Self-Supervised Learning in NLP

Minlie Huang

Tsinghua University

aihuang@tsinghua.edu.cn

Why SSL?

- Yann Lecun:
- Human and animal babies learning by observations

The Future is Self-Supervised

Yann LeCun
NYU - Courant Institute & Center for Data Science
Facebook AI Research

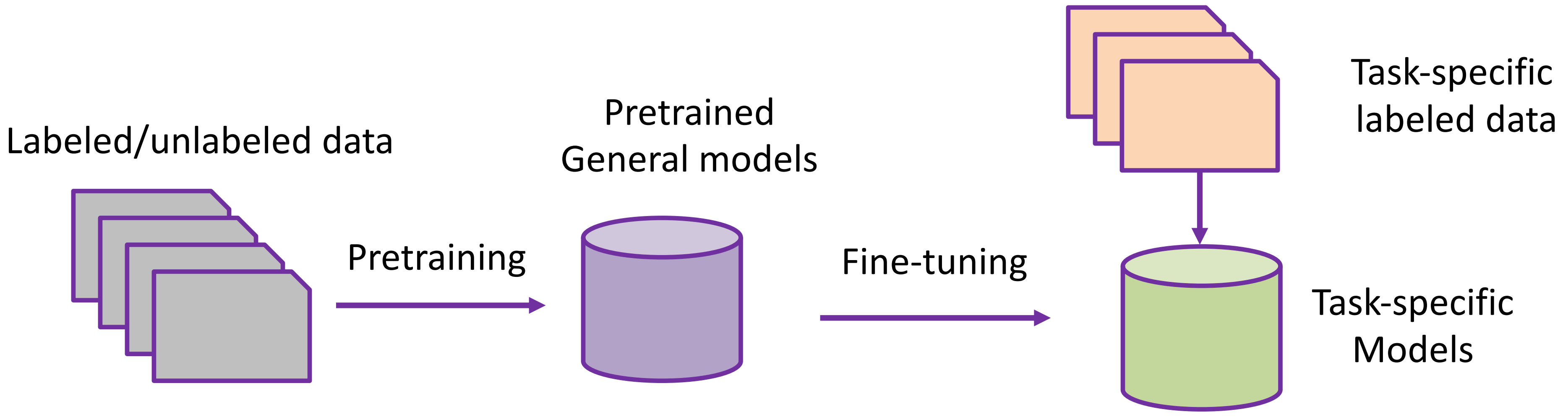
Learning Paradigms

- **Unsupervised learning**: P(X)
  - Autoencoder, VAE, Boltzmann Machine
- **Supervised learning**: P(y|X)
  - SVM, NB, DT, MLP, CNN, RNN
- **Semi-supervised learning**: labeled data + unlabeled data
  - Self-training
  - Self-supervised learning (sometimes)

Key Concepts

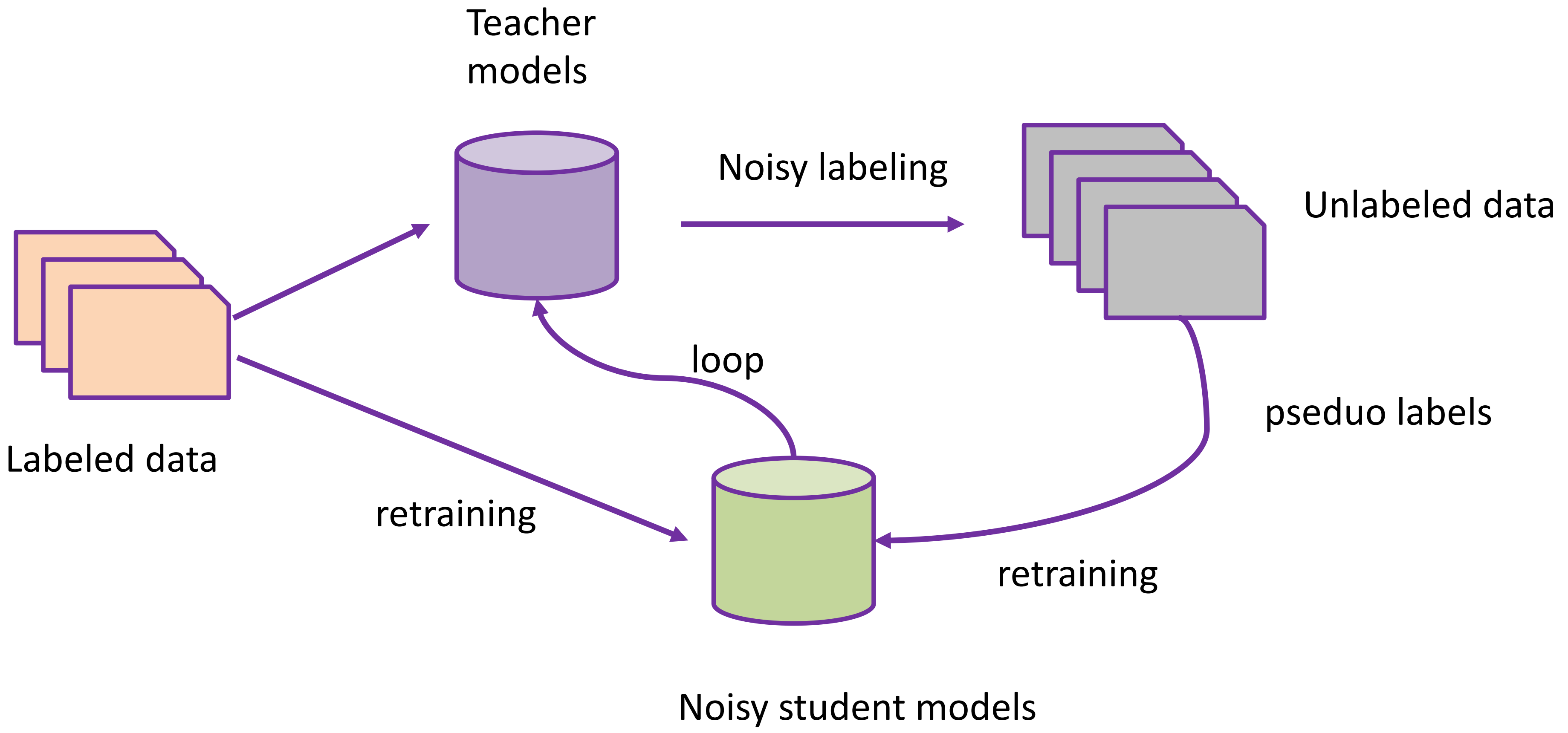
- Pretraining
- Self-training
- Self-supervised learning

Pretraining + Fine-tuning



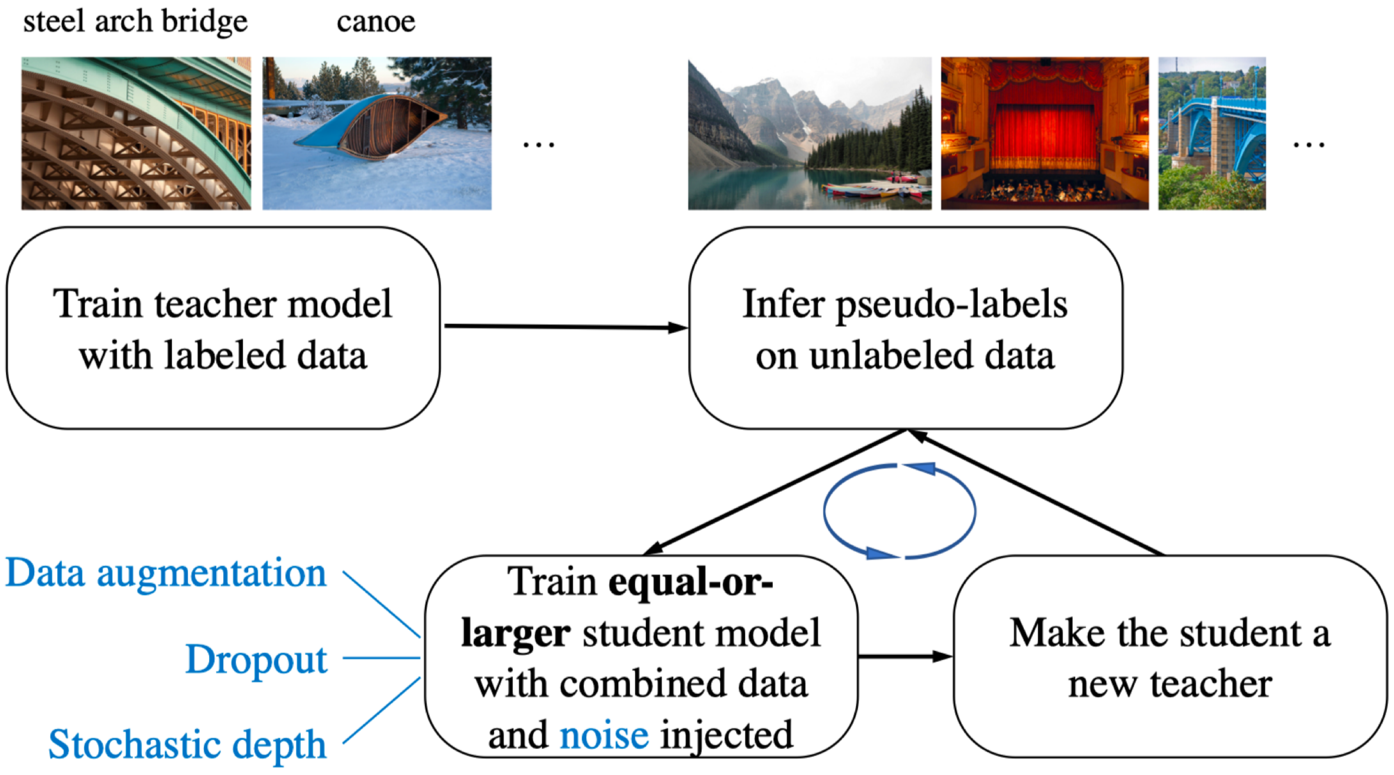
- Pretrained on ImageNet (labeled), fine-tuned on image classification, detection, segmentation
- Pretrained on large text corpora (unlabeled), fine-tuned for NLU tasks: BERT, GPT, etc.

Self-training



Self-training with Noisy Student

- Adding noise to the student (augmentation, dropout, stochastic depth)
- Using student models that are not smaller than the teacher



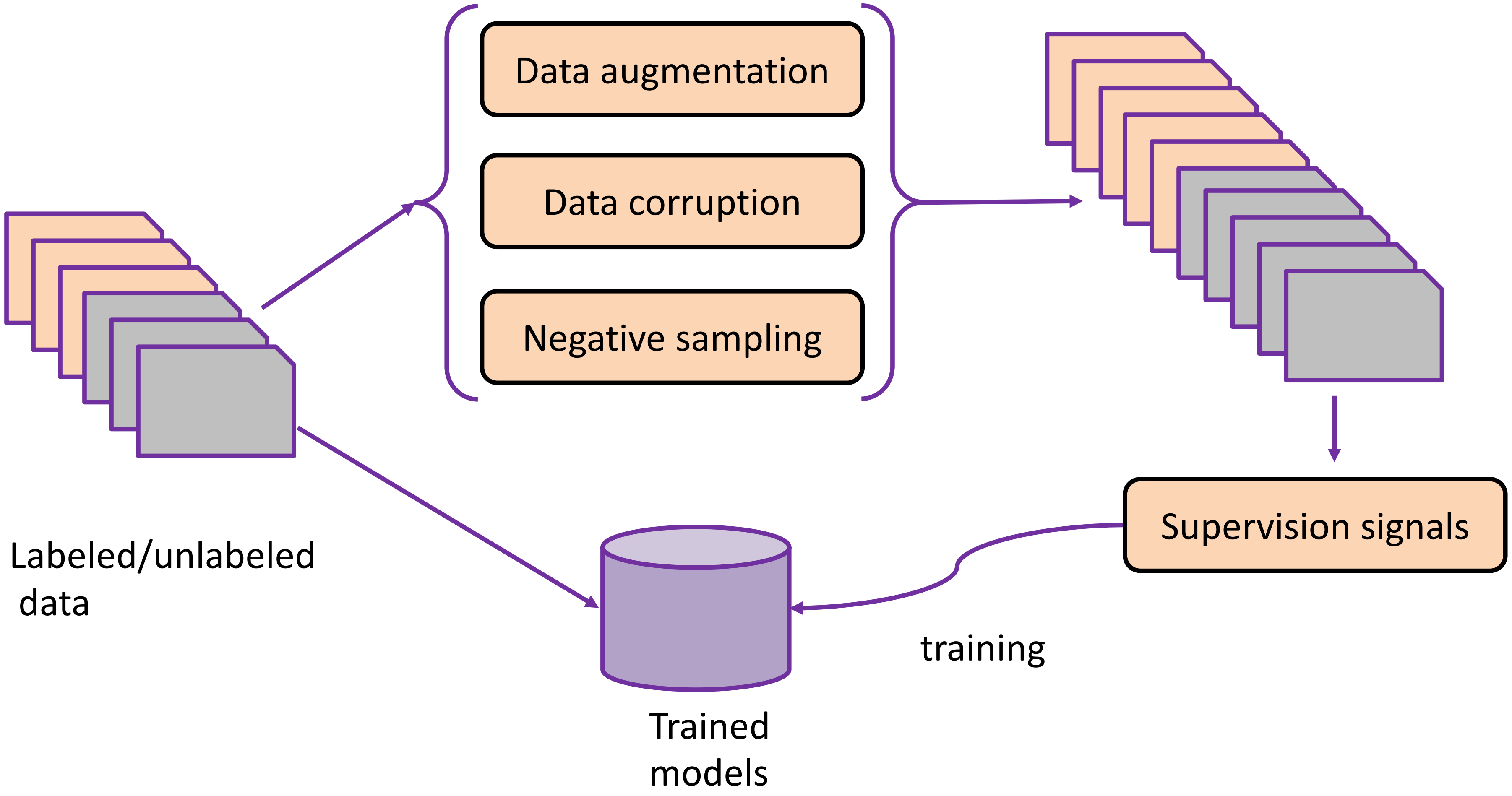
$$\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f^{noised}(x_i,\theta^t))$$

$$\tilde{y}_i = f(\tilde{x}_i,\theta_*^t), \forall i = 1,\cdots,m$$

$$\frac{1}{n}\sum_{i=1}^{n}\ell(y_i, f^{noised}(x_i,\theta^s)) + \frac{1}{m}\sum_{i=1}^{m}\ell(\tilde{y}_i, f^{noised}(\tilde{x}_i,\theta^s))$$

Self-training with Noisy Student improves ImageNet classification.

Self-Supervised Learning (SSL)



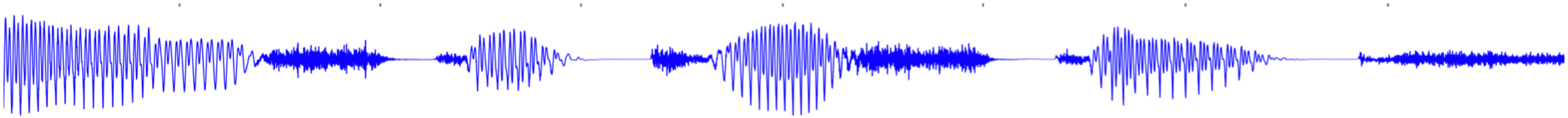
SSL: predicting future from past

Sequence/stream data

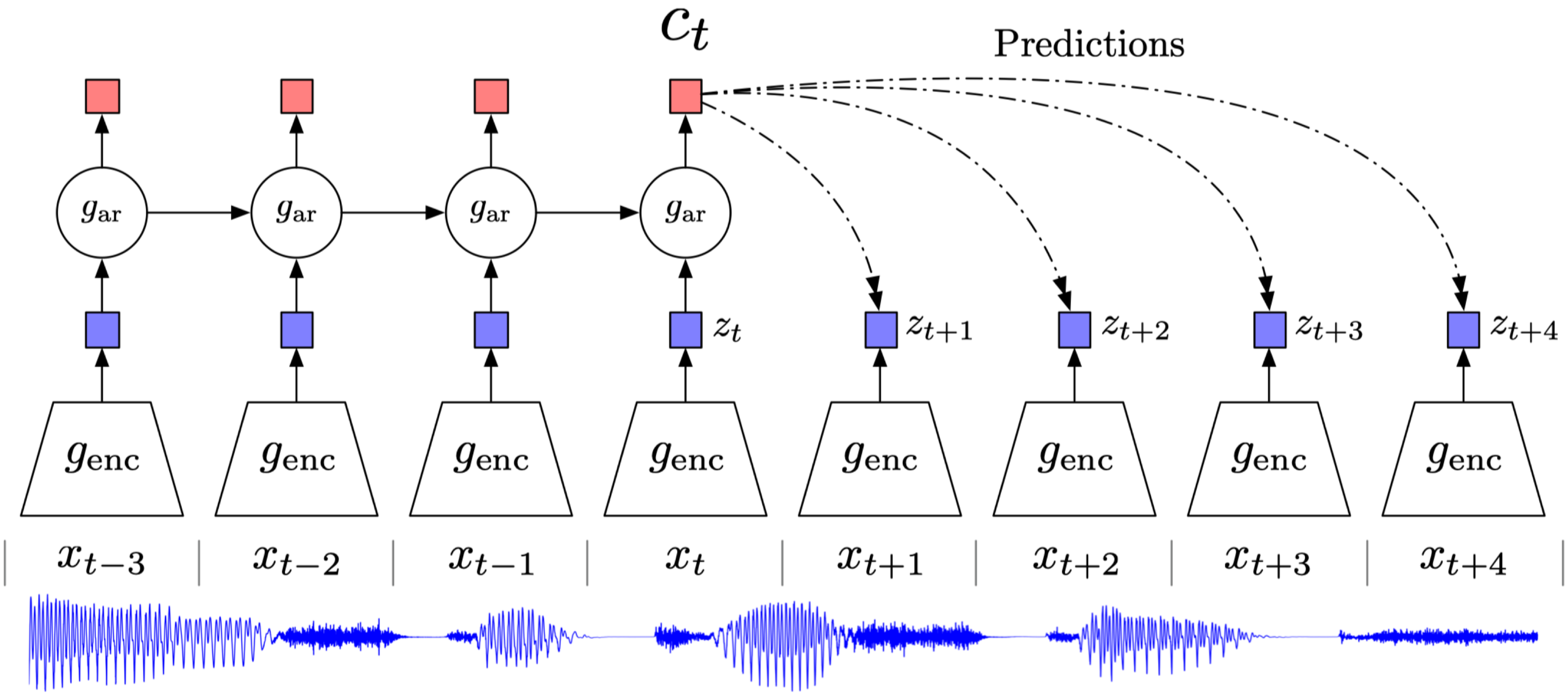
Context (human written): In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

Input frames

Ground truth

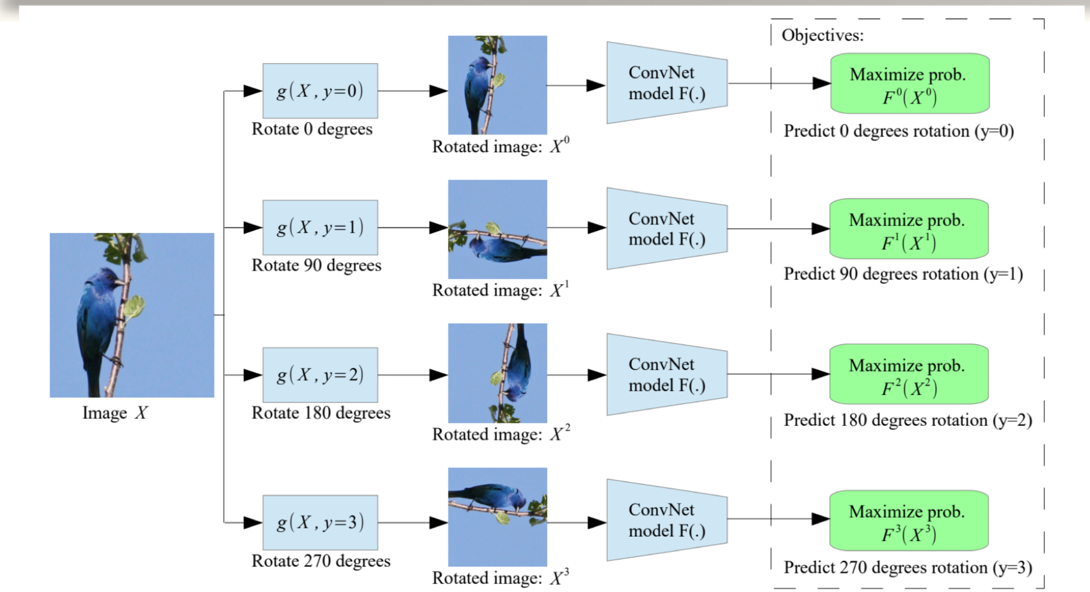


SSL: predicting future from past

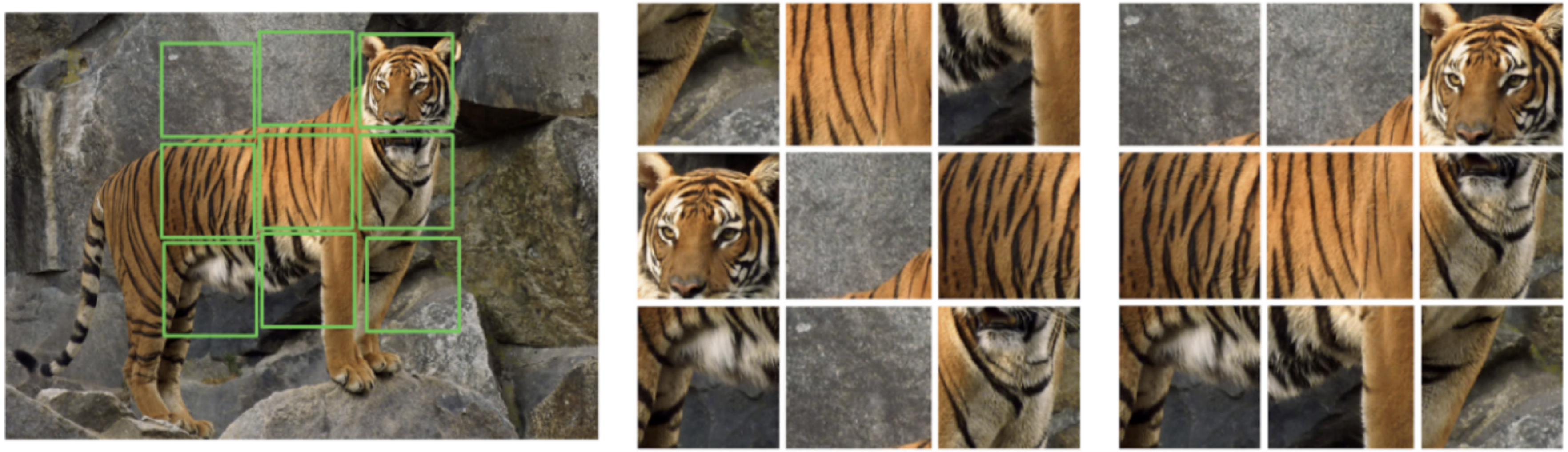


Representation Learning with Contrastive Predictive Coding. Van den Oord et al. 2018

SSL: recovering from corruption/perturbation



(g) Cutout (h) Gaussian noise (i) Gaussian blur



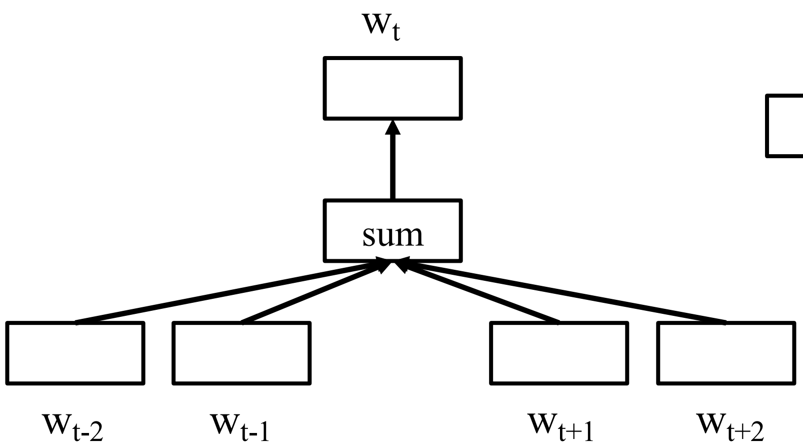
SSL in natural language processing

- **Self-supervised Learning**
  - Generative
    - Auto-encoding (AE): **Masked language modeling** from BERT-style models
    - Auto-regressive (AR): **Language modeling** from GPT-style models
    - AE+AR: **Permutation language modeling** from XLNet
  - Contrastive
    - Context-instance / Global-local contrast: **Text order prediction** from BERT-style models
    - Context-context / Global-global contrast: **Sentence distance prediction** from CONPONO
  - Generative-contrastive / Adversarial: **Replaced token detection** from ELECTRA

SSL in NLP: language modeling

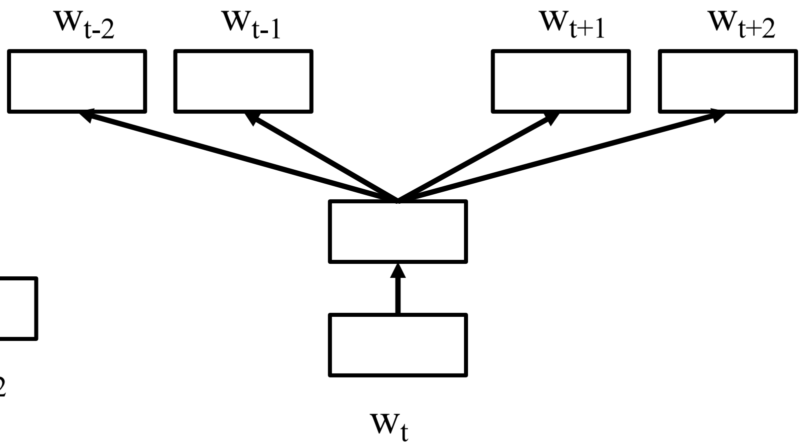
- Estimating P(**present**|**context**)

$Max\colon \log P(w_t|w_{t-2}w_{t-1}w_{t+1}w_{t+2})$



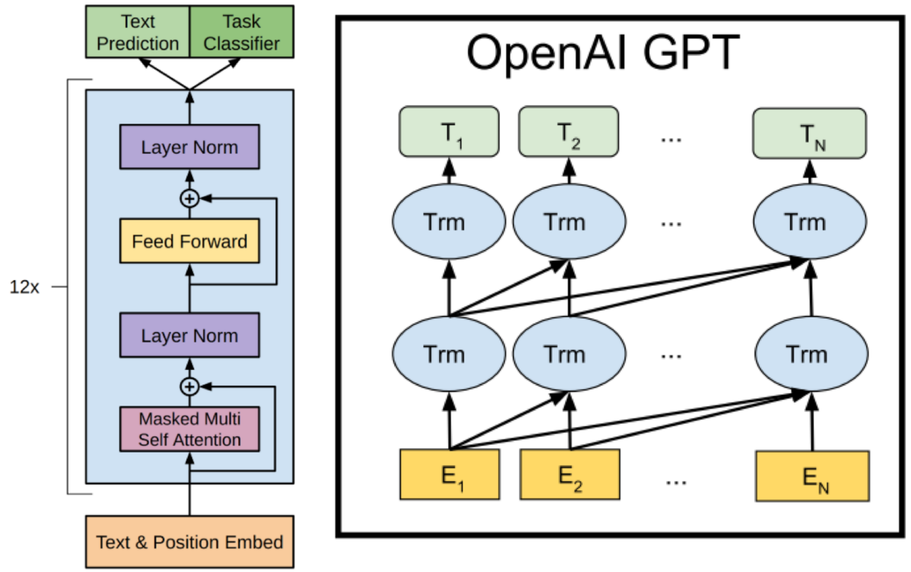
CBOW

$Max\colon \sum_j \log P(w_{t+j}|w_t)$

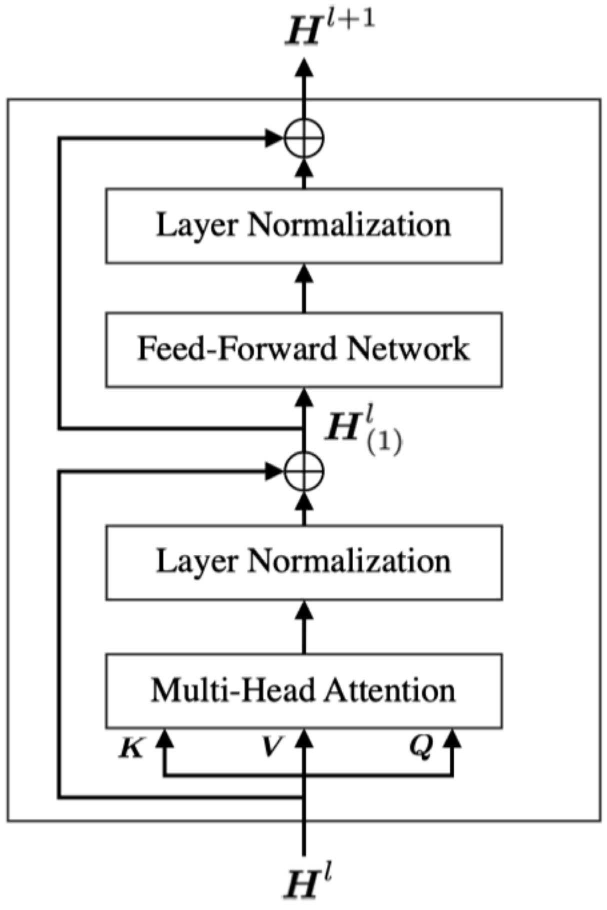


SkipGram

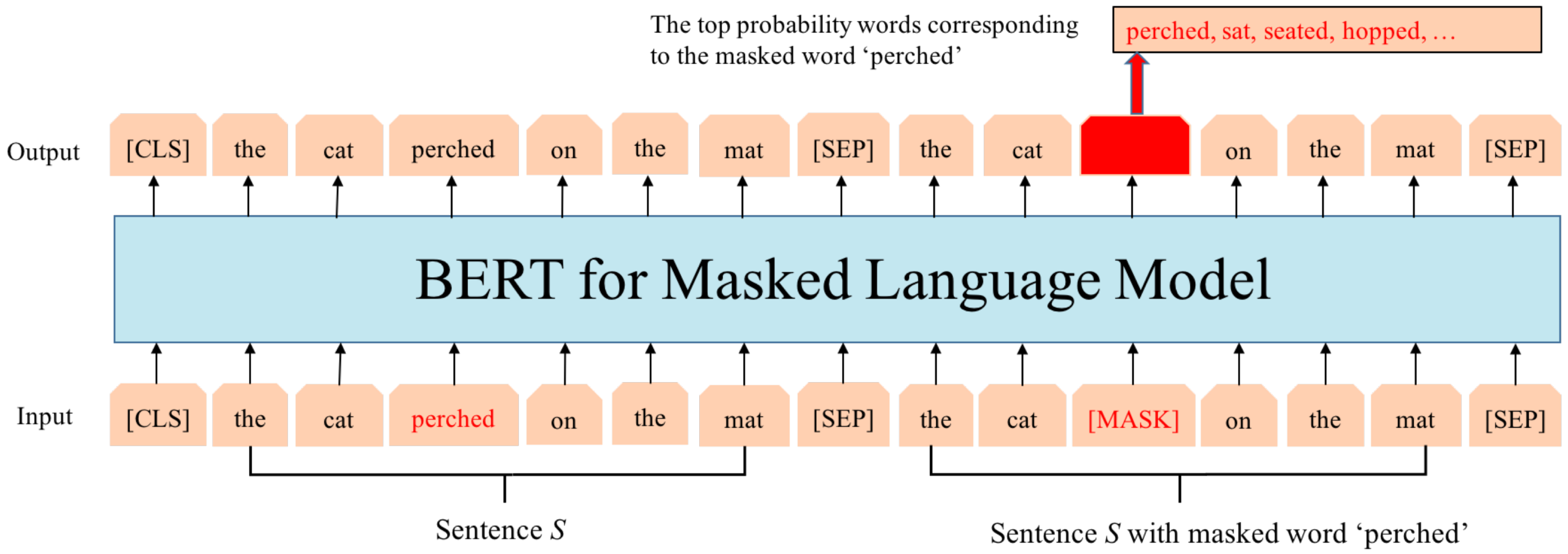
$Max\colon \sum_t \log P(w_t|w_1 w_2 \ldots w_{t-1})$



SSL in NLP: masked language modeling



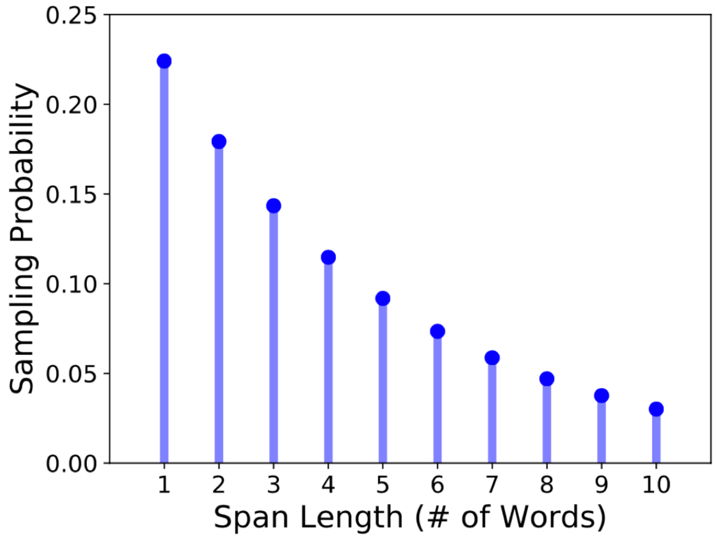
Recovering masked words in the input text

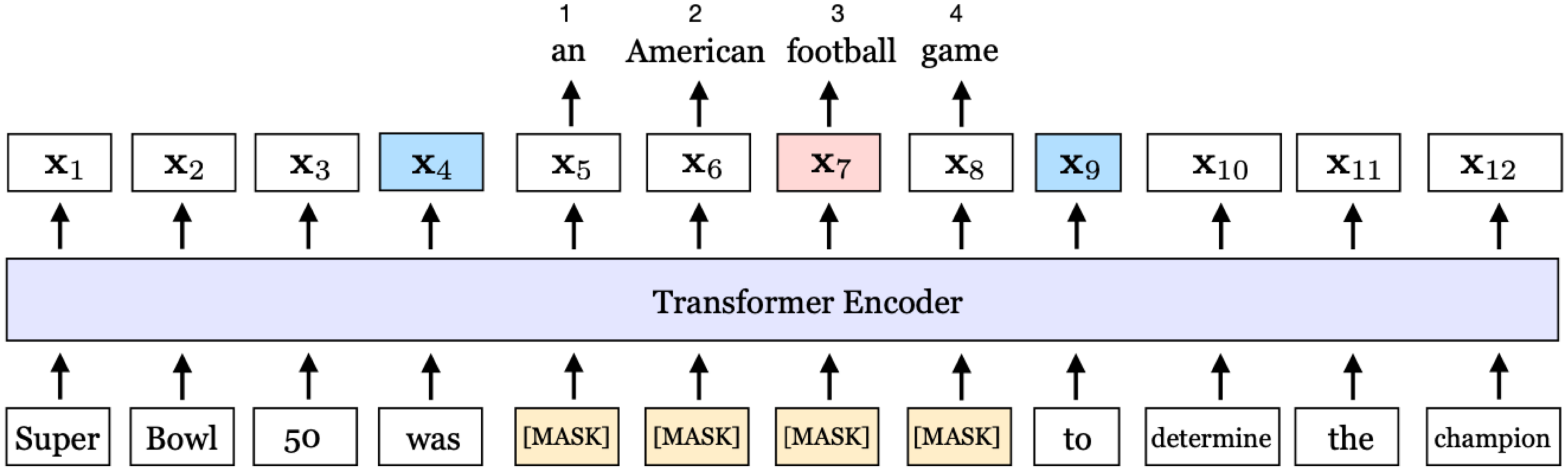


SSL in NLP: masked language modeling

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$

Span length distribution

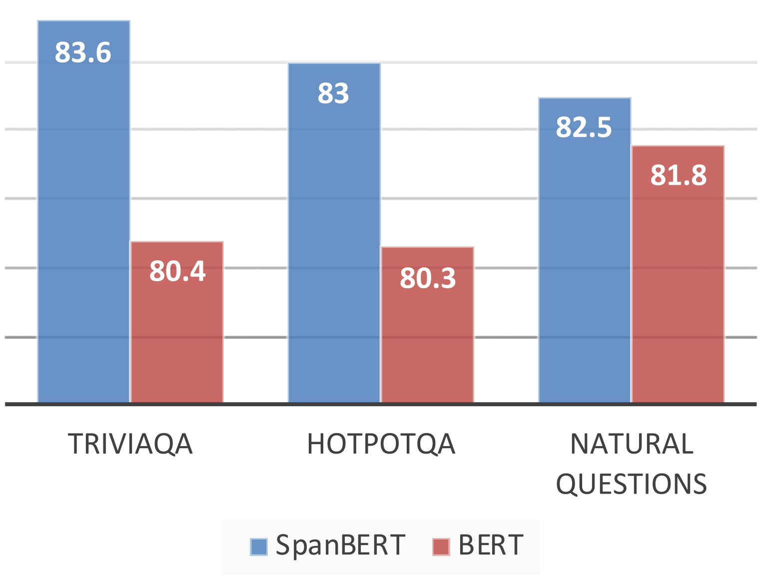




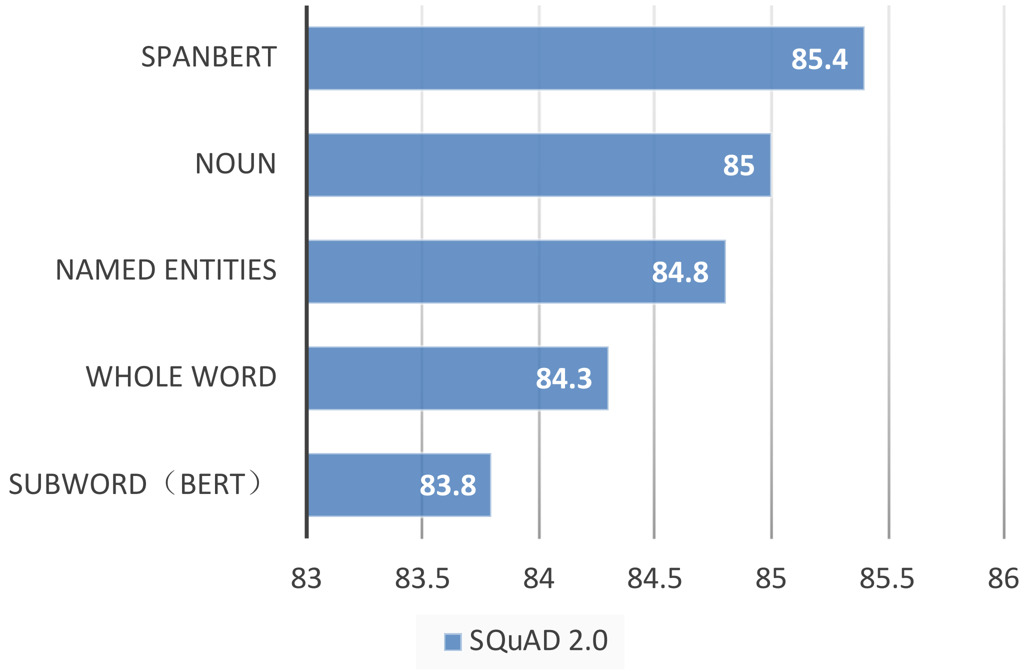
Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, Omer Levy: SpanBERT: Improving Pre-training by Representing and Predicting Spans. Trans. Assoc. Comput. Linguistics 8: 64-77 (2020).

SSL in NLP: masked language modeling

- Machine reading comprehension



- Masking strategies



SSL in NLP: permutation language modeling

- XLNet: Permutation Language Model

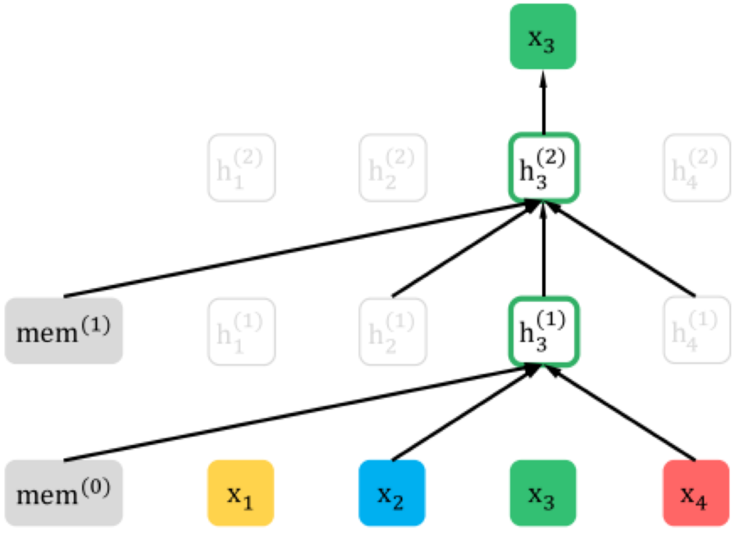
$$\max \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}_T}\left[\sum_{t=1}^{T}\log p_\theta(x_{z_t} \mid \mathbf{x}_{\mathbf{z}_{<t}})\right]$$

$$\mathcal{L}_{BERT} = -\,(\log P(deep|I\ like\ [\mathbf{MASK}]\ [\mathbf{MASK}]\ very\ much)$$
$$+\log P(learning|I\ like\ [\mathbf{MASK}]\ [\mathbf{MASK}]\ very\ much))$$

$$\mathcal{L}_{XLNet} = -\,(\log P(learning|I\ like\ [\mathbf{MASK}]\ [\mathbf{MASK}]\ very\ much)$$
$$+\log P(deep|I\ like\ [\mathbf{MASK}]\ \mathbf{learning}\ very\ much))$$
$$-\,(\log P(deep|I\ like\ [\mathbf{MASK}]\ [\mathbf{MASK}]\ very\ much)$$
$$+\log P(learning|I\ like\ \mathbf{deep}\ [\mathbf{MASK}]\ very\ much))$$

w 1 2 3 4 w'

Z_1: P(1)*P(2|1)*P(3|1,2)*P(4|1,2,3)
Z_2: P(3)*P(1|3)*P(2|3,1)*P(4|3,1,2)
Z_3: P(2)*P(4|2)*P(3|2,4)*P(1|2,4,3)



Factorization order: 2 → 4 → 3 → 1

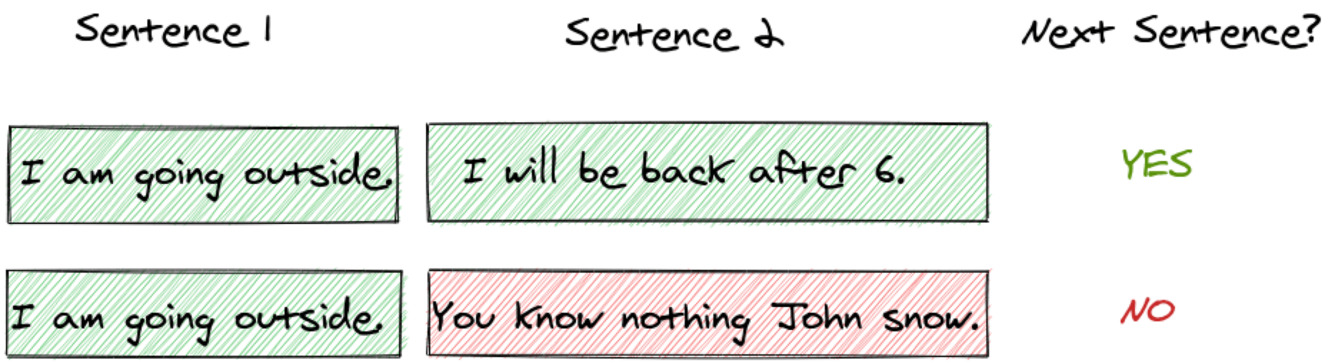
SSL in NLP: permutation language modeling

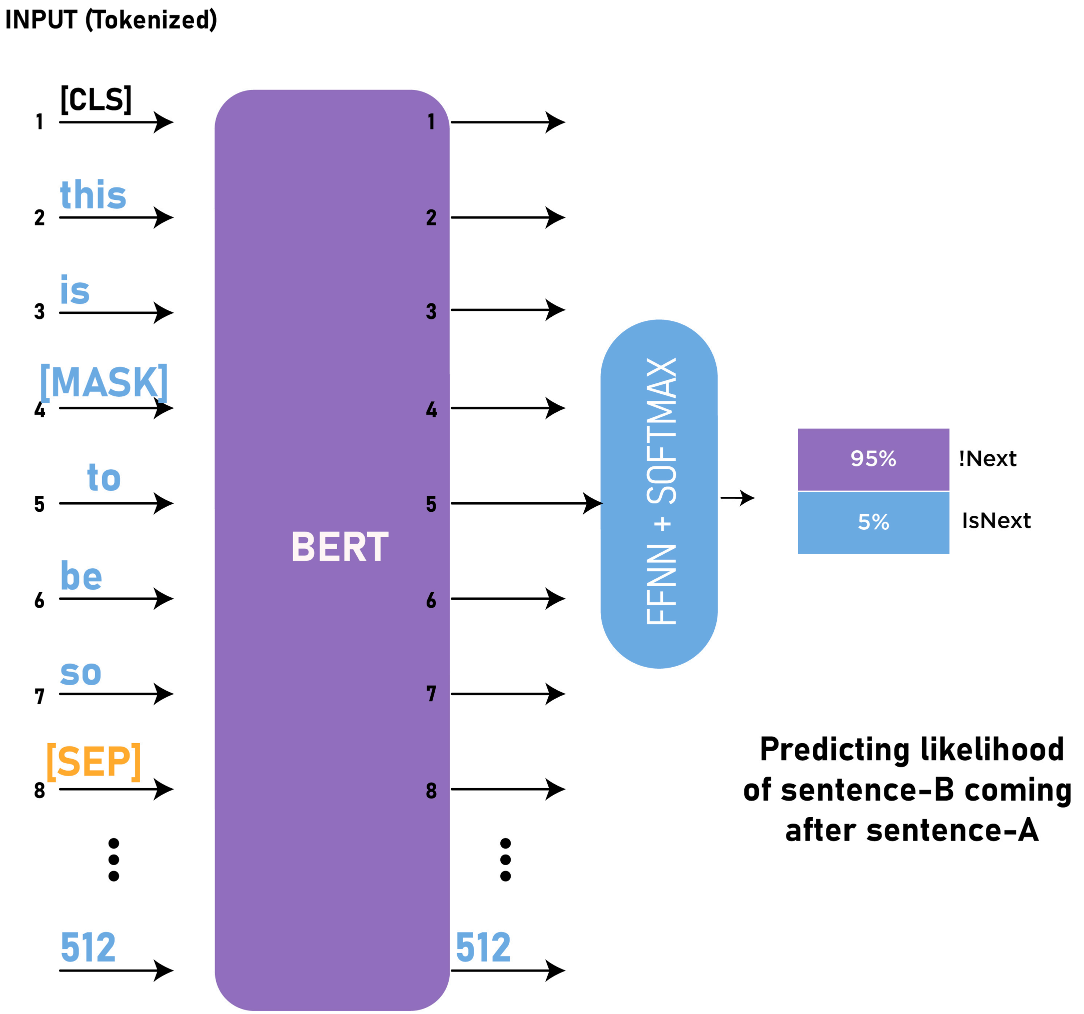
- Permutation language modeling outperforms MLM significantly (XLNet vs. BERT)
- The model structure Transformer-XL performs better than vanilla Transformer (XLNet vs. -memory)
- NSP seems to degrade the performance of XLNet (XLNet vs. +next-sent pred)

#	Model	RACE	SQuAD2.0 F1	SQuAD2.0 EM	MNLI m/mm	SST-2
1	BERT-Base	64.3	76.30	73.66	84.34/84.65	92.78
2	DAE + Transformer-XL	65.03	79.56	76.80	84.88/84.45	92.60
3	XLNet-Base ($K = 7$)	66.05	**81.33**	**78.46**	**85.84/85.43**	92.66
4	XLNet-Base ($K = 6$)	66.66	80.98	78.18	85.63/85.12	**93.35**
5	- memory	65.55	80.15	77.27	85.32/85.05	92.78
6	- span-based pred	65.95	80.61	77.91	85.49/85.02	93.12
7	- bidirectional data	66.34	80.65	77.87	85.31/84.99	92.66
8	+ next-sent pred	**66.76**	79.83	76.94	85.32/85.09	92.89

SSL in NLP: text order prediction

- Bert: Next Sentence Prediction
- Negative sampling





SSL in NLP: text order prediction

- Does next sentence prediction (NSP) work well?

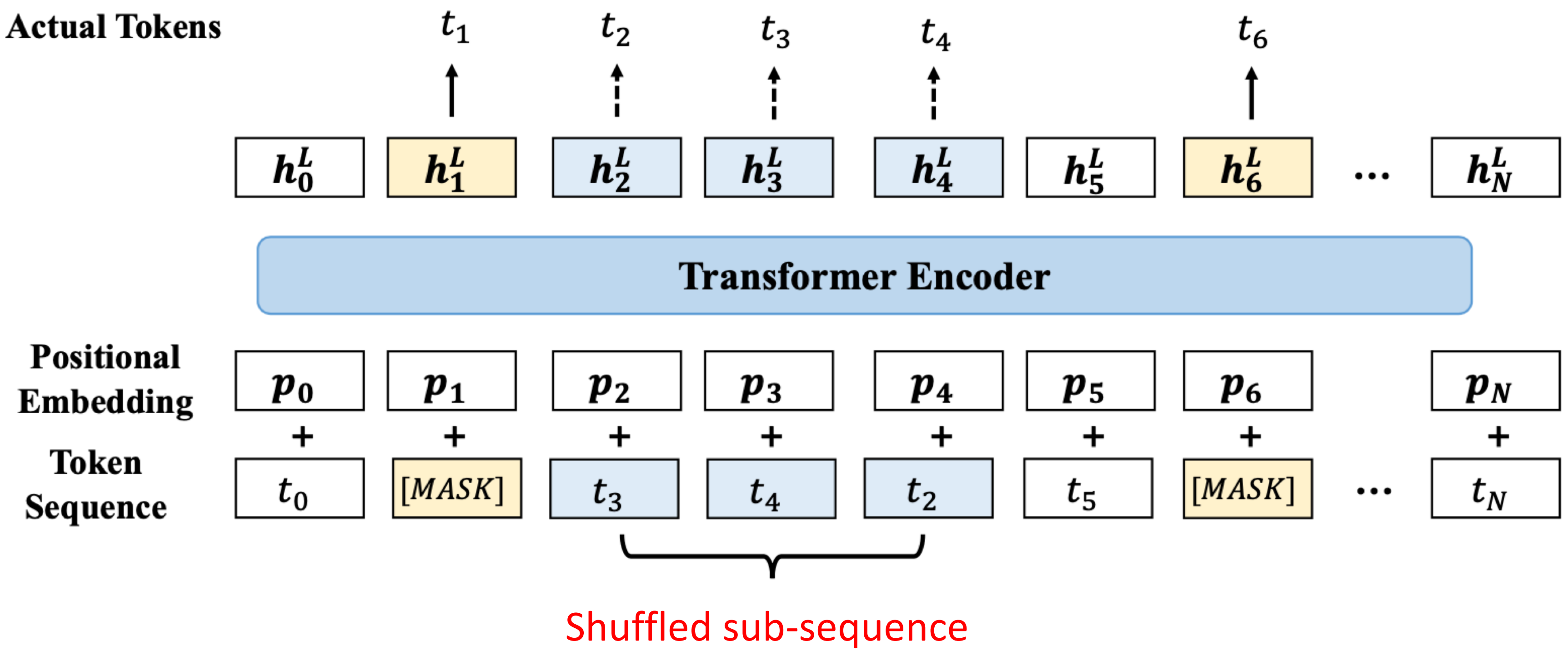
Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6

  - This 2-class classification task may be too easy for BERT to learn.
  - The input format of two sentence segments may not be consistent with downstream tasks.

Yinhan Liu et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019)

SSL in NLP: text order prediction

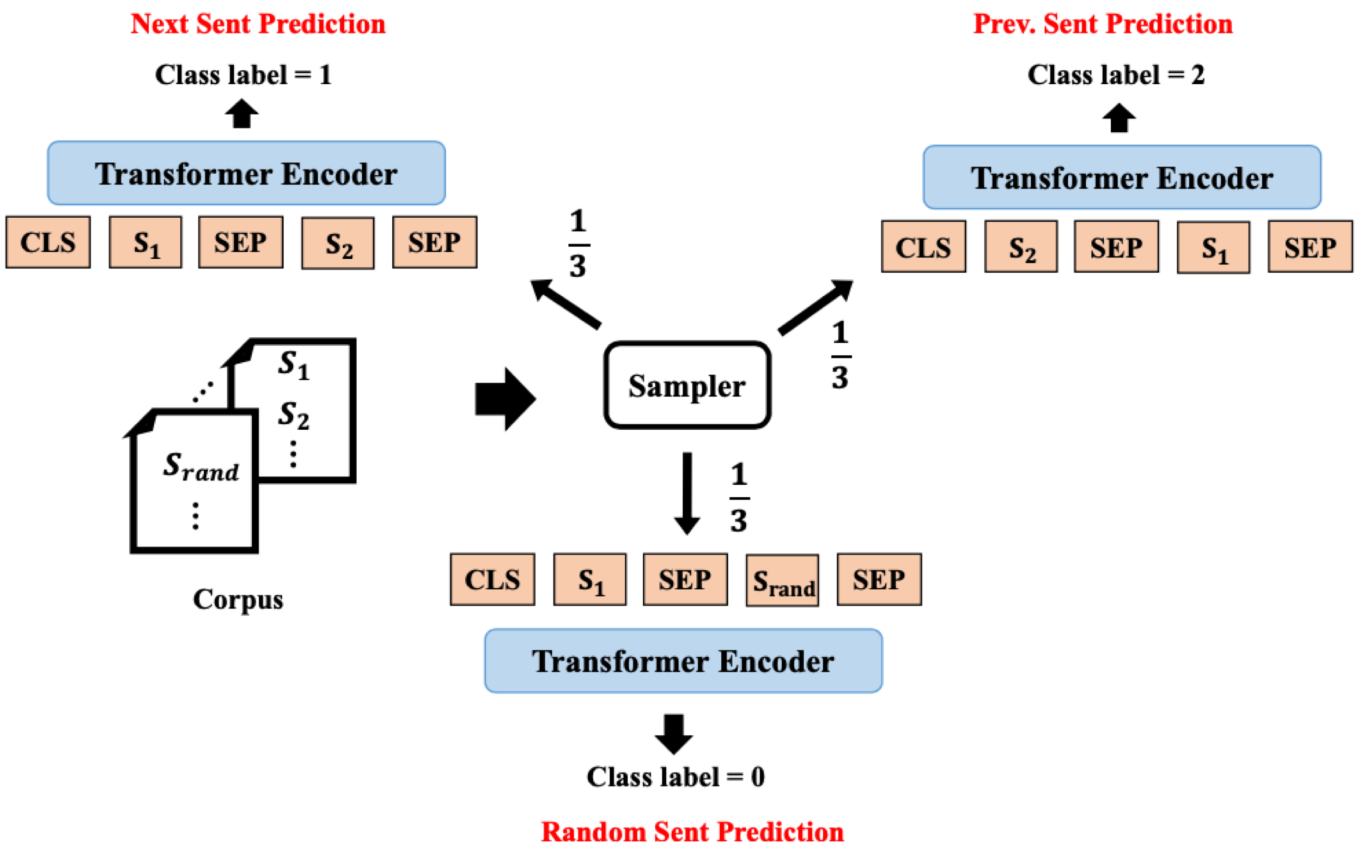
- StructBERT: word shuffle in subsequence



Wei Wang et al: StructBERT: Incorporating Language Structures into Pre-training for Deep Language Understanding. ICLR 2020.

SSL in NLP: text order prediction

StructBERT: sentence-order type prediction



SSL in NLP: text order prediction

- Fine-grained text order prediction tasks at the word level and the sentence level outperform vanilla NSP in BERT.

Task	CoLA (Acc)	SST-2 (Acc)	MNLI (Acc)	SNLI (Acc)	QQP (Acc)	SQuAD (F1)
StructBERTBase	**85.8**	**92.9**	**85.4**	91.5	**91.1**	**90.6**
-word structure	81.7	92.7	85.2	**91.6**	90.7	90.3
-sentence structure	84.9	**92.9**	84.1	91.1	90.5	89.1
BERTBase	80.9	92.7	84.1	91.3	90.4	88.5

SSL in NLP: sentence distance prediction

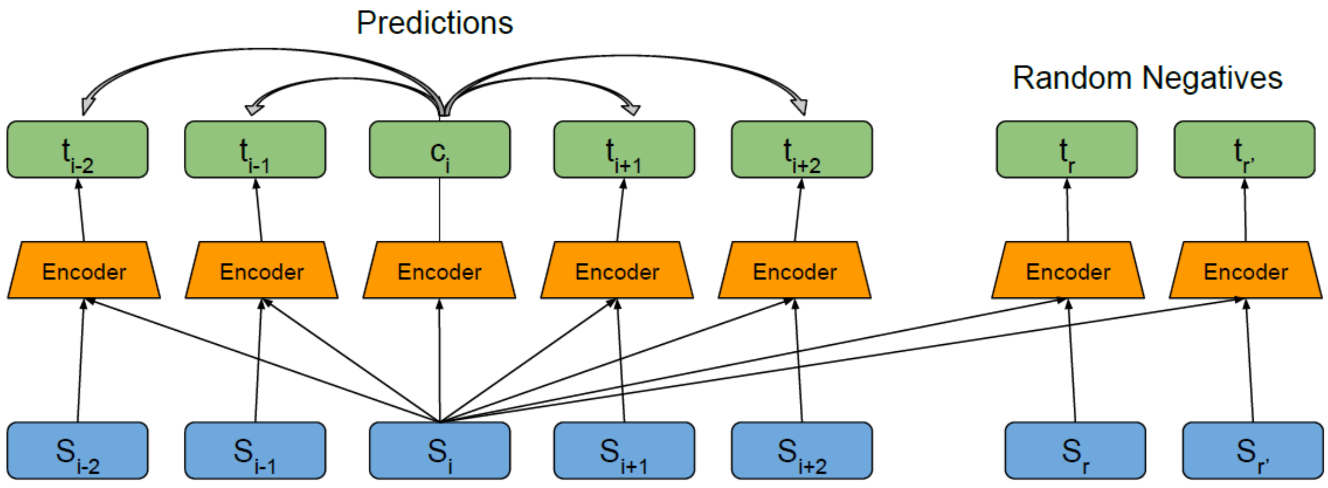
- CONPONO: Distance Prediction between Sentences

$$t_{i+k} = g_\theta(s_i, s_{i+k}), c_i = g_\theta(s_i)$$

$$\mathcal{L}_k = -\mathbb{E}_S\left[\log \frac{\exp(t_{i+k}^\top W_k c_i)}{\sum_{s_j \in S} \exp(t_j^\top W_k c_i)}\right]$$

Objective:
select candidate in S which has k-distance to s_i

Negative samples:
1) in the same document, but distance is not k to s_i
2) randomly sampled from other documents



SSL in NLP: sentence distance prediction

- CONPONO performs better than BERT-style models in the discourse-level representation tasks

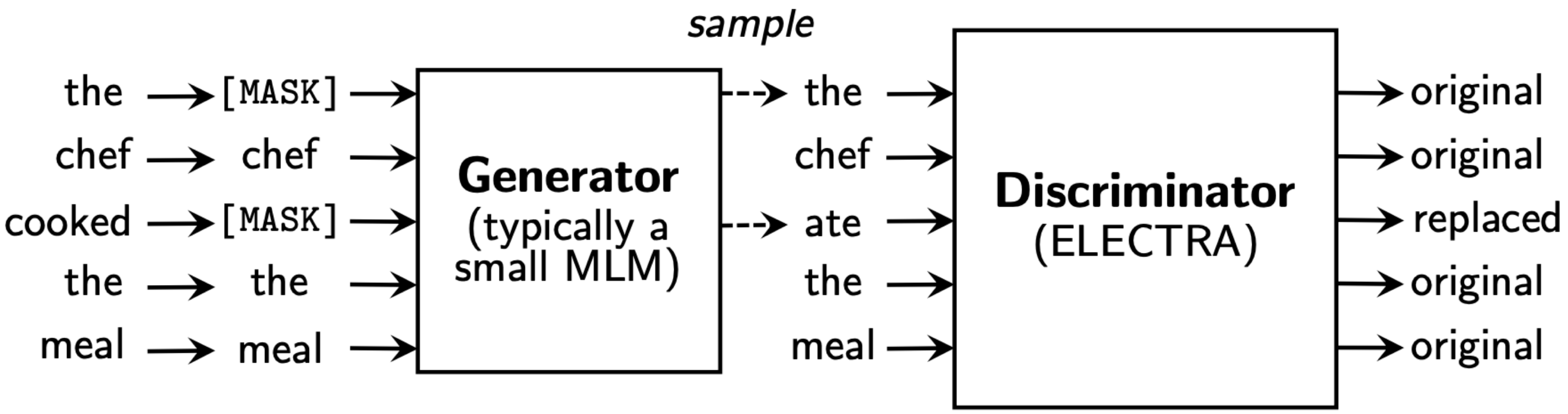
Model	SP	BSO	DC	SSP	PDTB-E	PDTB-I	RST-DT	avg.
BERT-Base	53.1	68.5	58.9	80.3	41.9	42.4	58.8	57.7
BERT-Large	53.8	69.3	59.6	**80.4**	**44.3**	43.6	59.1	58.6
RoBERTa-Base	38.7	58.7	58.4	79.7	39.4	40.6	44.1	51.4
BERT-Base BSO	53.7	72.0	71.9	80.0	42.7	40.5	**63.8**	60.6
CONPONO *isolated*	50.2	57.9	63.2	79.9	35.8	39.6	48.7	53.6
CONPONO *uni-encoder*	59.9	74.6	72.0	79.6	40.0	43.9	61.9	61.7
CONPONO (k=2)	**60.7**	**76.8**	**72.9**	**80.4**	42.9	**44.9**	63.1	**63.0**
CONPONO std.	±.3	±.1	±.3	±.1	±.7	±.6	±.2	-

- Different settings of k (window size) may work in different tasks.
  - k>1 seems key to downstream tasks, because there is more variation farther from the anchor sentence.
  - Larger distances (k>2) from the anchor sentence lead to more ambiguity.

SSL in NLP: replaced token detection

- ELECTRA: Replaced Token Detection

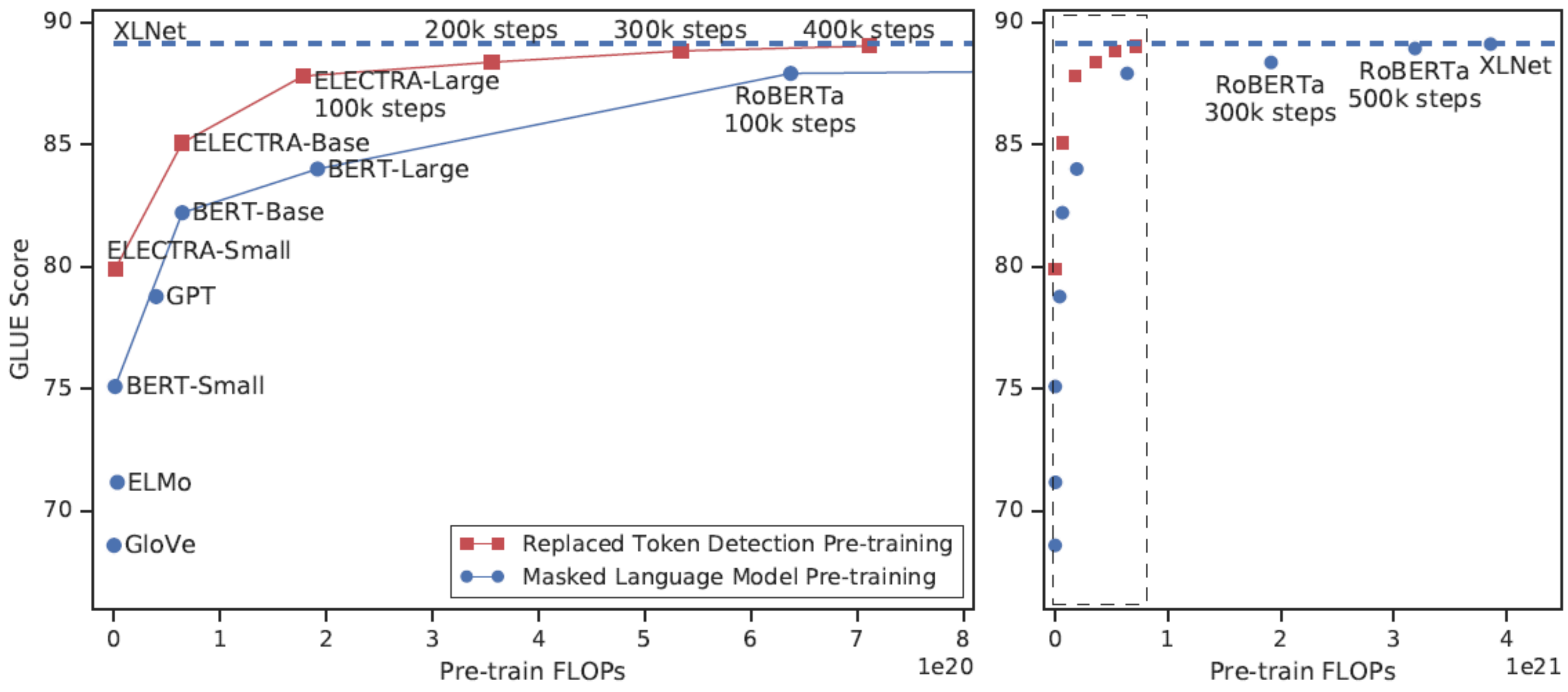
$$\mathcal{L} = \mathcal{L}_{MLM}(x, \theta_G) + \lambda \mathcal{L}_{Disc}(x, \theta_D)$$



Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning: ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. ICLR 2020.

SSL in NLP: replaced token detection

- Replaced token detection consistently outperforms language modeling (GPT) and masked language model (BERT, RoBERTa) given the same compute budget



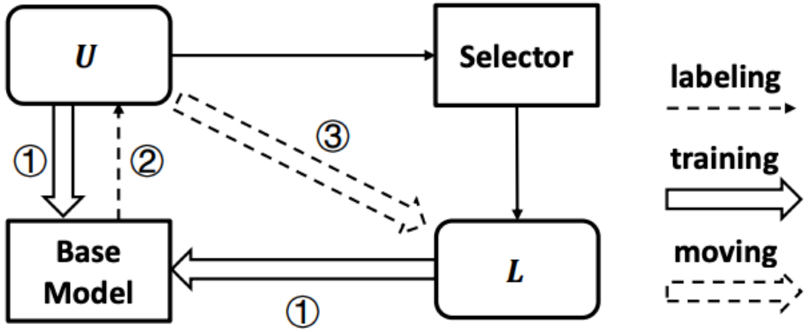
SSL in NLP：other tasks

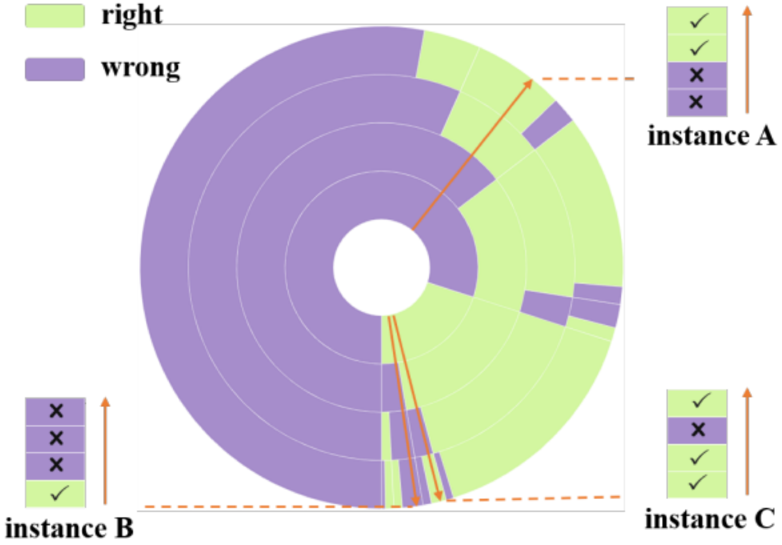
- Dialog modeling (Wu et al. ACL2019)
- Sequence-to-sequence generation (He et al. ICLR 2020)
- Machine reading comprehension (Niu et al. ACL 2020; Klein and Nabi ACL 2020)
- Text classification (5+ papers)

Evidence finding in MRC

Q: Did a little boy write the note?
D: ...**This note is from a little girl**. She wants to be your friend. If you want to be her friend, ...
A: No

Q: Is she carrying something?
D: ...On the step, I find the elderly Chinese lady, small and slight, holding the hand of a little boy. **In her other hand, she holds a paper carrier bag.** ...
A: Yes





A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction. Niu et al. ACL 2020

Unreferenced evaluation of NLG

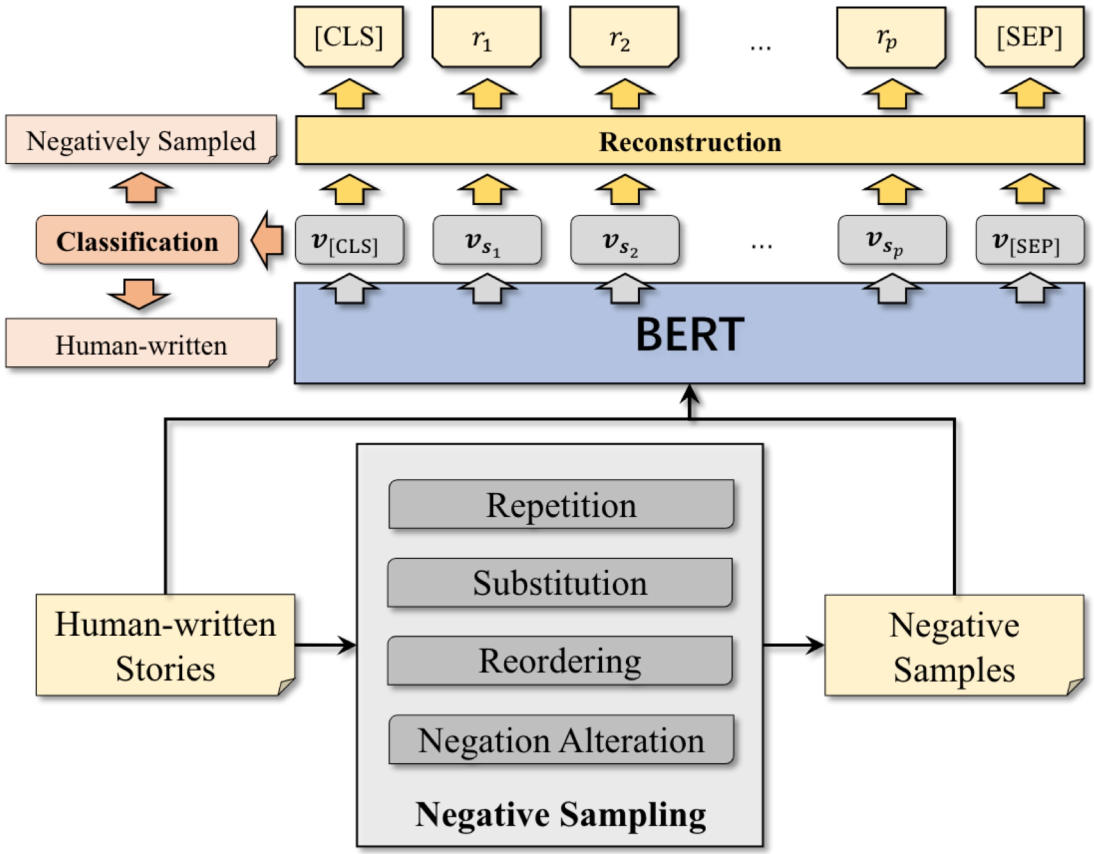
Leading Context
Jack was at the bar.

Reference By Human
He noticed a phone on the floor. He was going to take it to lost and found. But it started ringing on the way. Jack answered it and returned it to the owner's friends.

Sample 1 (Reasonable, B=0.29, M=0.49, U=1.00)
On the way out he noticed a phone on the floor. He asked around if anybody owned it. Eventually he gave it to the bartender. They put it into their lost and found box.

Sample 2 (Reasonable, B=0.14, M=0.27, U=1.00)
He had a drinking problem. He kept having more beers. After a while he passed out. When he waked up, he was surprised to find that he lost over a hundred dollars.

Sample 3 (Unreasonable, B=0.20, M=0.35, U=0.00)
He was going to get drunk and get drunk. The bartender told him it was already time to leave. Jack started drinking. Jack wound up returning but cops came on the way home.



UNION: An Unreferenced Metric for Evaluating Open-ended Story Generation. Guan&Huang.

Summary

- Self-Supervised Learning (SSL) is learning dependencies
  - Pixel-level, patch-level, word-level, sentence-level, discourse-level, etc.
  - Vector-level: making learned representations more predictive
  - Task-level: encoding task-agnostic information vs. task-specific information
- For NLP
  - Data augmentation is hard (label-preserving)
  - (Strong) Negative samples are hard to collect
  - Data perturbation seems to be very effective in many tasks

Thanks for your attention

- Recruiting post-docs, PhDs, & interns

- Minlie Huang, Tsinghua University
- aihuang@tsinghua.edu.cn
- http://coai.cs.tsinghua.edu.cn/hml