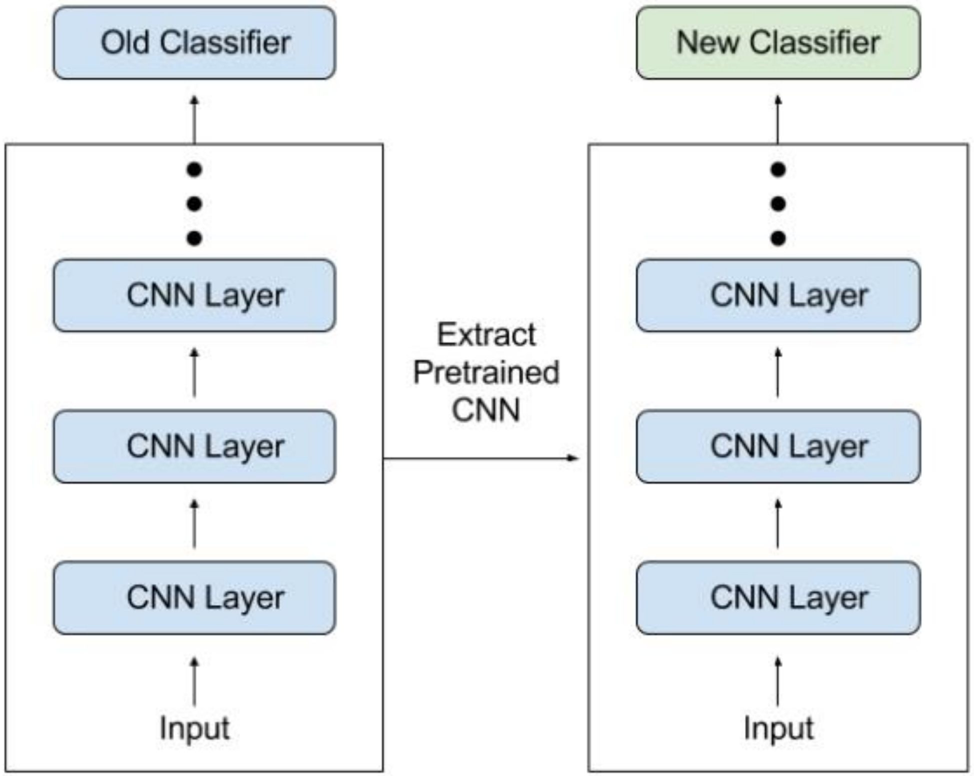
Universal Language Model Fine-tuning for Text Classification

Jeremy Howard*
fast.ai
University of San Francisco
j@fast.ai

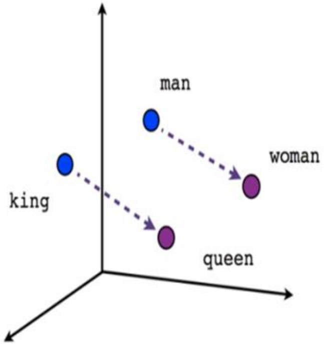
Sebastian Ruder*
Insight Centre, NUI Galway
Aylien Ltd., Dublin
sebastian@ruder.io

2019.02.11
발표자 : 김용규

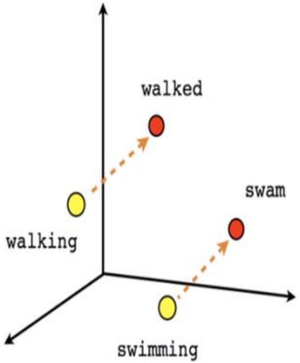
Transfer learning in CV



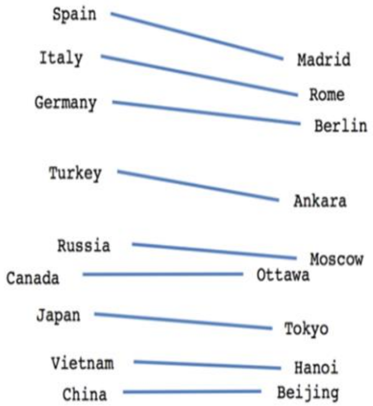
Transfer learning in NLP



Male-Female

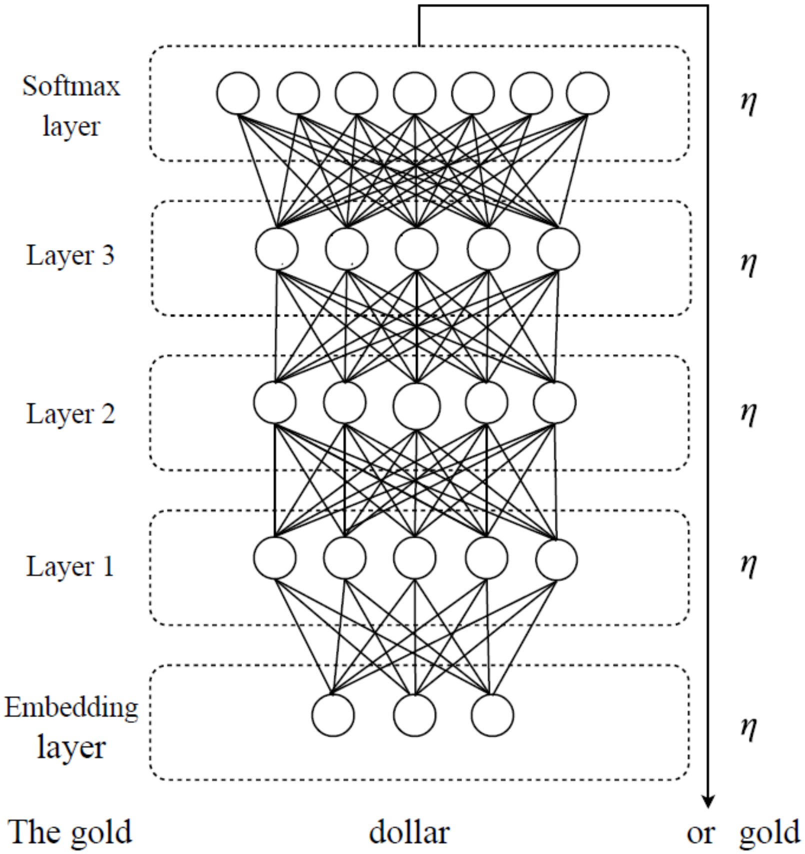


Verb tense



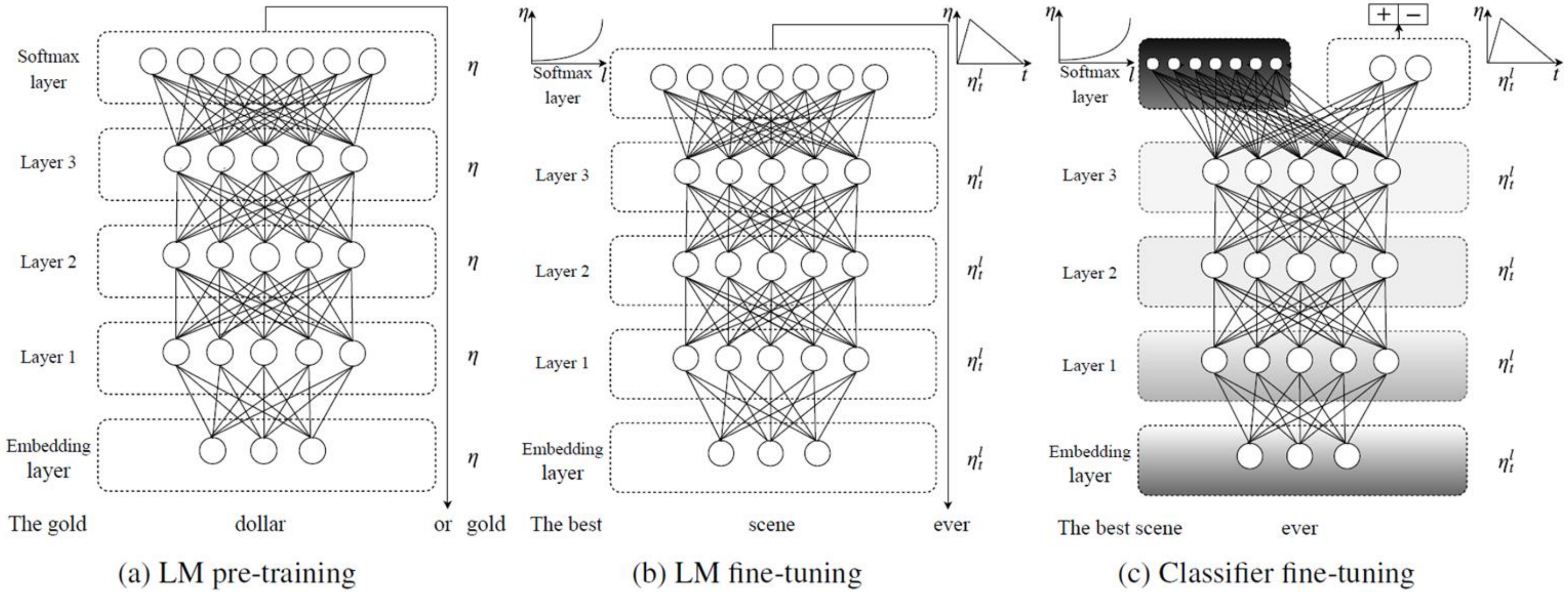
Country-Capital

Language Model



“The service was poor, but the food was _______”

Universal Language Model Fine-tuning for Text Classification



Universal Language Model Fine-tuning for Text Classification

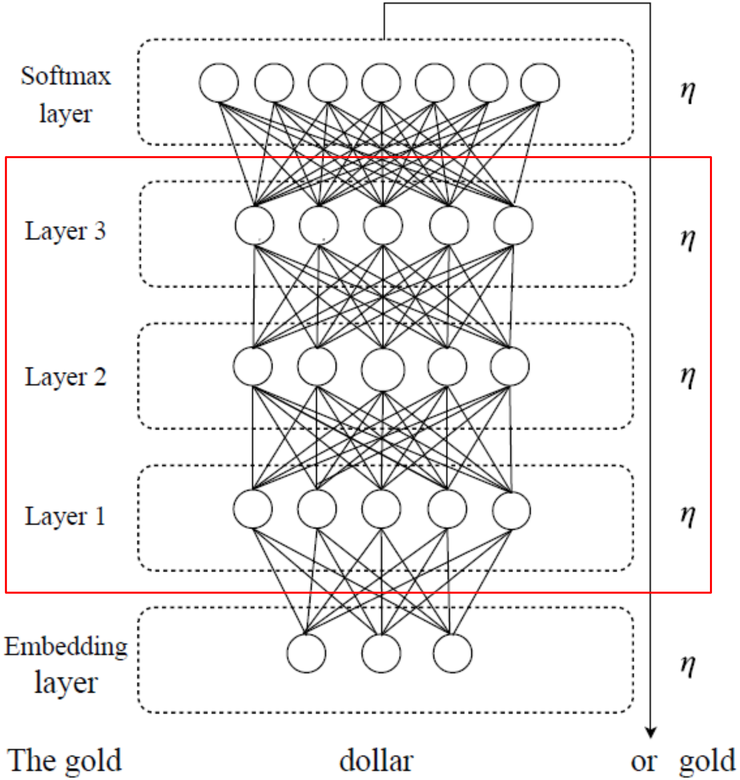
- It works across tasks varying in document size, number, and label type
- It uses a single architecture and training process
- It requires no custom feature engineering or preprocessing
- It does not require additional in-domain documents or labels

Contribution

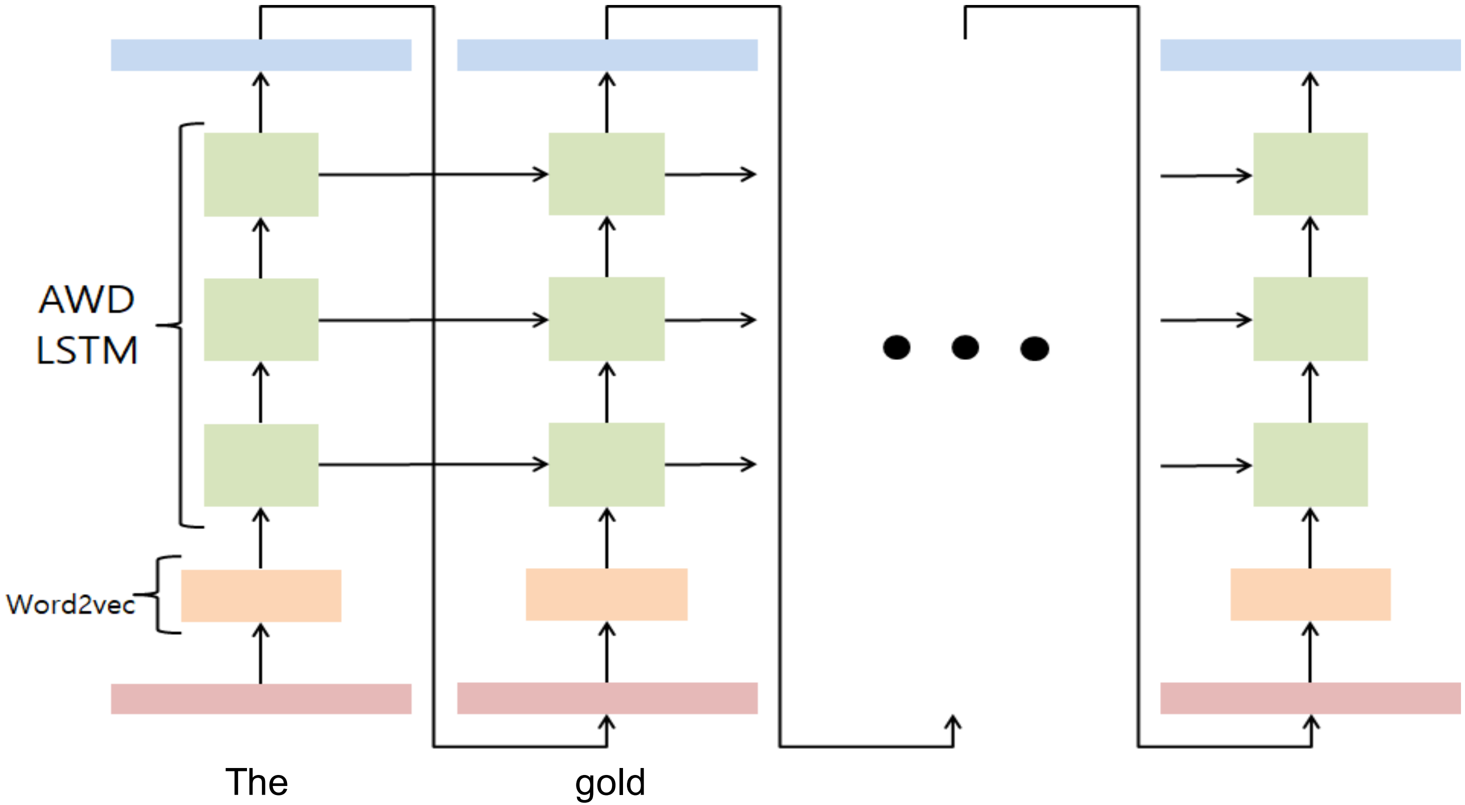
- transfer learning in NLP
- discriminative fine-tuning, slanted triangular learning rates, gradual unfreezing
- increase of text classification performance

(a) General-domain LM pre-training

- Wikipedia article
- AWD-LSTM(Average Stochastic Gradient weight dropped LSTM)



(a) LM pre-training

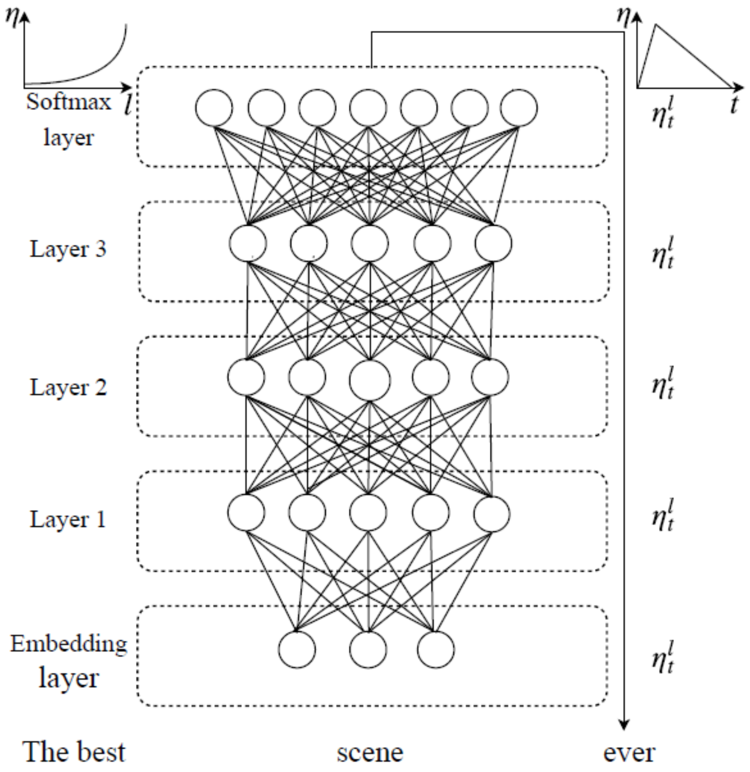

AWD
LSTM
Word2vec
The
gold

AWD-LSTM (Regularizing and Optimizing LSTM)

	AWD-LSTM	LSTM
Regularizing	DropConnect (recurrent connection)	Dropout
Optimizing	LT-ASGD	SGD

(b) Target task LM fine-tuning

- Task-specific data
- Discriminative fine-tuning
- Slanted triangular learning rates



(b) LM fine-tuning

(b)-1 Discriminative fine-tuning

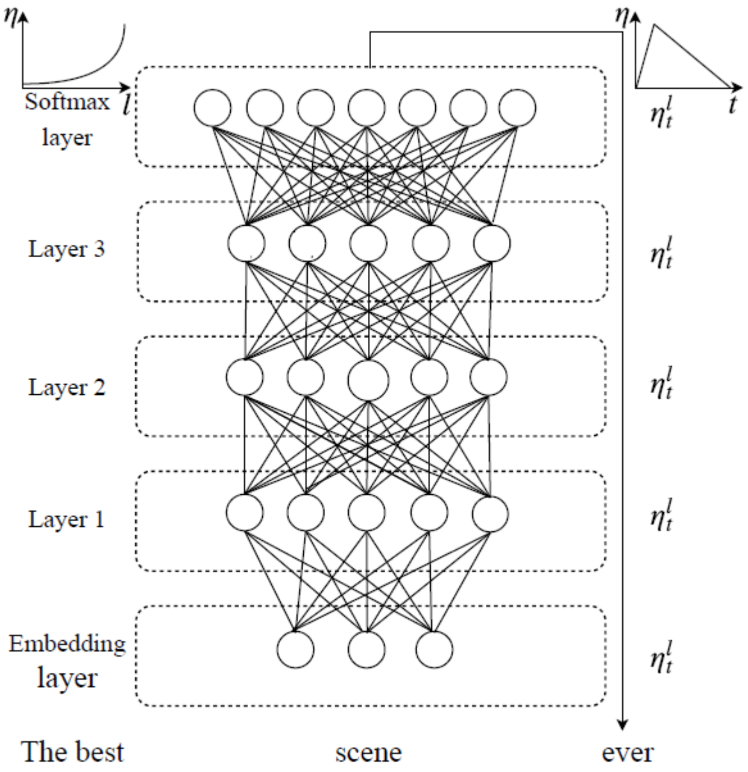
$$\theta_t = \theta_{t-1} - \eta \cdot \nabla_\theta J(\theta)$$

$$\theta = \{\theta^1, \ldots, \theta^L\}$$

$$\eta = \{\eta^1, \ldots, \eta^L\}$$

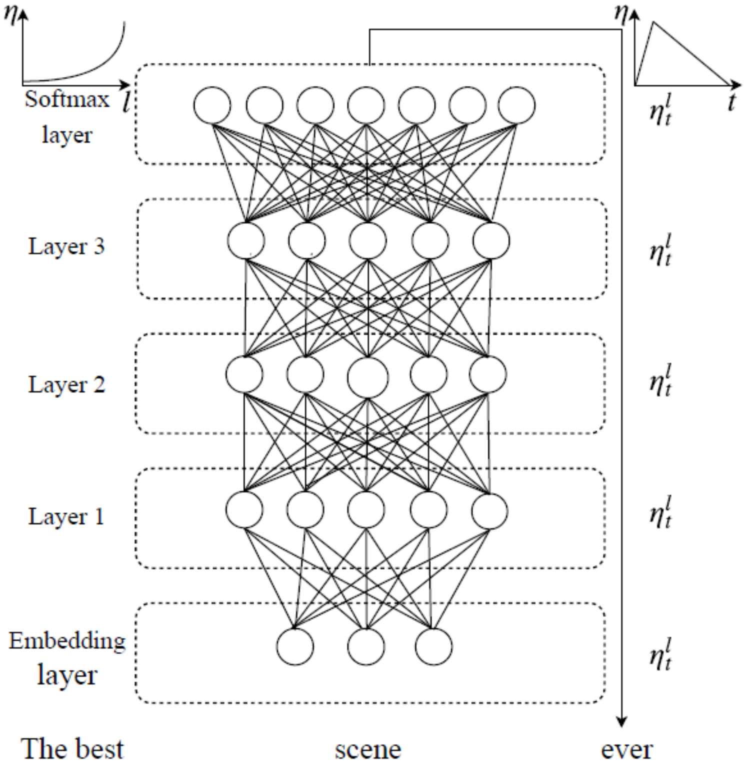
$$\theta_t^l = \theta_{t-1}^l - \eta^l \cdot \nabla_{\theta^l} J(\theta)$$

$$\eta^{l-1} = \eta^l / 2.6$$



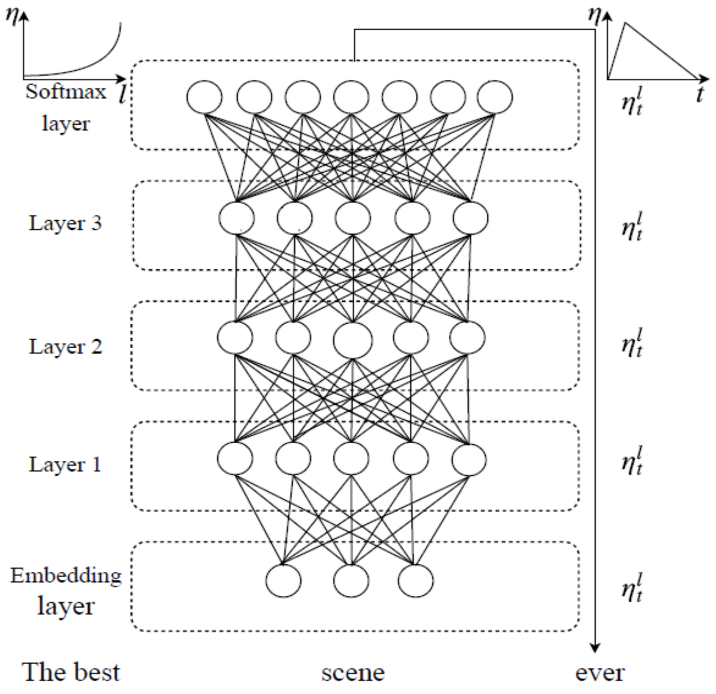
(b) LM fine-tuning

Catastrophic forgetting



(b) LM fine-tuning

(b)-2 Slanted triangular learning rates



(b) LM fine-tuning

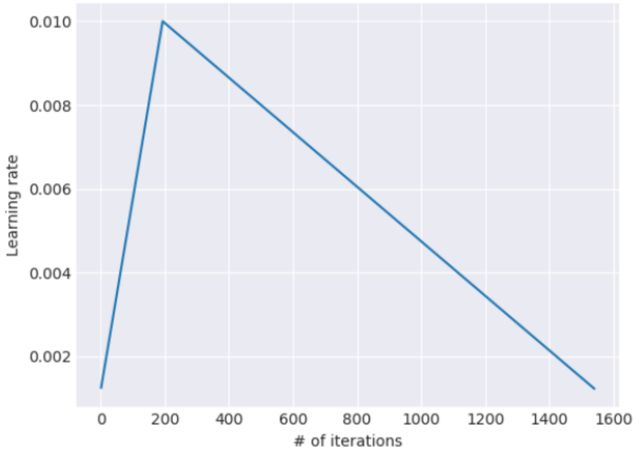
$$cut = \lfloor T \cdot cut\_frac \rfloor$$

$$p = \begin{cases} t/cut, & \text{if } t < cut \\ 1 - \frac{t-cut}{cut \cdot (1/cut\_frac - 1)}, & \text{otherwise} \end{cases}$$

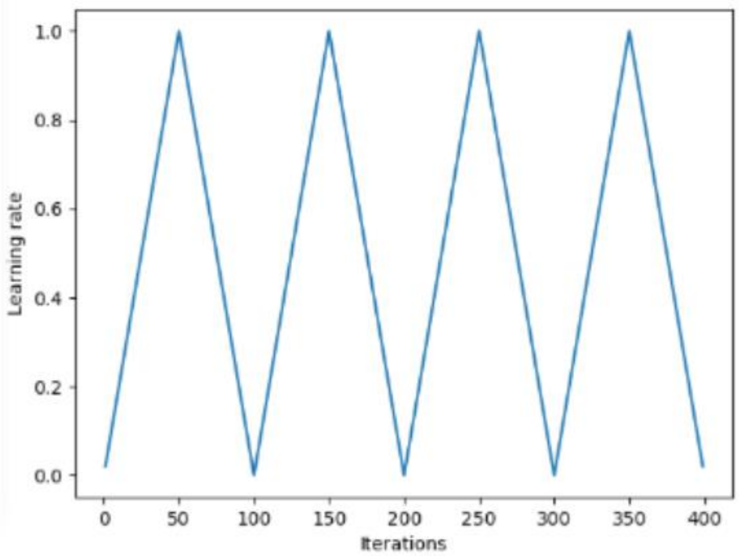
$$\eta_t = \eta_{max} \cdot \frac{1 + p \cdot (ratio - 1)}{ratio}$$

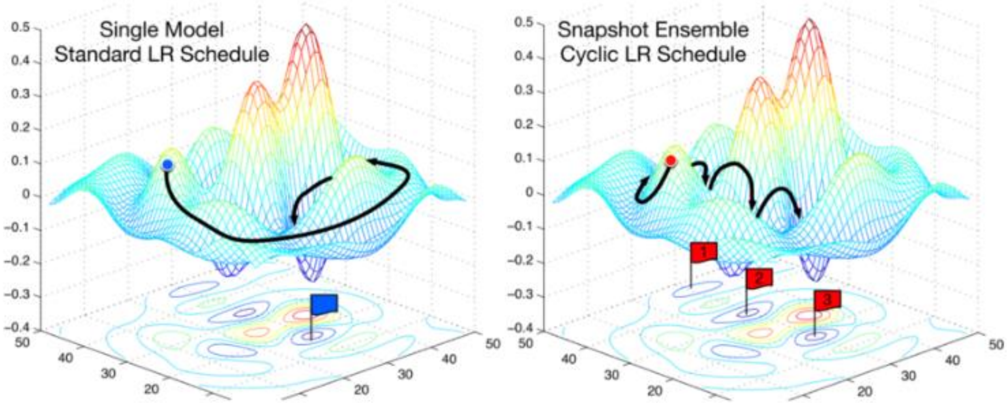
$$cut\_frac = 0.1, \quad \eta_{max} = 0.01$$

$$ratio = 32$$



cyclical triangular learning rate

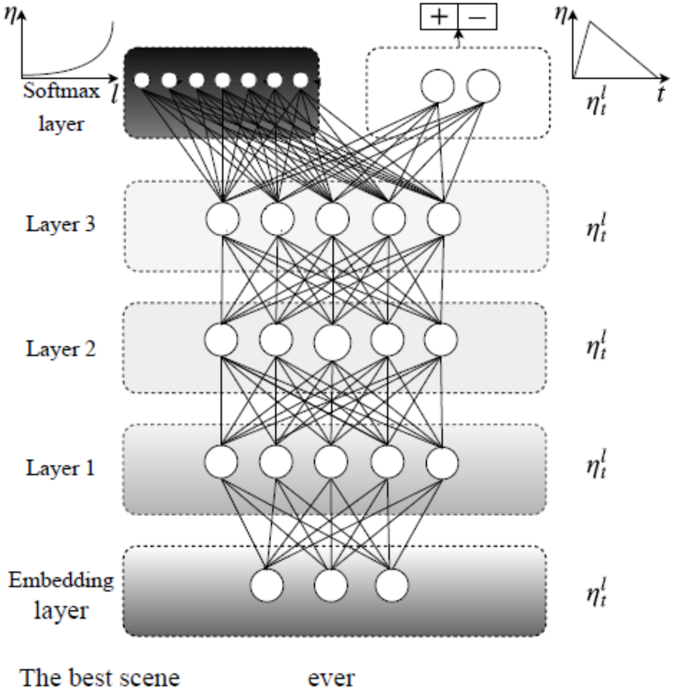




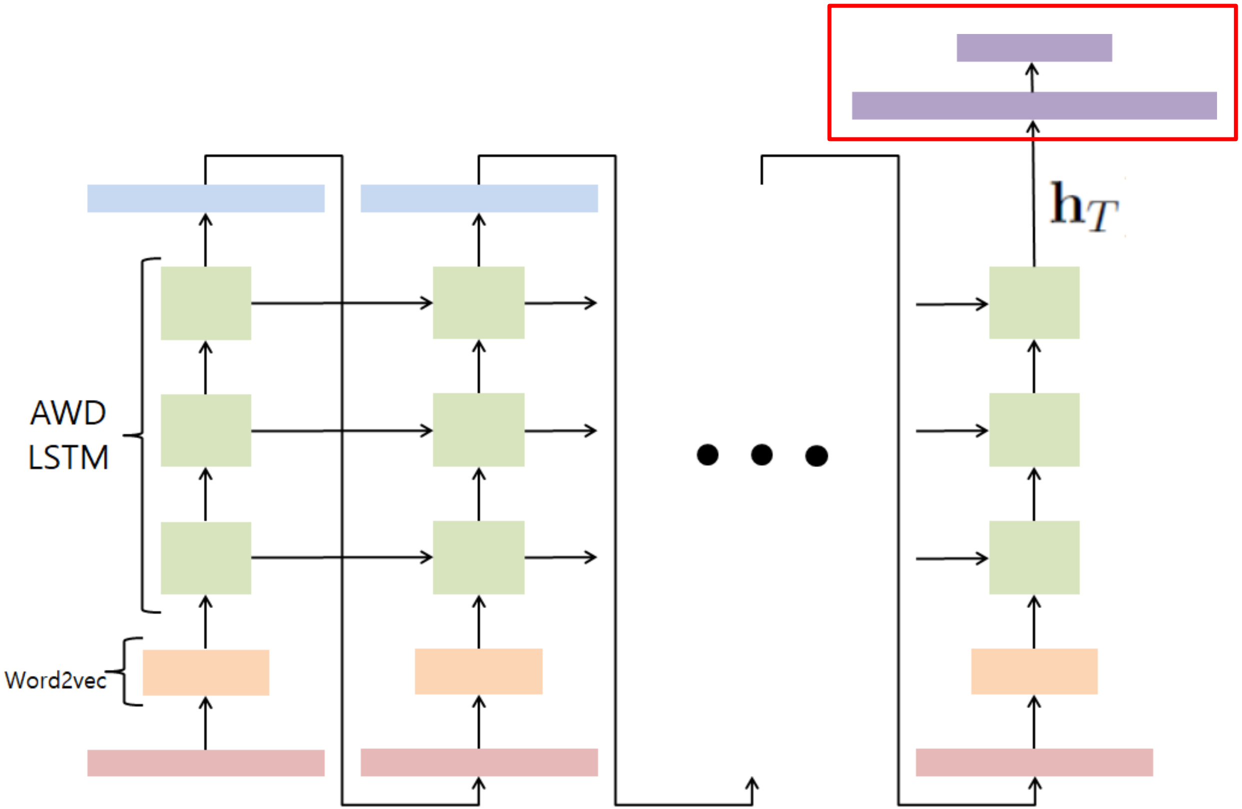
Text Classifier 추가 후 Fine-tuning

(c) Target task classifier fine-tuning

- Concat pooling
- Gradual unfreezing

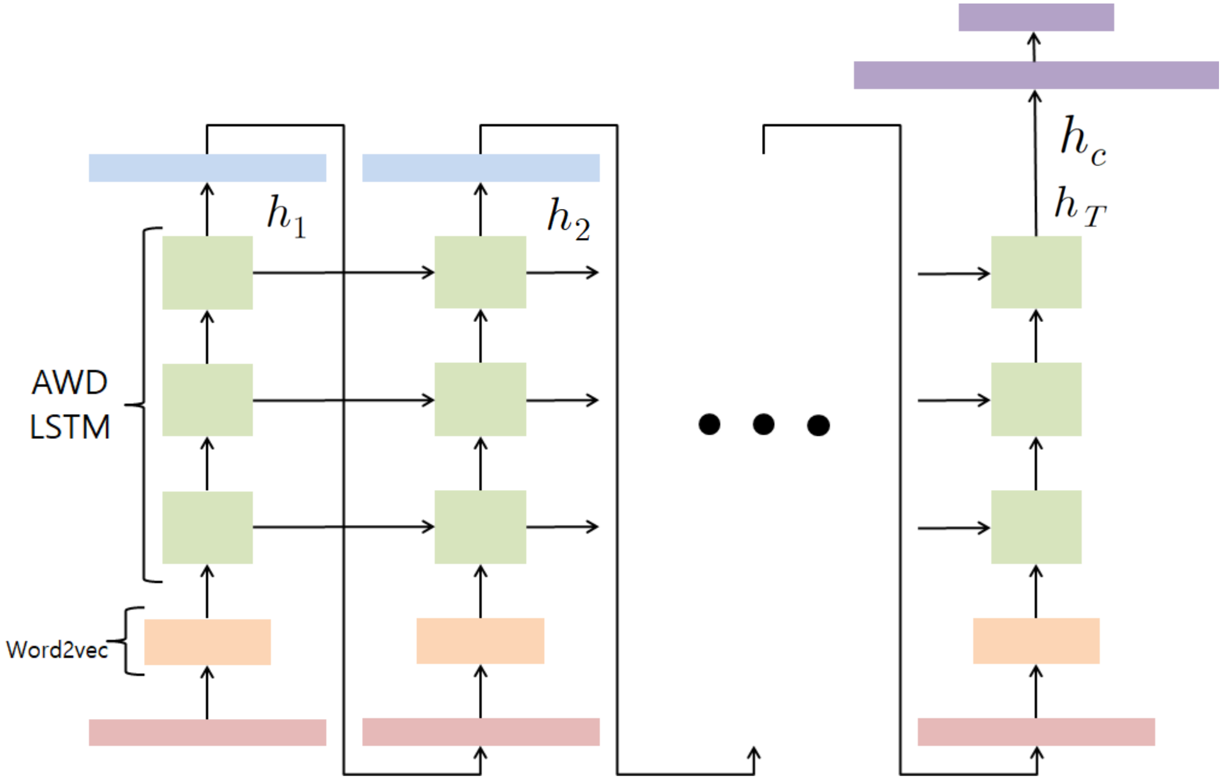


(c) Classifier fine-tuning



$\mathbf{h}_T$
AWD
LSTM
Word2vec

(c)-1 Concat pooling



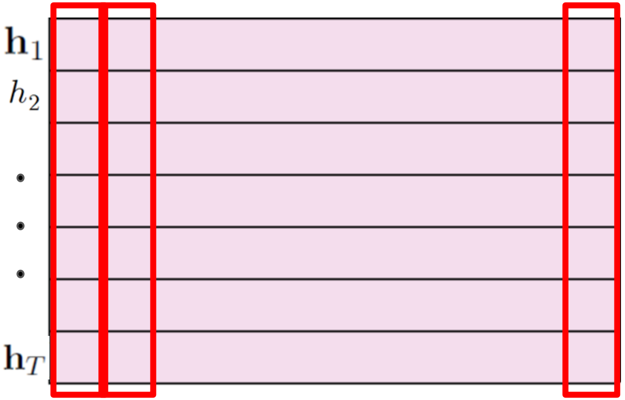
$$\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$$

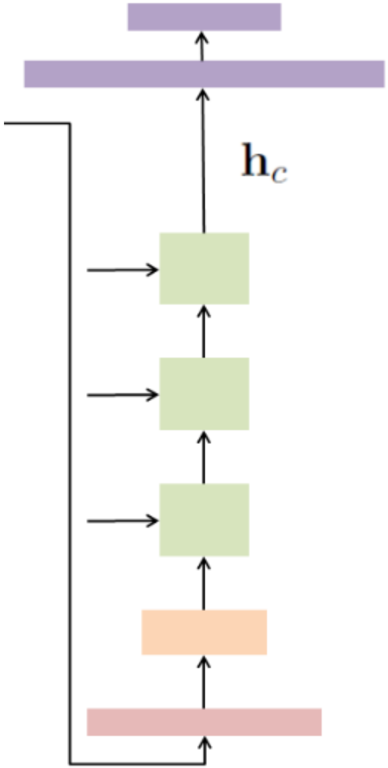
$$\mathbf{h}_c = [\mathbf{h}_T, \mathtt{maxpool}(\mathbf{H}), \mathtt{meanpool}(\mathbf{H})]$$

(c)-1 Concat pooling

$$\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$$

$$\mathbf{h}_c = [\mathbf{h}_T, \mathtt{maxpool}(\mathbf{H}), \mathtt{meanpool}(\mathbf{H})]$$



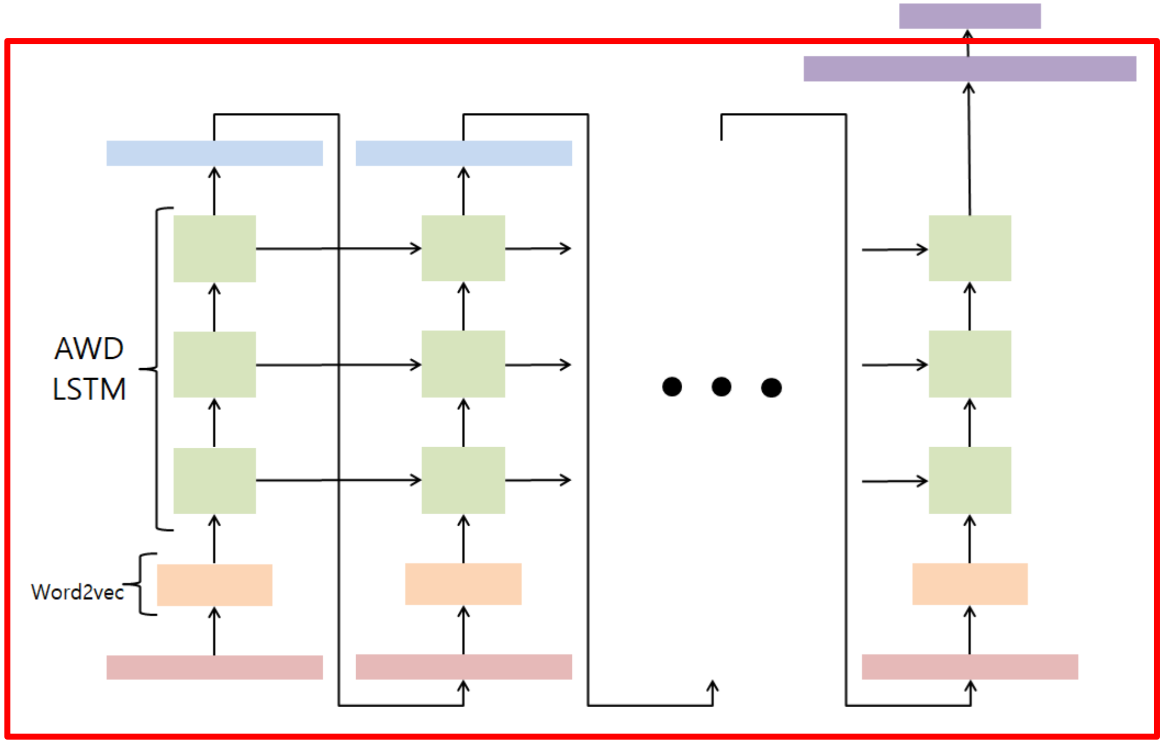


(c)-2 Gradual unfreezing

- Last layer에 least general knowledge 를 학습시키기 위한 작업이다.
- 각 layer가 수렴할 때까지 진행한다.

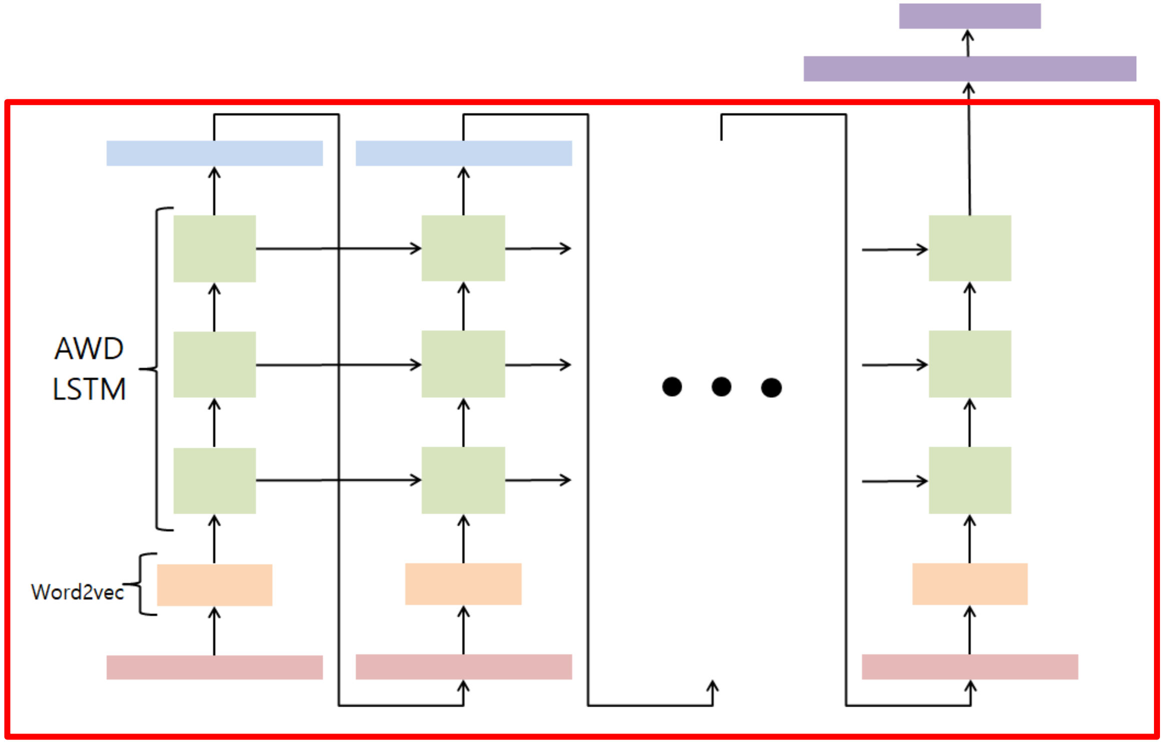
(c)-2 Gradual unfreezing

epoch
1



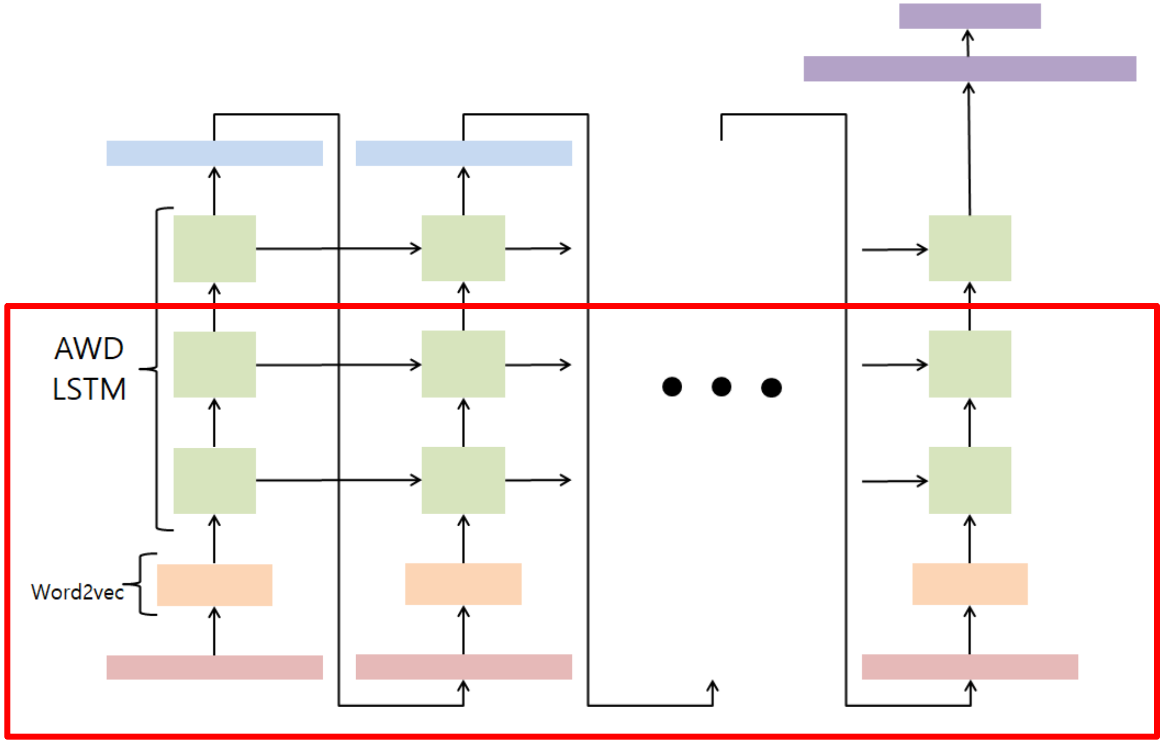
(c)-2 Gradual unfreezing

epoch
1



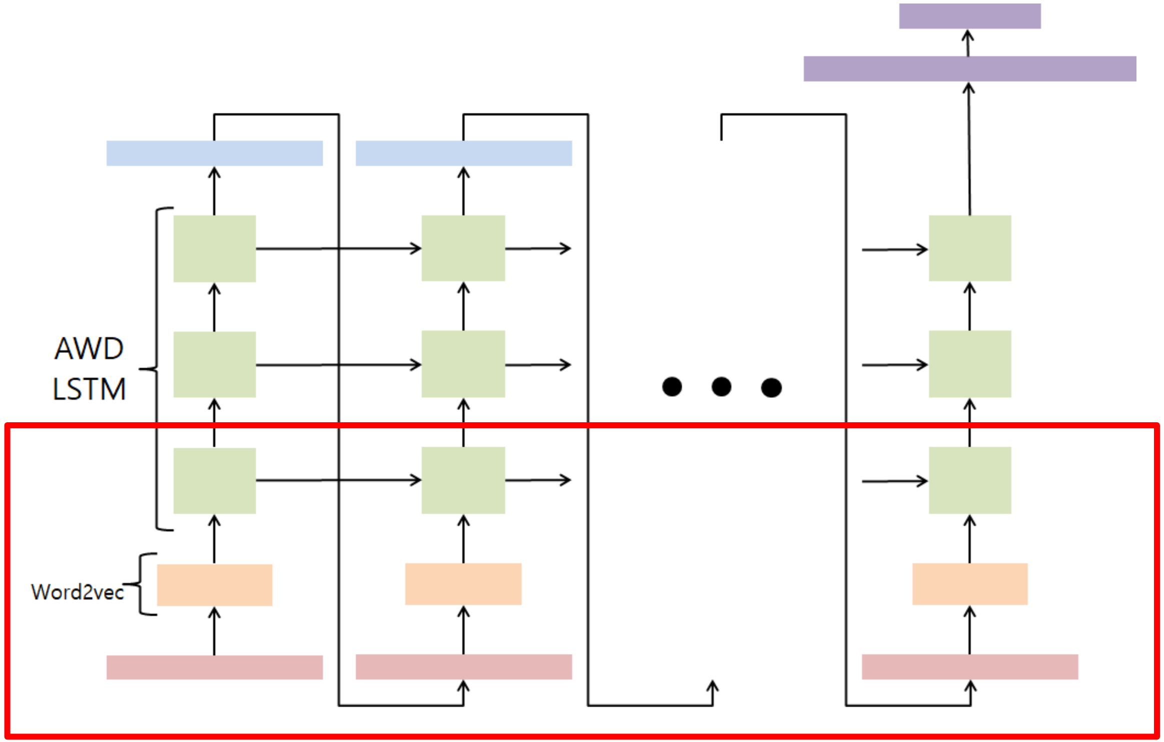
(c)-2 Gradual unfreezing

epoch
1

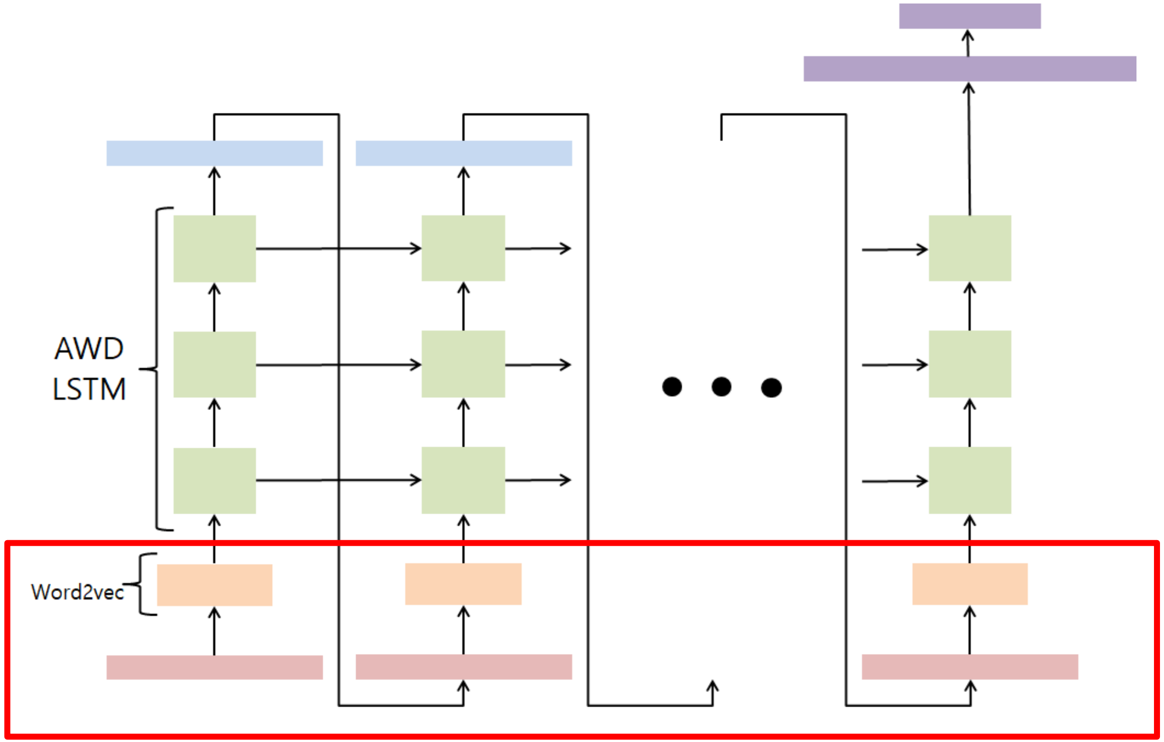


(c)-2 Gradual unfreezing

epoch
1

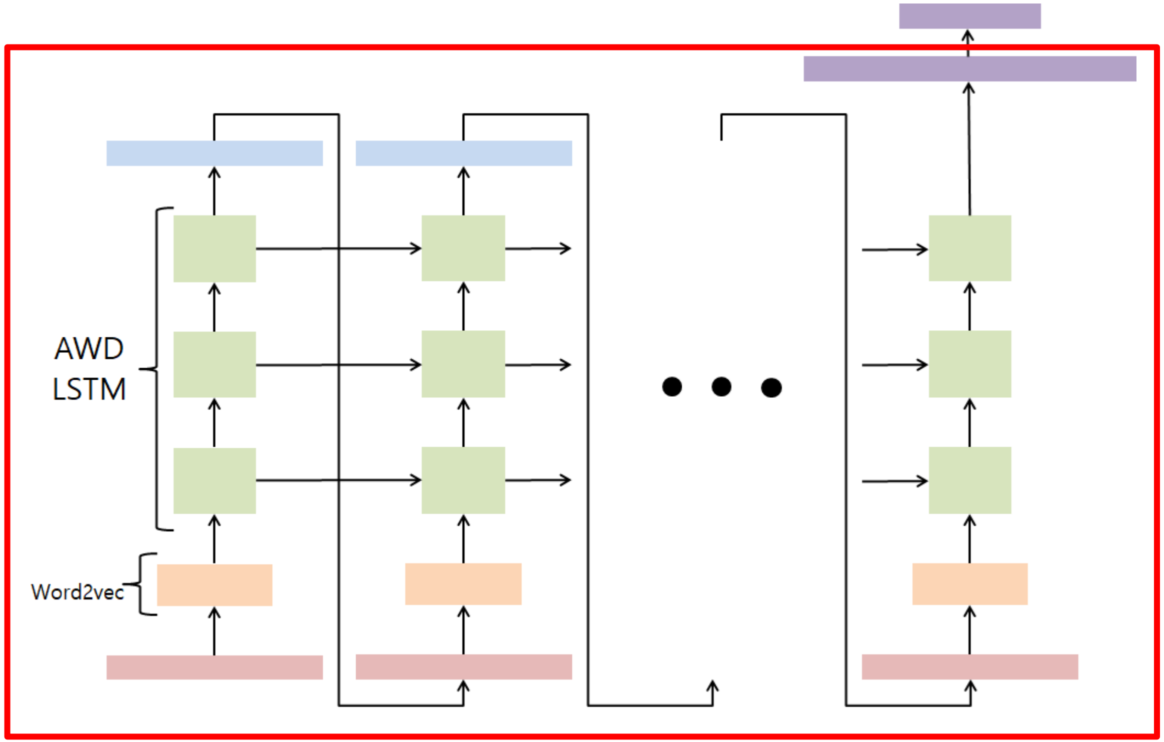


(c)-2 Gradual unfreezing



epoch
1

catastrophic forgetting



Result

	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**

	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**

Result

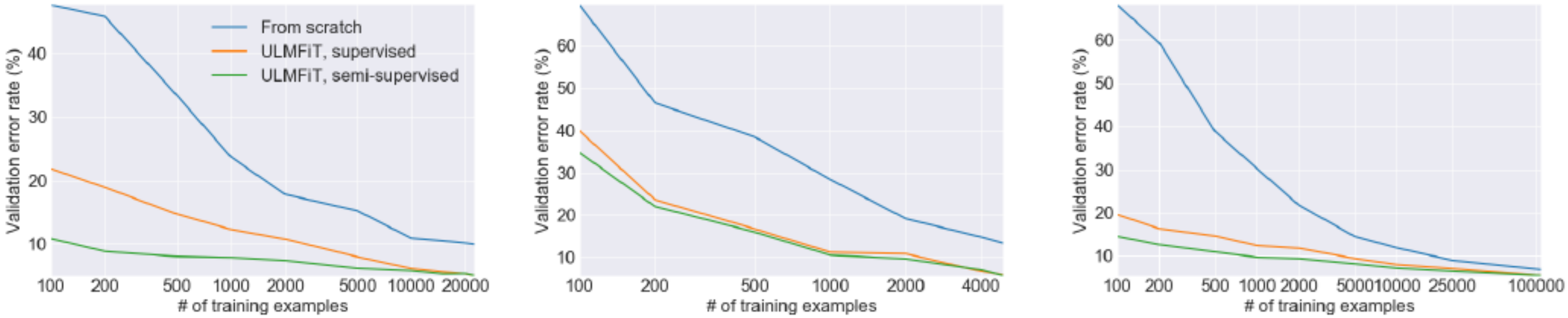


Figure 3: Validation error rates for supervised and semi-supervised ULMFiT vs. training from scratch with different numbers of training examples on IMDb, TREC-6, and AG (from left to right).

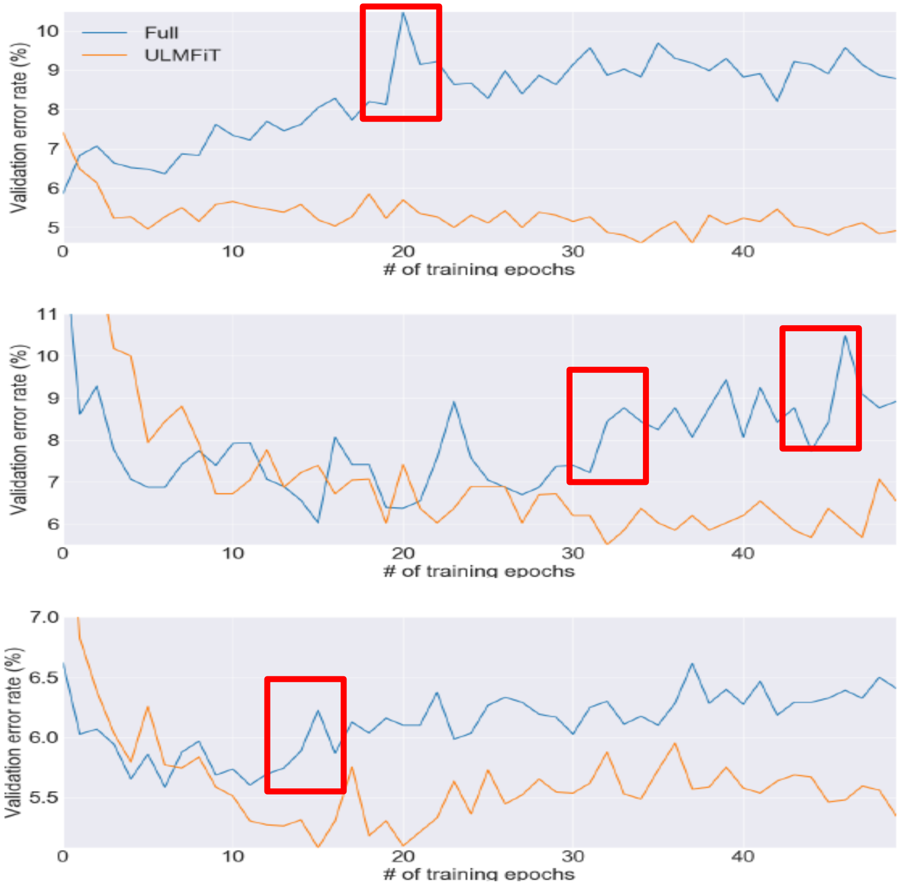
ablation

LM	IMDb	TREC-6	AG
Vanilla LM	5.98	7.41	5.76
AWD-LSTM LM	**5.00**	**5.69**	**5.38**

LM fine-tuning	IMDb	TREC-6	AG
No LM fine-tuning	6.99	6.38	6.09
Full	5.86	6.54	5.61
Full + discr	5.55	6.36	5.47
Full + discr + stlr	**5.00**	**5.69**	**5.38**

Classifier fine-tuning	IMDb	TREC-6	AG
From scratch	9.93	13.36	6.81
Full	6.87	6.86	5.81
Full + discr	5.57	6.21	5.62
Last	6.49	16.09	8.38
Chain-thaw	5.39	6.71	5.90
Freez	6.37	6.86	5.81
Freez + discr	5.39	5.86	6.04
Freez + stlr	5.04	6.02	5.35
Freez + cos	5.70	6.38	**5.29**
Freez + discr + stlr	**5.00**	**5.69**	5.38

Result


Full
ULMFiT
Validation error rate (%)
# of training epochs
Validation error rate (%)
# of training epochs
Validation error rate (%)
# of training epochs