Contrastive Learning

(continued)

TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning (Su et al., arXiv'21)

Problem: Token embeddings from BERT output an anisotropic distribution of token representations that occupies a narrow subset of the entire representation space

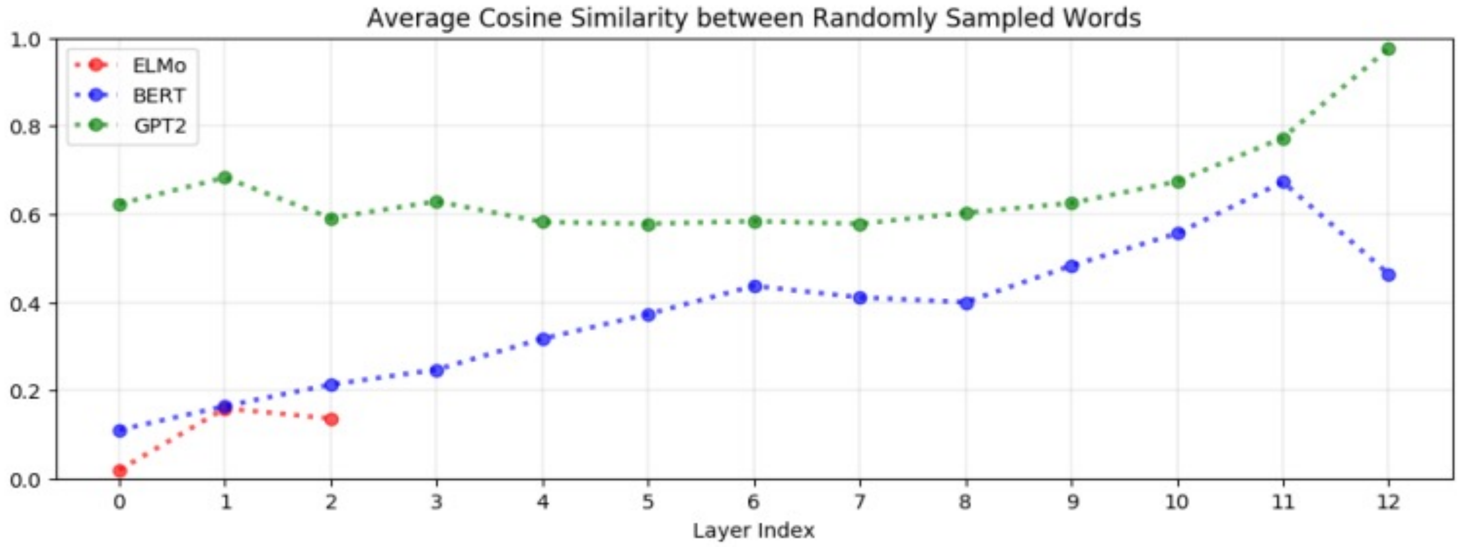


Figure 1: In almost all layers of BERT, ELMo, and GPT-2, the word representations are anisotropic (i.e., not directionally uniform): the average cosine similarity between uniformly randomly sampled words is non-zero. The one exception is ELMo's input layer; this is not surprising given that it generates character-level embeddings without using context. Representations in higher layers are generally more anisotropic than those in lower ones.

TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning (Su et al., arXiv'21)

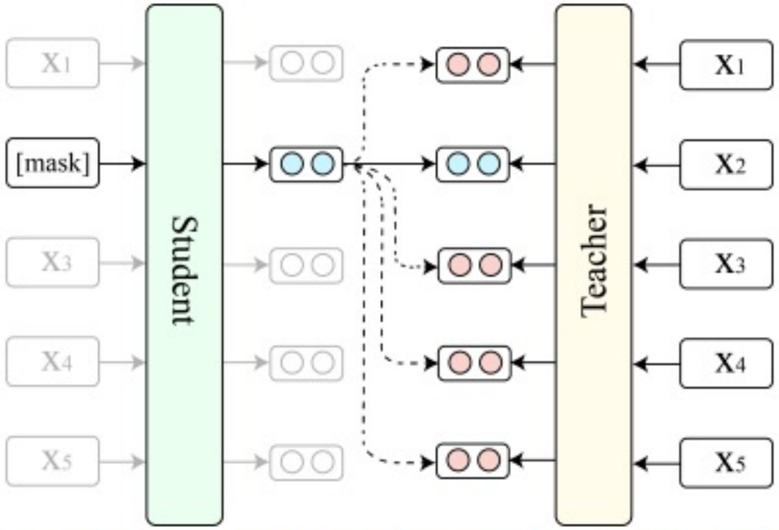


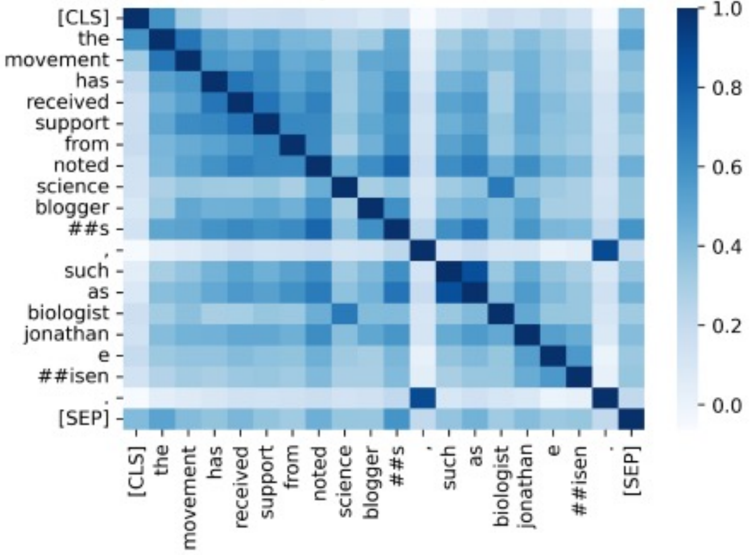
Figure 1: An overview of TaCL. The student learns to make the representation of a masked token closer to its "reference" representation produced by the teacher (solid arrow) and away from the representations of other tokens in the same sequence (dashed arrows).

$$-\sum_{i=1}^{n} \mathbb{1}(\tilde{x}_i) \log \frac{\exp(\text{sim}(\tilde{h}_i, h_i)/\tau)}{\sum_{j=1}^{n} \exp(\text{sim}(\tilde{h}_i, h_j)/\tau)},$$

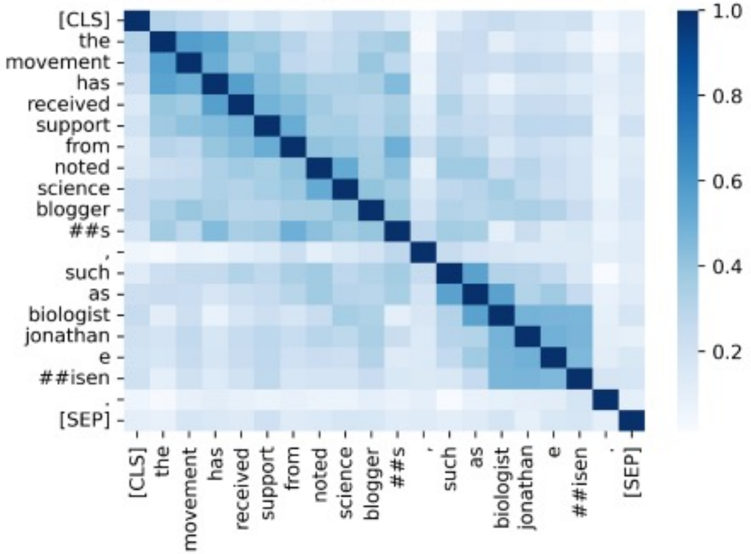
$$\mathcal{L} = \mathcal{L}_{\text{TaCL}} + \mathcal{L}_{\text{MLM}} + \mathcal{L}_{\text{NSP}}$$

TaCL: Improving BERT Pre-training with Token-aware Contrastive Learning (Su et al., arXiv’21)

(a) Self-similarity Visualization of BERT



(b) Self-similarity Visualization of TaCL



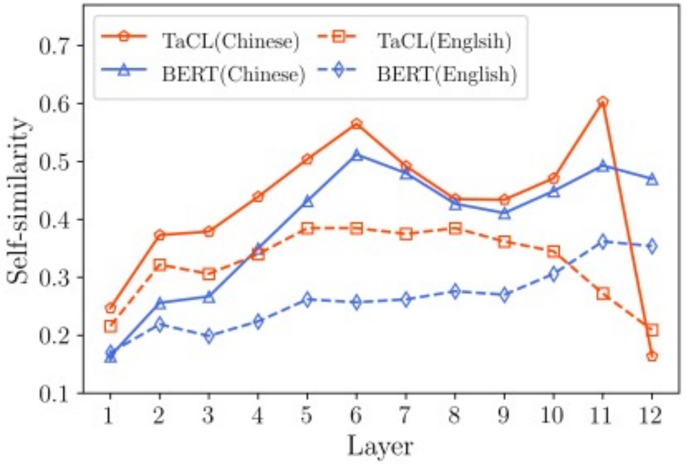


Figure 2: Layer-wise representation self-similarity.

$$s(x) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \text{cosine}(h_i, h_j),$$

Supervised TaCL loss for token-level tasks

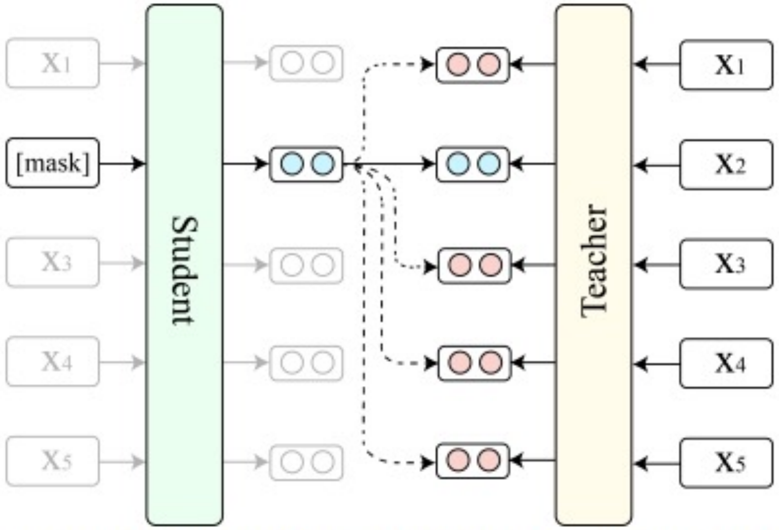


Figure 1: An overview of TaCL. The student learns to make the representation of a masked token closer to its “reference" representation produced by the teacher (solid arrow) and away from the representations of other tokens in the same sequence (dashed arrows).

Unsupervised TaCL loss (original)

$$-\sum_{i=1}^{n} \mathbb{1}(\tilde{x}_i) \log \frac{\exp(\mathrm{sim}(\tilde{h}_i, h_i)/\tau)}{\sum_{j=1}^{n} \exp(\mathrm{sim}(\tilde{h}_i, h_j)/\tau)},$$

Supervised TaCL loss (mine)

$$-\sum_{i=1}^{n} 1(\tilde{x}_i) \log \frac{\exp(sim(\tilde{h}_i, h_i)/\tau)}{\sum_{j=1}^{n} 1(x_j) \exp(sim(\tilde{h}_i, h_j)/\tau)}, \tag{1}$$

where $1(\tilde{x}_i) = 1$, if x_i is a masked token, otherwise $1(\tilde{x}_i) = 0$ and $1(x_j) = 1$, if x_j's class is not same as x_i's class, otherwise $1(x_j) = 0$.

COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining (Meng et al., NeurIPS'21)

- Challenge: Efficiency, Anisotropy
- BERT

$$\left[x_1^{\text{orig}}, \ldots, [\text{MASK}]_i, \ldots, x_n^{\text{orig}}\right] \xrightarrow{\text{Transformer}} \boldsymbol{H} \xrightarrow{\text{MLM Head}} p_{\text{MLM}}(x|\boldsymbol{h}_i)$$

$$p_{\text{MLM}}(x|\boldsymbol{h}_i) = \frac{\exp(\boldsymbol{x}^\top \boldsymbol{h}_i)}{\sum_{x_t \in V} \exp(\boldsymbol{x}_t^\top \boldsymbol{h}_i)}; \quad \mathcal{L}_{\text{MLM}} = \mathbb{E}\left(-\sum_{i \in \mathcal{M}} \log p_{\text{MLM}}\left(x_i^{\text{orig}} \middle| \boldsymbol{h}_i\right)\right)$$

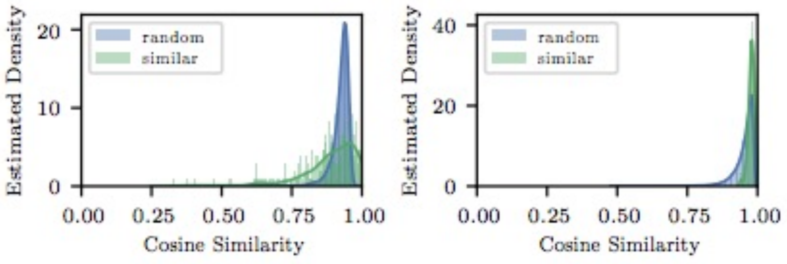
- ELECTRA

$$x_i^{\text{MLM}} \sim p_{\text{MLM}}\left(x|\boldsymbol{h}_i\right), \text{ if } i \in \mathcal{M}\,; \quad x_i^{\text{MLM}} = x_i^{\text{orig}}, \text{ else.}$$

$$X^{\text{MLM}} \xrightarrow{\text{Main Transformer}} \boldsymbol{H} \xrightarrow{\text{RTD Head}} p_{\text{RTD}}\left(\mathbb{1}(x_i^{\text{MLM}} = x_i^{\text{orig}}) \middle| \boldsymbol{h}_i\right)$$

COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining (Meng et al., NeurIPS'21)

- Challenges of ELECTRA-style pretraining
  - Missing LM benefits
    - Lack of LM capability (necessary for prompt based learning)
    - Binary task not sufficient to capture word-level semantics
  - Squeezing representation space
    - Two random sentences have high similarity score (lack of uniformity)
    - Closely related sentence have different representations (lack of alignment)
- COCO-LM: matches MNLI accuracy of RoBERTa and ELECTRA with 60% and 50% of GPU hours in pretraining



(a) RoBERTa. (b) ELECTRA.

Figure 1: Cosine similarity distributions of random/similar sequence pairs using [CLS] embeddings from pretrained models. Histograms/curves are distribution bins/kernel density estimates.

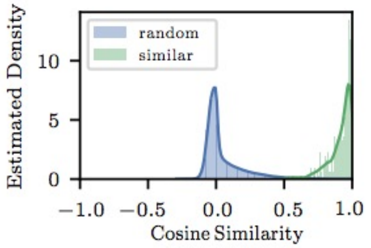


Figure 5: Cosine similarity of sequence pairs randomly sampled from pretraining corpus and most similar pairs from STS-B using [CLS] from COCO-LM$_{\text{Base}}$.

COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining (Meng et al., NeurIPS’21)

- COCO-LM introduces 2 pretraining tasks

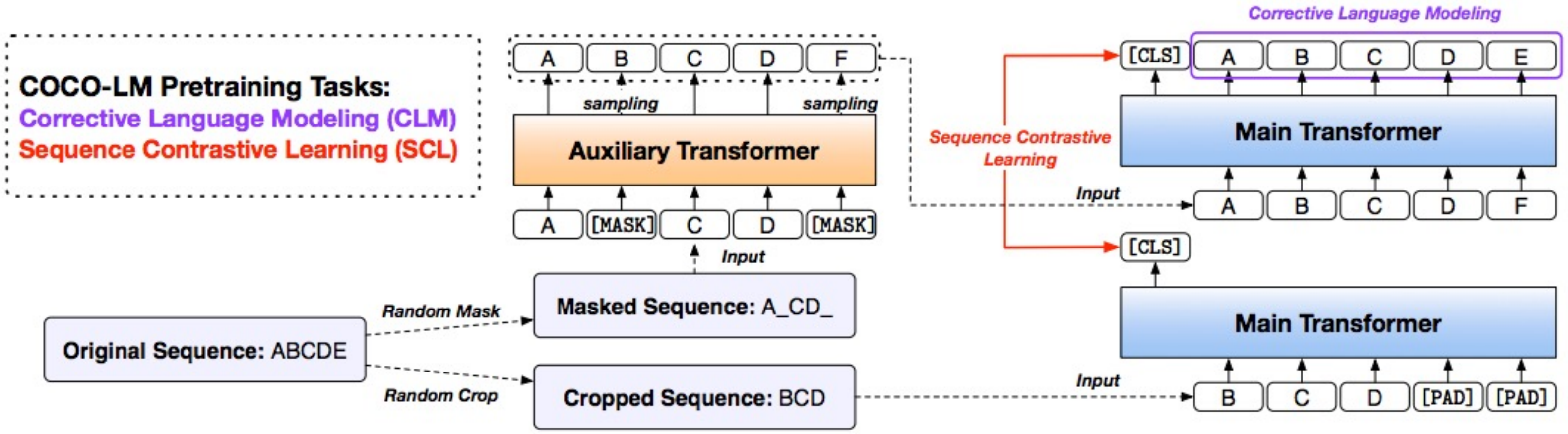
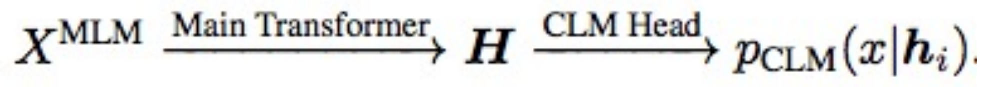


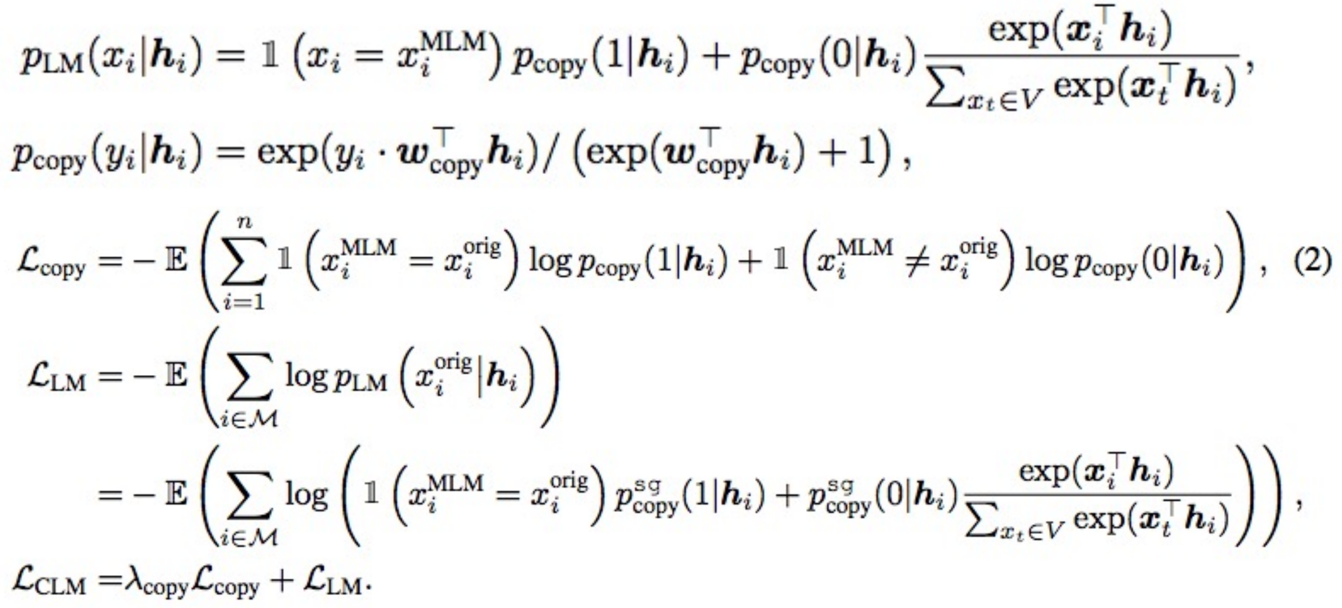
Figure 2: The overview of COCO-LM. The auxiliary Transformer is pretrained by MLM. Its corrupted text sequence is used as the main Transformer’s pretraining input in Corrective Language Modeling and paired with the cropped original sequence for Sequence Contrastive Learning.

COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining (Meng et al., NeurIPS’21)

- COCO-LM introduces 2 pretraining tasks
- Task-1: Corrective LM (CLM)



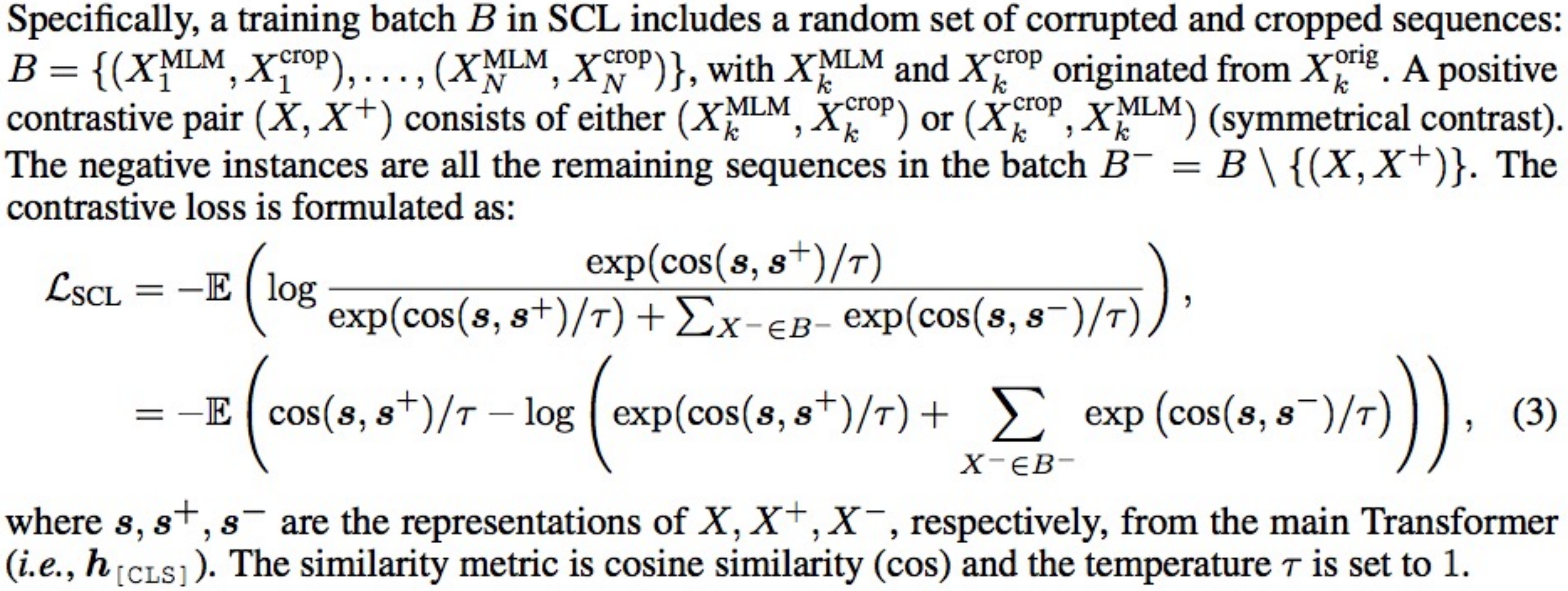
$$\boldsymbol{X}^{\text{MLM}} \xrightarrow{\text{Main Transformer}} \boldsymbol{H} \xrightarrow{\text{CLM Head}} p_{\text{CLM}}(x|\boldsymbol{h}_i).$$



$$\begin{aligned}
p_{\text{LM}}(x_i|\boldsymbol{h}_i) &= \mathbb{1}\left(x_i = x_i^{\text{MLM}}\right) p_{\text{copy}}(1|\boldsymbol{h}_i) + p_{\text{copy}}(0|\boldsymbol{h}_i) \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{h}_i)}{\sum_{x_t \in V} \exp(\boldsymbol{x}_t^\top \boldsymbol{h}_i)}, \\
p_{\text{copy}}(y_i|\boldsymbol{h}_i) &= \exp(y_i \cdot \boldsymbol{w}_{\text{copy}}^\top \boldsymbol{h}_i) / \left(\exp(\boldsymbol{w}_{\text{copy}}^\top \boldsymbol{h}_i) + 1\right), \\
\mathcal{L}_{\text{copy}} &= -\mathbb{E}\left(\sum_{i=1}^{n} \mathbb{1}\left(x_i^{\text{MLM}} = x_i^{\text{orig}}\right) \log p_{\text{copy}}(1|\boldsymbol{h}_i) + \mathbb{1}\left(x_i^{\text{MLM}} \neq x_i^{\text{orig}}\right) \log p_{\text{copy}}(0|\boldsymbol{h}_i)\right), \quad (2) \\
\mathcal{L}_{\text{LM}} &= -\mathbb{E}\left(\sum_{i \in \mathcal{M}} \log p_{\text{LM}}\left(x_i^{\text{orig}}|\boldsymbol{h}_i\right)\right) \\
&= -\mathbb{E}\left(\sum_{i \in \mathcal{M}} \log\left(\mathbb{1}\left(x_i^{\text{MLM}} = x_i^{\text{orig}}\right) p_{\text{copy}}^{\text{sg}}(1|\boldsymbol{h}_i) + p_{\text{copy}}^{\text{sg}}(0|\boldsymbol{h}_i) \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{h}_i)}{\sum_{x_t \in V} \exp(\boldsymbol{x}_t^\top \boldsymbol{h}_i)}\right)\right), \\
\mathcal{L}_{\text{CLM}} &= \lambda_{\text{copy}} \mathcal{L}_{\text{copy}} + \mathcal{L}_{\text{LM}}.
\end{aligned}$$

COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining (Meng et al., NeurIPS’21)

- Task-2: Sequence Contrastive Learning (SCL)

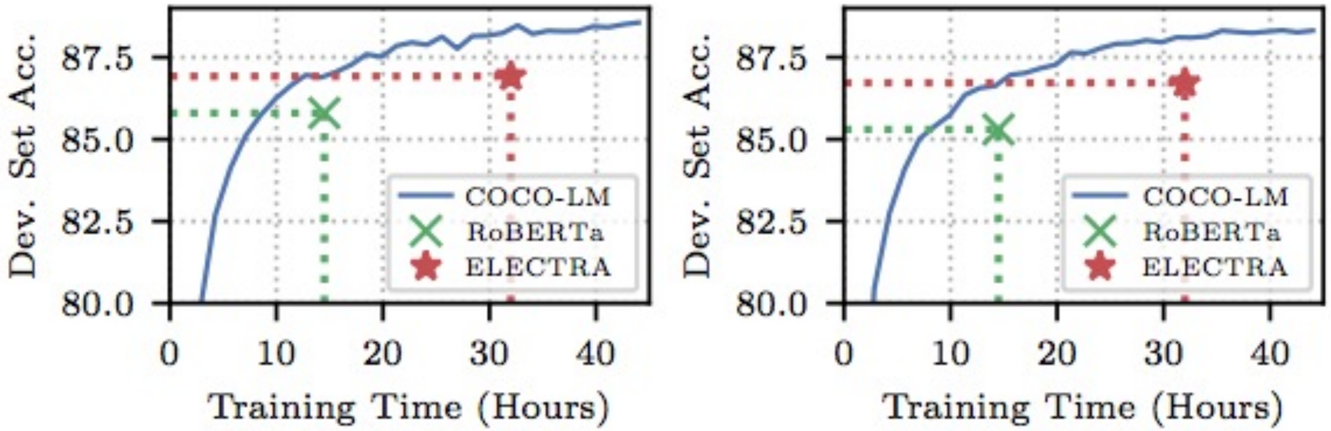


Specifically, a training batch B in SCL includes a random set of corrupted and cropped sequences: $B = \{(X_1^{\text{MLM}}, X_1^{\text{crop}}), \dots, (X_N^{\text{MLM}}, X_N^{\text{crop}})\}$, with X_k^{MLM} and X_k^{crop} originated from X_k^{orig}. A positive contrastive pair (X, X^+) consists of either $(X_k^{\text{MLM}}, X_k^{\text{crop}})$ or $(X_k^{\text{crop}}, X_k^{\text{MLM}})$ (symmetrical contrast). The negative instances are all the remaining sequences in the batch $B^- = B \setminus \{(X, X^+)\}$. The contrastive loss is formulated as:

$$\begin{aligned}\mathcal{L}_{\text{SCL}} &= -\mathbb{E}\left(\log \frac{\exp(\cos(\boldsymbol{s}, \boldsymbol{s}^+)/\tau)}{\exp(\cos(\boldsymbol{s}, \boldsymbol{s}^+)/\tau) + \sum_{X^- \in B^-} \exp(\cos(\boldsymbol{s}, \boldsymbol{s}^-)/\tau)}\right), \\ &= -\mathbb{E}\left(\cos(\boldsymbol{s}, \boldsymbol{s}^+)/\tau - \log\left(\exp(\cos(\boldsymbol{s}, \boldsymbol{s}^+)/\tau) + \sum_{X^- \in B^-} \exp\left(\cos(\boldsymbol{s}, \boldsymbol{s}^-)/\tau\right)\right)\right), \quad (3)\end{aligned}$$

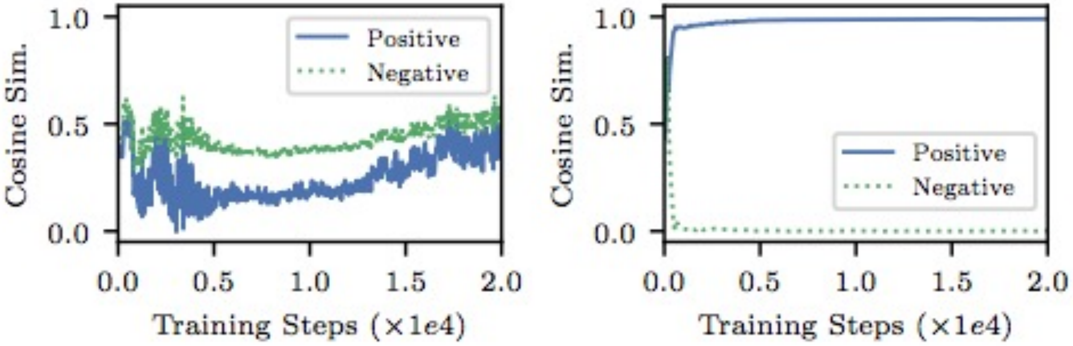
where $\boldsymbol{s}, \boldsymbol{s}^+, \boldsymbol{s}^-$ are the representations of X, X^+, X^-, respectively, from the main Transformer (*i.e.*, $\boldsymbol{h}_{\texttt{[CLS]}}$). The similarity metric is cosine similarity (cos) and the temperature τ is set to 1.

COCO-LM: Correcting and Contrasting Text Sequences for Language Model Pretraining (Meng et al., NeurIPS’21)



(a) MNLI-m

(b) MNLI-mm



(a) Without SCL

(b) With SCL