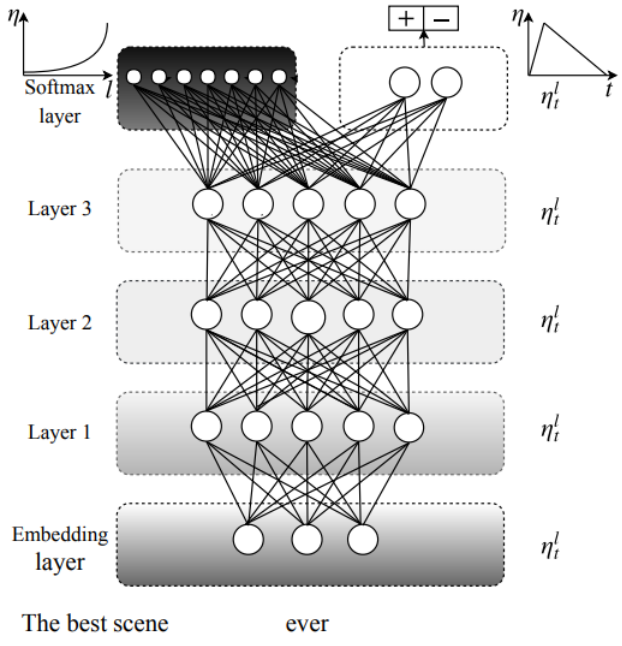
Transfer Learning

ULMFiT (Howard & Ruder, 2018)



Learning goals

- tbd

CONTEXTUALITY

1st Generation of neural embeddings are "context-free":

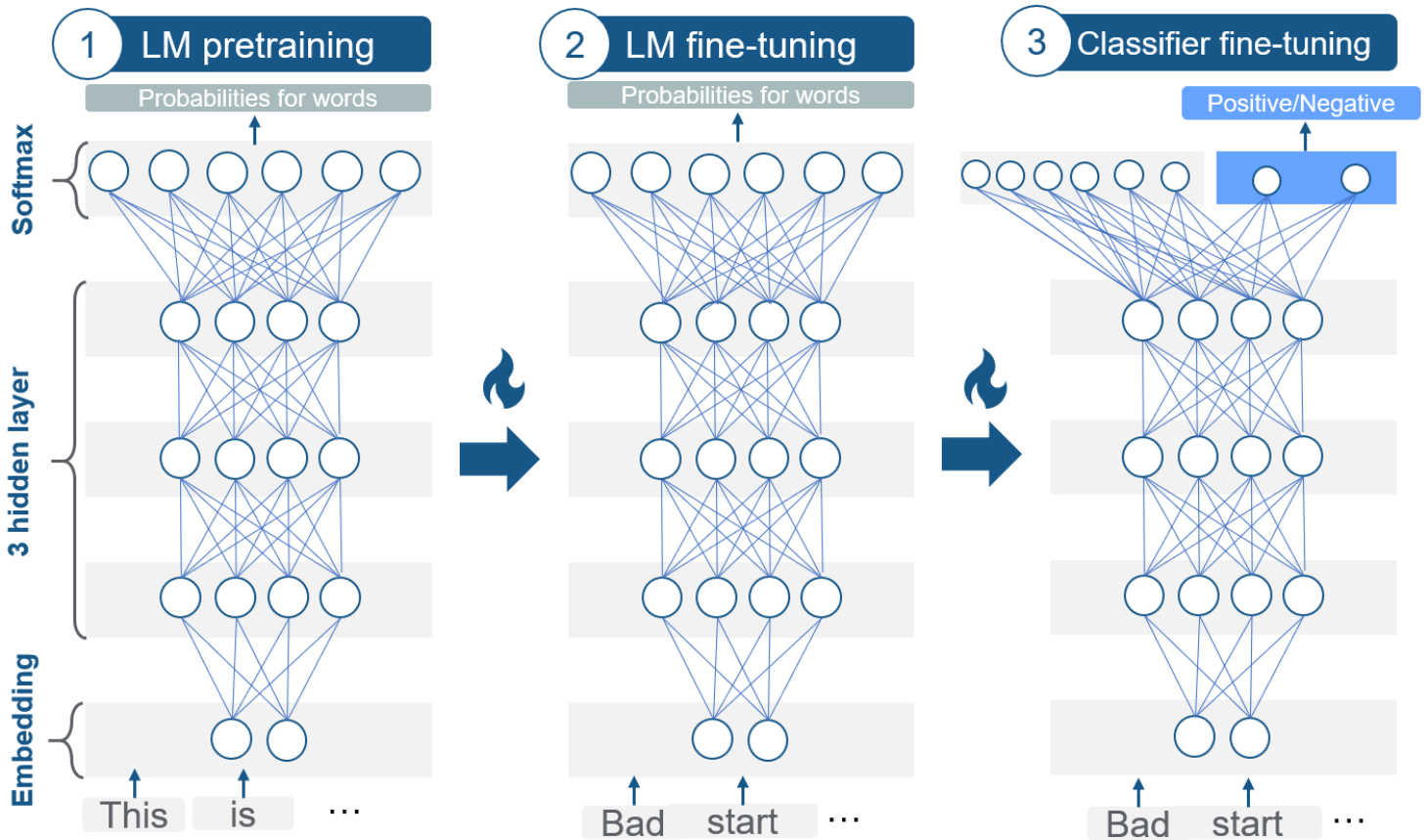
- Breakthrough paper by Mikolov et al, 2013 (Word2Vec)
- Followed by Pennington et al, 2014 (GloVe)
- Extension of Word2Vec by Bojanowski et al, 2016 (FastText)

Why "Context-free"?

- Models learn *one single* embedding for each word
- Why could this possibly be problematic?
  - "The *default* setting of the function is xyz."
  - "The probability of *default* is rather high."
- Would be nice to have different embeddings for these two occurrences

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 + three LSTM layers + Softmax Layer
- **LM fine-tuning:**
  - Discriminative fine-tuning
- **Classifier fine-tuning:**
  - Concat Pooling
  - Gradual unfreezing