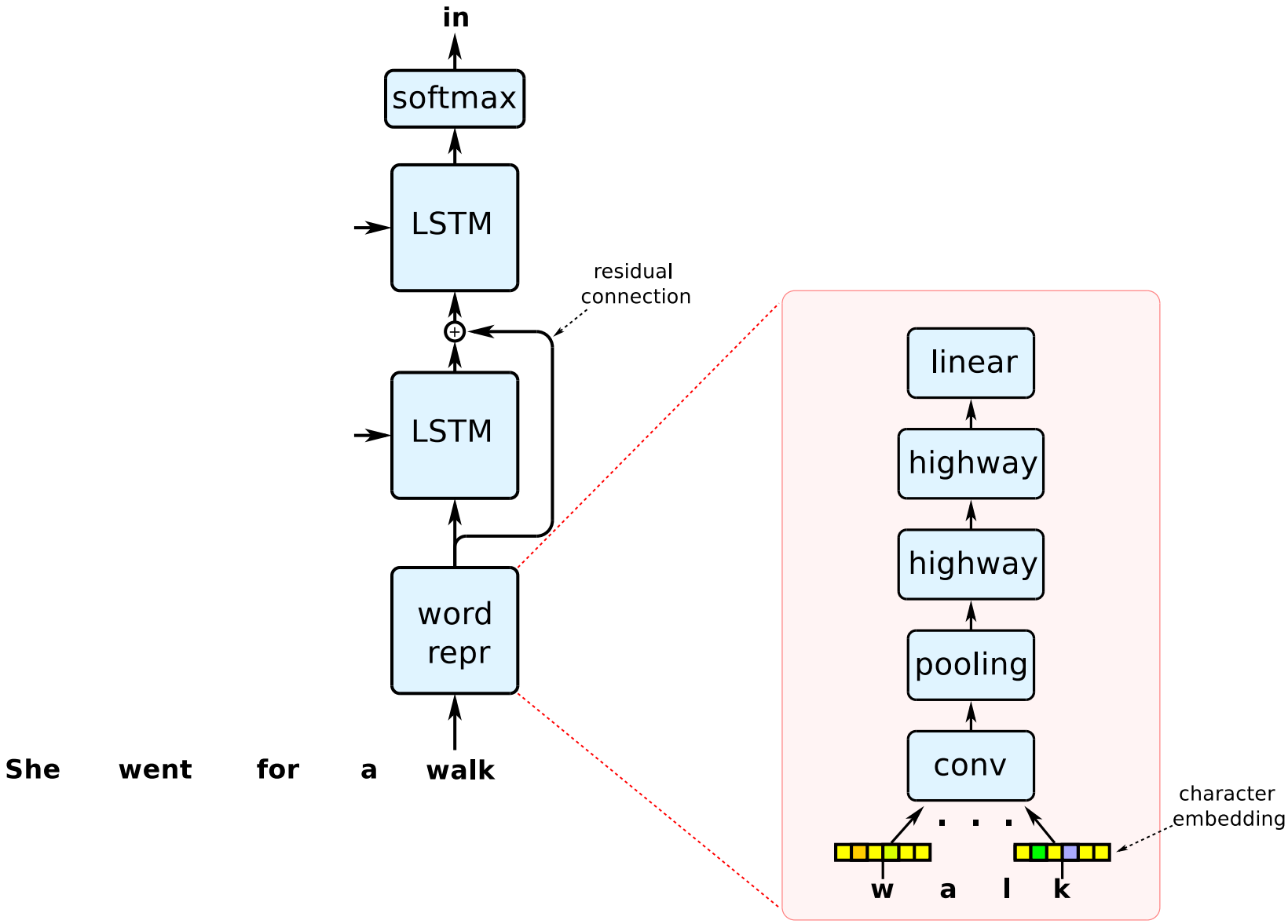
  `https://allennlp.org/elmo`

- its core is a combination of two neural language models

language models in ELMo



language models in ELMo



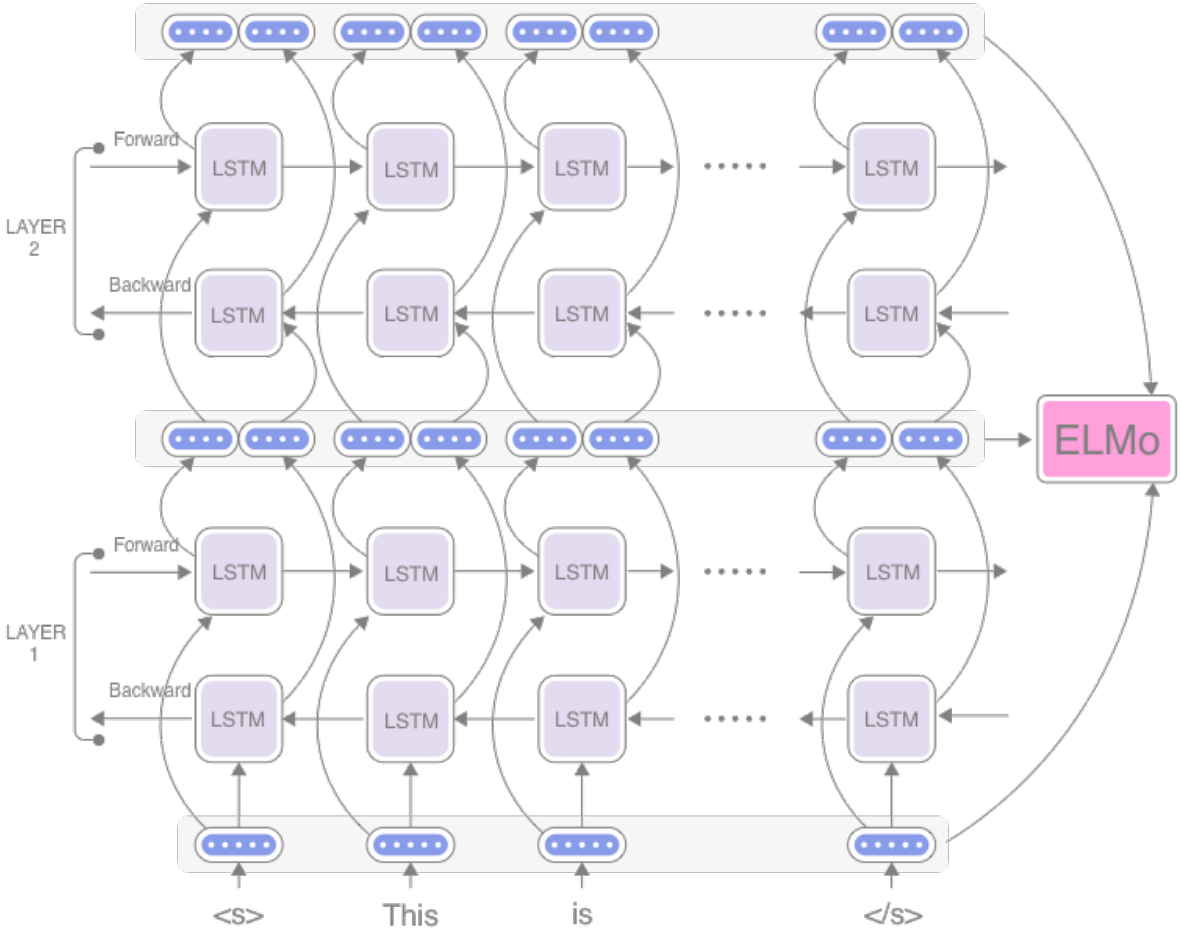
using ELMo in applications

- when used for transfer learning, ELMo computes a weighted sum of the outputs of all the LM's layers

$$\text{ELMo}_k = \gamma \sum_{j=0}^{L} s_j \boldsymbol{h}_{k,j}$$

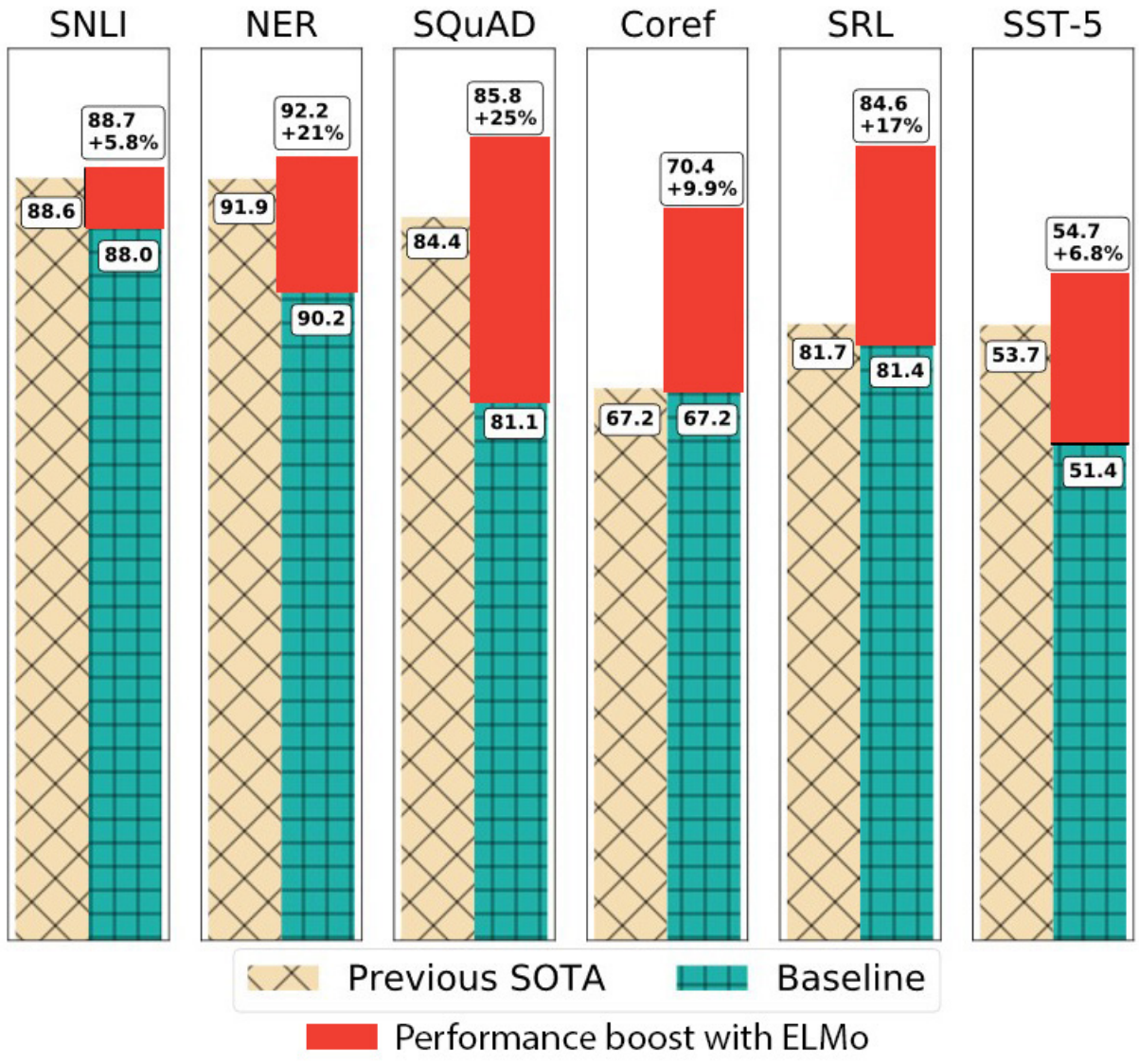
  where γ and s_j are learned weights
- ELMo complements or replaces a standard word embedding layer
- the basic ELMo model is frozen after pre-training
  - but **domain-specific fine-tuning** of LMs can improve

computing word representations in ELMo



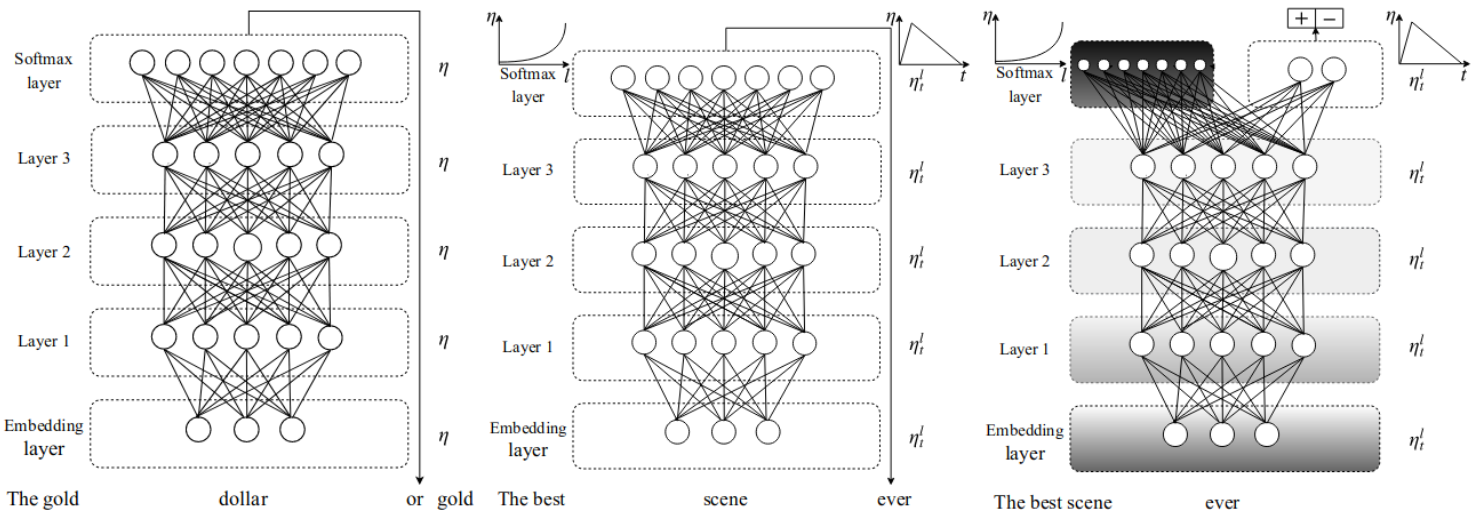
[source]

ELMo: results



ULMFiT: careful fine-tuning of layers

- Howard and Ruder (2018) also used bidirectional language models for transfer learning
- case study: text categorization
- they **fine-tune** using **different learning rates**



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

references

Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. 2003. A neural probabilistic language model. *JMLR* 3:1137–1155.

A. Dai and Q. Le. 2015. Semi-supervised sequence learning. In *NIPS 28*.

J. Howard and S. Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL*.

R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu. 2016. Exploring the limits of language modeling. arXiv:1602.02410.

M. McCloskey and N. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. *The Psychology of Learning and Motivation* 24:109–165.

T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*.

M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.