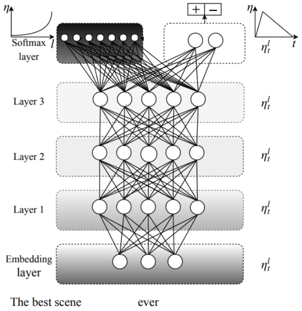
Transfer Learning

ULMFiT (Howard & Ruder, 2018)



Learning goals

- Understand the paradigm of fine-tuning for Transfer Learning
- Get the intuition of the subtleties of training ULMFiT

FINE-TUNING APPROACH

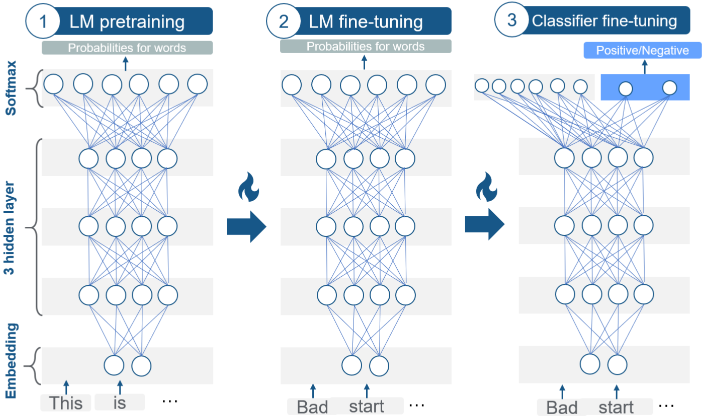
Shortcomings of ELMo:

- No adaption of the Embeddings to target domain/task
  - Source & target domain/task might be pretty different
  - No good representations for domain-/task-specific words
- Sequential nature of LSTMs:
  - Not fully parallelizable (compared to Transformers, see next chapter)
  - Fails to capture long-range dependency during contextualization

Alleviations/Alternatives:

- ULMFiT ▸ Howard and Ruder, 2018 is a uni-directional LSTM which is fine-tuned as a whole model on data from the target domain/task.
- GPT ▸ Radford et al., 2018 is a Transformer (decoder) which is fine-tuned as a whole model on data from the target domain/task.

ULMFIT ▸ HOWARD AND RUDER, 2018



Source: *Carolin Becker*

ARCHITECTURAL DETAILS

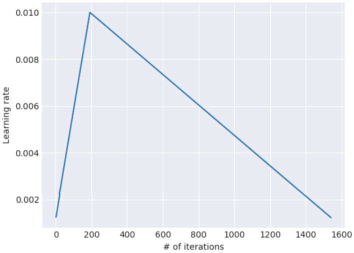
- AWD-LSTMs ▸ Merity et al., 2017 as backbone of the architecture
  - DropConnect ▸ Wan et al., 2013
  - Averaged stochastic gradient descent (ASGD) for optimiziation
- Embedding layer
  - $E = 400$
  - Some word-level tokenization (not entirely clear)
- Three LSTM layers ($H = 1150$) + Softmax Layer

LM PRE-TRAINING DETAILS

- *Corpus: Wikitext-103*
  - 28.595 Wikipedia articles
  - $\approx$ 103M words
- *Objective:* Language Modeling

LM FINE-TUNING DETAILS

- *Discriminative fine-tuning*
  - Different learning rates for each layer
  - First choose learning rate for the last layer
  - Use this to determine the learning rates for lower layers
- *Slanted triangular learning rates*



Source: *Howard and Ruder (2018)*

CLASSIFIER FINE-TUNING

- *Concat Pooling*
  - Consider not only last hidden state for classification
  - $\mathbf{h}_c = [\mathbf{h}_T, \texttt{maxpool}(\mathbf{H}), \texttt{meanpool}(\mathbf{H})]$
    with $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$
- *Gradual unfreezing*
  - Minimizing risk of 'catastrophic forgetting'
  - Gradually make layers eligible for gradient updates (from top to bottom)
- *BPT3C*
  - Divide documents in chunks of pre-defined length (here: 70)
  - Initialize model with the final hidden state of previous chunk
- *Bidirectionality*
  - Same procedure for a backward model

PERFORMANCE

	Model	Test		Model	Test
IMDb	CoVe (McCann et al., 2017)	8.2	TREC-6	CoVe (McCann et al., 2017)	4.2
	oh-LSTM (Johnson and Zhang, 2016)	5.9		TBCNN (Mou et al., 2015)	4.0
	Virtual (Miyato et al., 2016)	5.9		LSTM-CNN (Zhou et al., 2016)	3.9
	ULMFiT (ours)	**4.6**		ULMFiT (ours)	**3.6**

Table 2: Test error rates (%) on two text classification datasets used by McCann et al. (2017).

	AG	DBpedia	Yelp-bi	Yelp-full
Char-level CNN (Zhang et al., 2015)	9.51	1.55	4.88	37.95
CNN (Johnson and Zhang, 2016)	6.57	0.84	2.90	32.39
DPCNN (Johnson and Zhang, 2017)	6.87	0.88	2.64	30.58
ULMFiT (ours)	**5.01**	**0.80**	**2.16**	**29.98**

Table 3: Test error rates (%) on text classification datasets used by Johnson and Zhang (2017).

Source: *Howard and Ruder (2018)*