Contrastive Self-supervised BERT for Vision and Language Pre-training

Shentong Mo*, Jingfei Xia, Ihor Markevych
Carnegie Mellon University

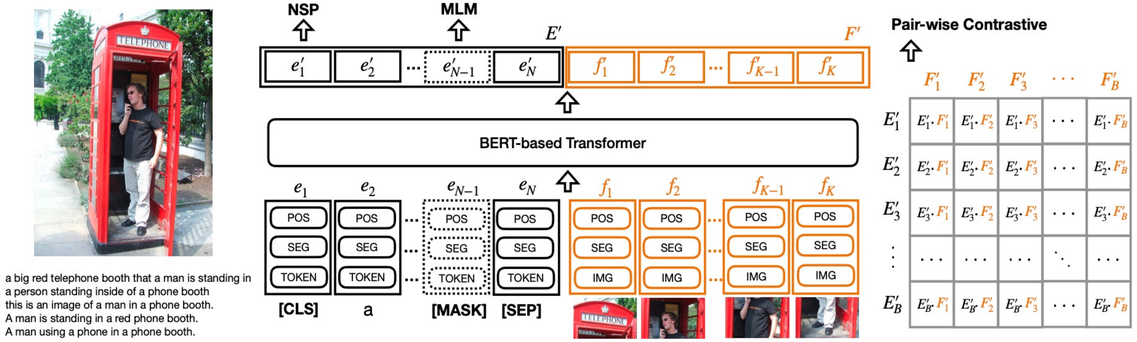
ICML International Conference On Machine Learning

Introduction

We present a simple but effective Contrastive Self-Supervised framework based on BERT for visual and linguistic representations learning, namely **CSS-BERT**, which applies a pair-wise contrastive loss to learn alignments between the whole sentence and each image.

- ✓ Our CSS-BERT mitigates the semantic confusion between the whole sentence and each image during pre-training.
- ✓ Our CSS-BERT achieves competitive performance when transferred to four main downstream tasks.
- ✓ Extensive ablation studies demonstrate the effectiveness of the pair-wise contrastive loss in our CSS-BERT.

CSS-BERT



- Linguistic pre-training (MLM + NSP)

1) a token embedding $\mathbf{e}^t$ for each sub-word in a sentence;

2) a segment embedding $\mathbf{e}^s$ indicating which part of the text the token is from;

3) a position embedding $\mathbf{e}^p$ for the position of the token in the sentence.

- Visual pre-training

1) an image feature embedding $\mathbf{f}^i$ for each image ROI;

2) a segment embedding $\mathbf{f}^s$ indicating which token embedding the image embedding is opposed to;

3) a position embedding $\mathbf{f}^p$ for alignments between tokens and each image ROI.

- Visual-linguistic pre-training

The Pair-wise Contrastive Loss (PwCL) between linguistic embeddings $\mathbf{E}'_i$ and visual embeddings $\mathbf{F}'_i$ is defined as

$$\mathcal{L}_{\text{PwCL}} = -\log \frac{\sum_{i=1}^{B}(\mathbf{E}'_i \cdot \mathbf{F}'_i)}{\sum_{i=1}^{B}\sum_{j=1}^{B}\mathbb{1}_{i\neq j}(\mathbf{E}'_i \cdot \mathbf{F}'_j)}$$

Downstream Tasks

Table 1. Comparison results on the VQA dataset.

Model	test-dev	test-std
LXMERT (Tan & Bansal, 2019)	72.42	72.54
ViLBERT (Lu et al., 2019)	70.55	70.92
VisualBERT (Li et al., 2019)	70.08	71.00
UNITER (Chen et al., 2019)	72.27	72.46
VL-BERT (Su et al., 2020)	71.79	72.22
CVLP (Shi et al., 2020)	72.77	72.90
CSS-BERT (ours)	**72.88**	**73.05**

Table 2. Comparison results on the VCR dataset.

Model	Q → A		QA → R		Q → AR	
	val	test	val	test	val	test
ViLBERT (Lu et al., 2019)	72.40	73.30	74.50	74.60	54.00	54.80
VisualBERT(Li et al., 2019)	70.80	71.60	73.20	73.20	52.20	52.40
UNITER (Chen et al., 2019)	-	75.00	-	77.20	-	58.20
VL-BERT$_{\text{base}}$ (Su et al., 2020)	73.77	-	74.36	-	55.20	-
CSS-BERT (ours)	**75.33**	**75.65**	**76.52**	**77.87**	**58.65**	**59.47**
VL-BERT$_{\text{large}}$ (Su et al., 2020)	75.50	75.80	77.90	78.40	58.90	59.70

Ablation Study

Table 3. Ablation study on contrastive pre-training and batch size. MLM, NSP, and PwCL denote Masked Language Modeling, Next Sentence Prediction, and Pair-wise Contrastive Loss.

MLM	NSP	PwCL	batch size	test-dev (↑)	test-std (↑)	APS (↑)
✓	✓		64	70.11±0.12	71.03±0.15	0.43±0.08
✓		✓	64	70.82±0.13	71.56±0.15	0.52±0.06
	✓	✓	64	70.06±0.12	70.95±0.13	0.41±0.06
✓	✓	✓	64	71.73±0.15	72.08±0.17	0.68±0.04
✓	✓	✓	128	72.18±0.13	72.32±0.16	0.72±0.04
✓	✓	✓	256	72.42±0.11	72.67±0.13	0.76±0.03
✓	✓	✓	512	**72.88±0.08**	**73.05±0.07**	**0.83±0.02**
✓	✓	✓	1024	72.83±0.05	72.96±0.06	0.81±0.02

Conclusion

We propose a simple but effective contrastive self-supervised approach based on BERT for pre-training visual and linguistic representations jointly, namely CSS- BERT.

Acknowledgements

We greatly thank Po-yao Huang and Prof. Graham Neubig for insightful discussions.