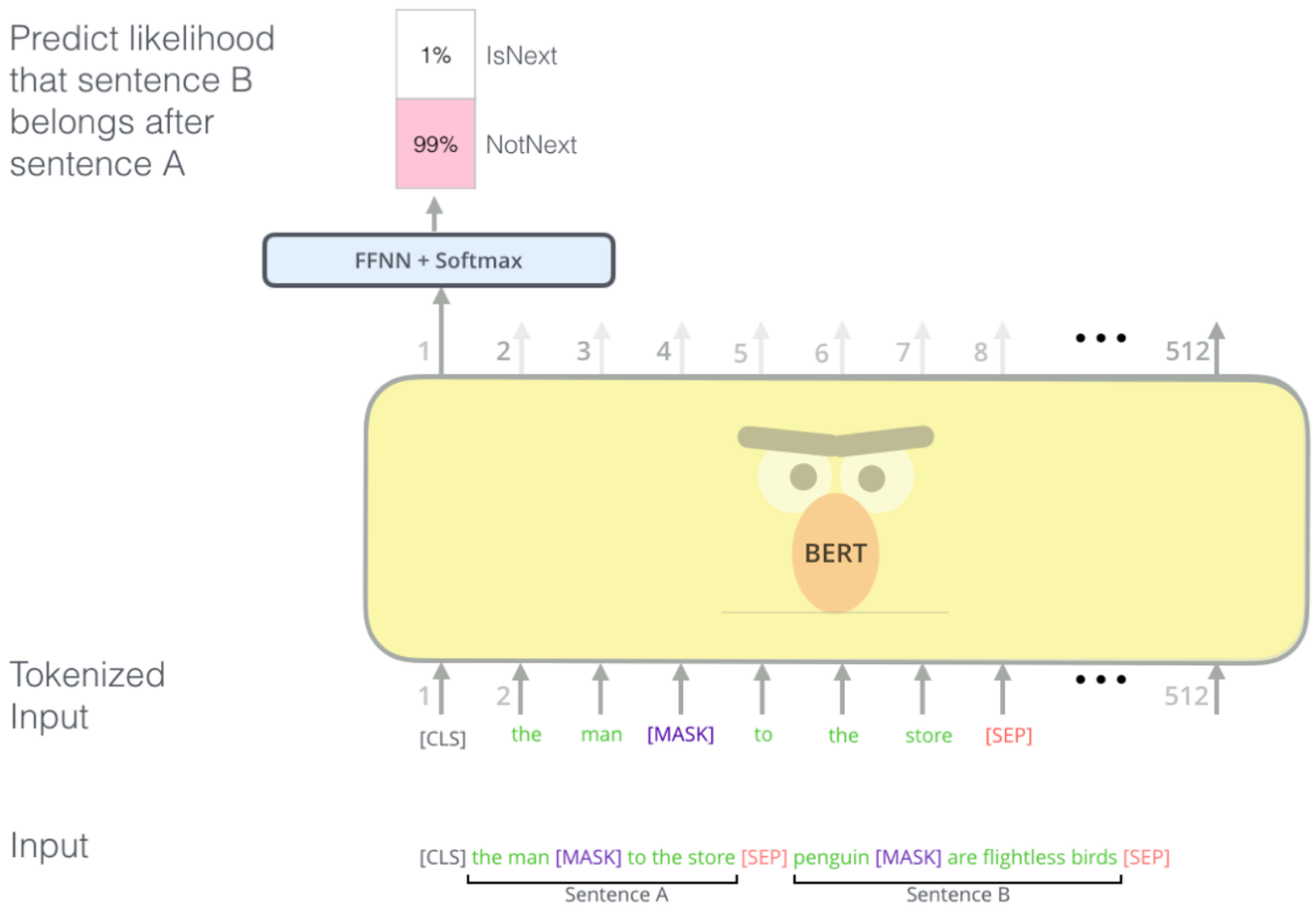
COS 598C Advanced Topics in Computer Science: Deep Learning for Natural Language Processing

Pre-training and Fine-tuning

Winter 2020

A note on next sentence prediction

- NSP could hurt the MLM training objective.
- **Recommended reading**: SpanBERT: Improving Pre-training by Representing and Predicting Spans



Fine-tuning vs Feature-based

- **Recommended reading**: To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks

Pretraining	Adaptation	NER	SA	Nat. lang. inference		Semantic textual similarity		
		CoNLL 2003	SST-2	MNLI	SICK-E	SICK-R	MRPC	STS-B
Skip-thoughts	❄	-	81.8	62.9	-	86.6	75.8	71.8
ELMo	❄	91.7	**91.8**	**79.6**	**86.3**	**86.1**	**76.0**	**75.9**
	🔥	**91.9**	91.2	76.4	83.3	83.3	74.7	75.5
	Δ=🔥-❄	0.2	-0.6	-3.2	-3.3	-2.8	-1.3	-0.4
BERT-base	❄	92.2	93.0	**84.6**	84.8	86.4	78.1	82.9
	🔥	**92.4**	**93.5**	**84.6**	**85.8**	**88.7**	**84.8**	**87.1**
	Δ=🔥-❄	0.2	0.5	0.0	1.0	2.3	6.7	4.2

Fine-tuning vs Feature-based

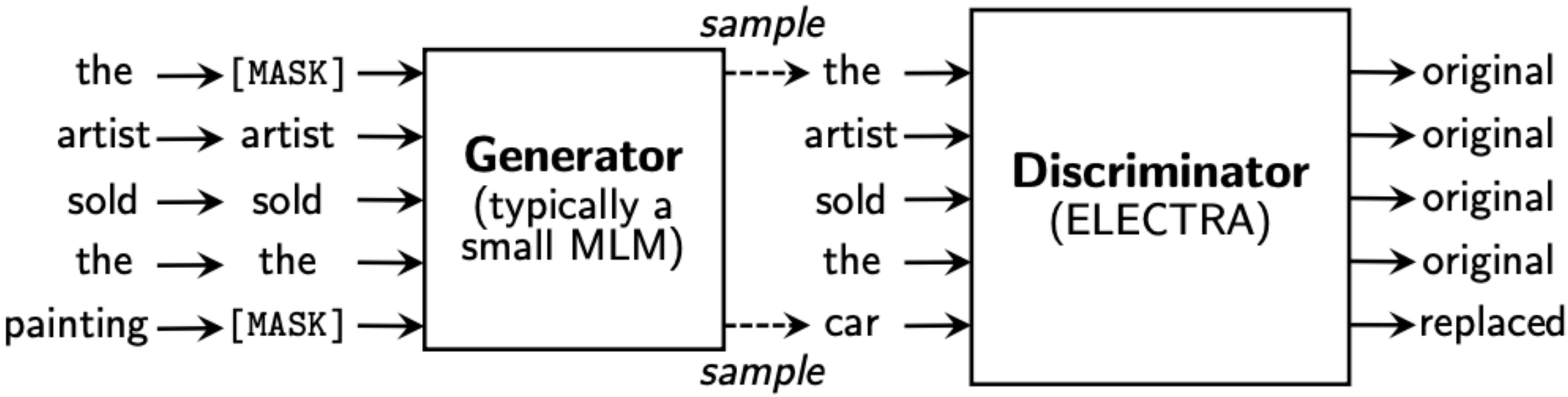
- **Recommended reading**: Incorporating BERT into Neural Machine Translation

Table 1: Preliminary explorations on IWSLT'14 English→German translation.

Algorithm	BLEU score
Standard Transformer	28.57
Use BERT to initialize the encoder of NMT	27.14
Use XLM to initialize the encoder of NMT	28.22
Use XLM to initialize the decoder of NMT	26.13
Use XLM to initialize both the encoder and decoder of NMT	28.99
Leveraging the output of BERT as embeddings	29.67

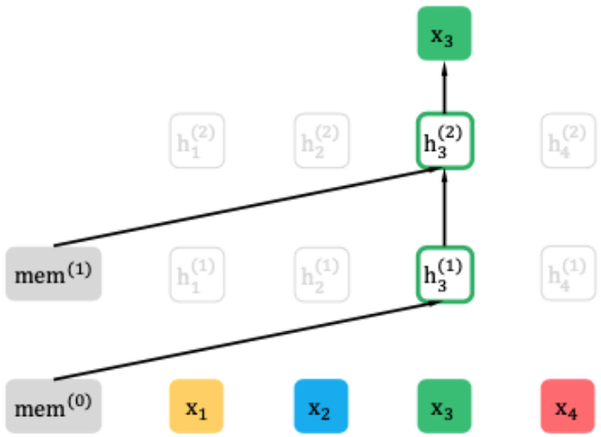
How to fix the 15% masking?

- **Recommended reading**: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators

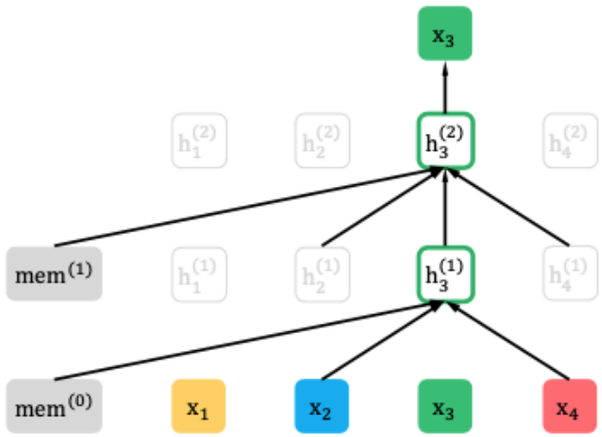


How to fix the 15% masking?

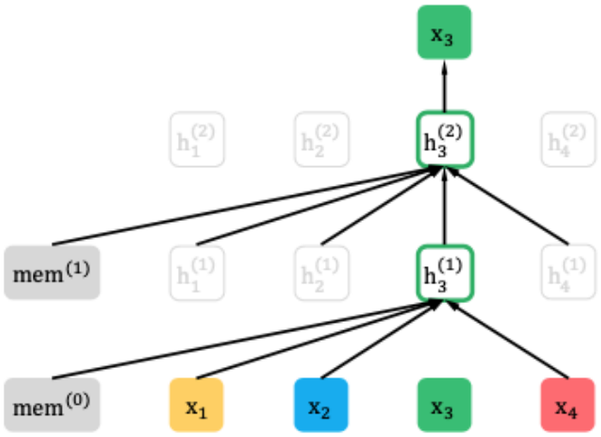
- **Next lecture:** XLNet: Generalized Autoregressive Pretraining for Language Understanding



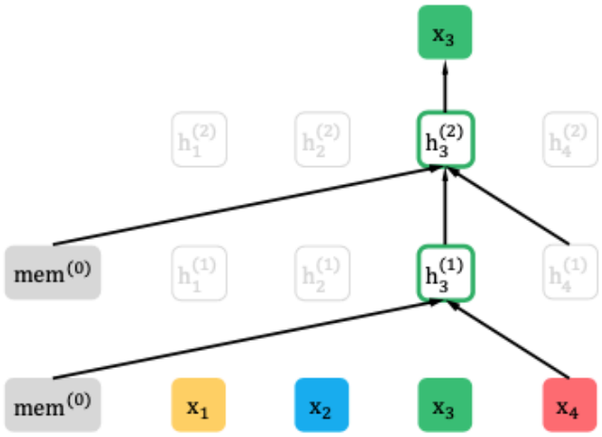
Factorization order: 3 → 2 → 4 → 1



Factorization order: 2 → 4 → 3 → 1



Factorization order: 1 → 4 → 2 → 3



Factorization order: 4 → 3 → 1 → 2