XDBERT: Distilling Visual Information to BERT from Cross-Modal Systems to Improve Language Understanding

Chan-Jan Hsu[1,2], **Hung-yi Lee**[1], **Yu Tsao**[2]
[1]National Taiwan University, Taiwan
[2]Academia Sinica, Taiwan
{r09946011, hungyilee}@ntu.edu.tw,
yu.tsao@citi.sinica.edu.tw

Abstract

Transformer-based models are widely used in natural language understanding (NLU) tasks, and multimodal transformers have been effective in visual-language tasks. This study explores distilling visual information from pretrained multimodal transformers to pretrained language encoders. Our framework is inspired by cross-modal encoders' success in visual-language tasks while we alter the learning objective to cater to the language-heavy characteristics of NLU. After training with a small number of extra adapting steps and finetuned, the proposed XDBERT (cross-modal distilled BERT) outperforms pretrained-BERT in general language understanding evaluation (GLUE), situations with adversarial generations (SWAG) benchmarks, and readability benchmarks. We analyze the performance of XDBERT on GLUE to show that the improvement is likely visually grounded.

1 Introduction

Transformer-based models are extensively used in natural language understanding (NLU) tasks, and some prominent pretraining strategies include BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), ALBERT (Lan et al., 2020), and ELECTRA (Clark et al., 2020). Despite their differences in curating the learning objectives, they all utilize text-based datasets only. In the real world, however, humans can benefit from the visual modality when acquiring knowledge from language; an obvious example is learning visually grounded words, such as colors and shapes.

Some studies have succeeded with visually grounded information used in NLU. ViCo (Gupta et al., 2019) learned visual co-occurrences in text and reported superior performance to GloVe in word analogy problems. Zhang et al. (2020) and Huang et al. (2020) used images to boost translation performance in supervised and unsupervised settings. Tan and Bansal (2020) reported improvements over BERT on NLU by proposing the concept of vokenization.

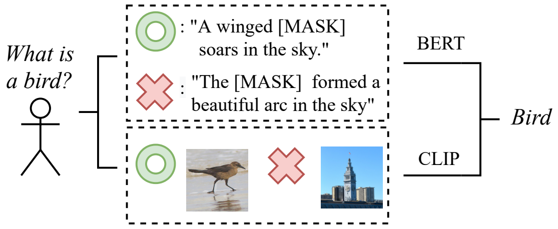
What is a bird?
"A winged [MASK] soars in the sky."
"The [MASK] formed a beautiful arc in the sky"
BERT
CLIP
Bird

Figure 1: Humans can answer cloze questions and match a word with an image, and the multi-views of a word could be simulated by neural networks. While BERT excels in masked word reconstruction, CLIP (Section 3) specializes at image-text matching. The two modalities have different collocations of concepts, which incentivize joint learning from the two systems.

Another branch of research focuses on solving multimodal downstream tasks such as visual question answering and image retrieval. Li et al. (2019); Lu et al. (2019); Su et al. (2020); Li et al. (2020) trained visual-text transformers, while LXMERT (Tan and Bansal, 2019) used different encoders for text and image and a cross-modal encoder. Tan and Bansal (2020) tested these models with general language understanding evaluation (GLUE Wang et al. (2018)) and found that the performance does not exceed using BERT (Appendix A), drawing the conclusion that vision-and-language pretraining on visually-grounded language dataset failed to distill useful information for general NLU. CLIP (Radford et al., 2021) utilizes contrastive loss to reach SOTA on zero-shot image classification in a retrieval fashion.

In this work, we establish the link between pretrained multimodal transformers and visually-grounded language learning. We devise a way to distill visual information from components of a pretrained multimodal transformer (CLIP text-transfomer, abbreviated as CLIP-T) to pretrained

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 479 - 489
May 22-27, 2022 ©2022 Association for Computational Linguistics

language transformers (BERT/ELECTRA), to incorporate versatile perception of words into the model (Figure 1). The usage of a visually grounded text-transformer as a teacher allows us to implement straightforward and non-fuzzy adapting tasks for distillation. We show that it is mathematically logical that the CLIP-T output approximates visual features (Sec. 2.2), and also the linguistic competence of CLIP-T is low (Sec. 3), to prove that the distilled information is predominantly visual and thus non-trivial to the pretrained-language transformer despite having textual inputs.

Methodologically, we use the cross-modal encoder structure inspired by Tan and Bansal (2019), to concatenate the two models and further adapt the ensemble for some extra steps (a lot fewer than the original pretraining steps). While adapting pretrained-BERT, we favor a document-level corpus (wiki103) over a vision-language corpus (MSCOCO) due to claims from Devlin et al. (2019)[1] and results from Tan and Bansal (2020) (Appendix A). The adapting tasks are joint masked language modeling (MLM), same sentence prediction, and CLIP token classification tasks, which are resemblant of BERT pretraining tasks to cater to the language-heavy characteristics of NLU. We do ablation studies to show that each of the task provides improvement (Section 5).

During finetuning, we finetune XDBERT (cross-modal distilled BERT), which is the language encoder after adaptation. We evaluate the linguistic capabilities of the model by finetuning on GLUE, situations with adversarial generations (SWAG (Zellers et al., 2018)) benchmarks, and readability benchmarks[2]. The resulting XDBERT outperforms pretrained BERT, proving that our adaptation strategy distills useful visual knowledge into BERT (right of Figure 2). We provide analysis to show that the improvements are visually grounded.

We summarize our contribution as follow:

- We explore distilling visual information from a pretrained multimodal transformer to a pretrained language transformer and improved NLU performance.
- Our adapting method is efficient and extensible to different combinations of pretrained-language encoders (BERT/ELECTRA).

2 Proposed Method

The training process consists of three phases: pretraining, adaptation, and finetuning (Figure 2). Our proposed method focuses on the adaptation phase with pretrained models, so pretraining is not a part of our experiment, but we explain all three phases for completeness. The adaptation phase incorporates the cross-modal transformer structure to jointly learn from CLIP-T and BERT outputs.

2.1 Model Architecture

The cross-modal transformer (middle of Figure 2) consists of a cross-modal encoder, CLIP-T and BERT. CLIP-T has the same module connections as BERT with only parameter differences (specifications in Appendix B). The cross-modal encoder consists of repeating cross-modal encoder layers, which is an extension to single-modality encoder layers (layers of BERT/CLIP-T) in Figure 3. The added cross-attention module follows the attention formula (Vaswani et al., 2017):

$$Attention\ output = softmax\left(\mathbf{Q} * \mathbf{K}^T / \sqrt{D}\right)\mathbf{V} \quad (1)$$

for queries ($\mathbf{Q}$), keys ($\mathbf{K}$) and values ($\mathbf{V}$) of dimension D, however, $\mathbf{Q}$ is generated from a modality other than $\mathbf{K}$ and $\mathbf{V}$. We choose the number of cross-modal encoder layers to be 2.

2.2 Pretraining

BERT is trained using the next sentence prediction and masked language modeling. CLIP is an image-text matching system with two components, a text encoder (CLIP-T), and an image encoder (CLIP-ViT), which learn to encode paired inputs to closer output embeddings via contrastive loss. The trained representation has the following properties:

$$cos(H_i, V_i) >> cos(H_i, V_j)(i \neq j) \quad (2)$$

$$cos(H_i, V_i) >> cos(H_j, V_i)(i \neq j) \quad (3)$$

where H_i is the CLIP text encoder output of X_i, and V_i is the CLIP image encoder output of Y_i. The text-image input (X_i, Y_i) is paired, and every (X_j, Y_k) ($j \neq k$) is a non-pair. Since H_i and V_i are normalized and have a length of 1, H_i can be used to approximate V_i. The similarity of H_i and V_i is also shown in multi-modal arithmetic propreties discovered in Tewel et al. (2021) Therefore, we use the CLIP text encoder output to approximate CLIP image encoder output for a straightforward adaptation process.

[1] "It is critical to use a document-level corpus rather than a shuffled sentence-level corpus such as the BillionWord Benchmark in order to extract long contiguous sequences"

[2] https://www.kaggle.com/c/commonlitreadabilityprize

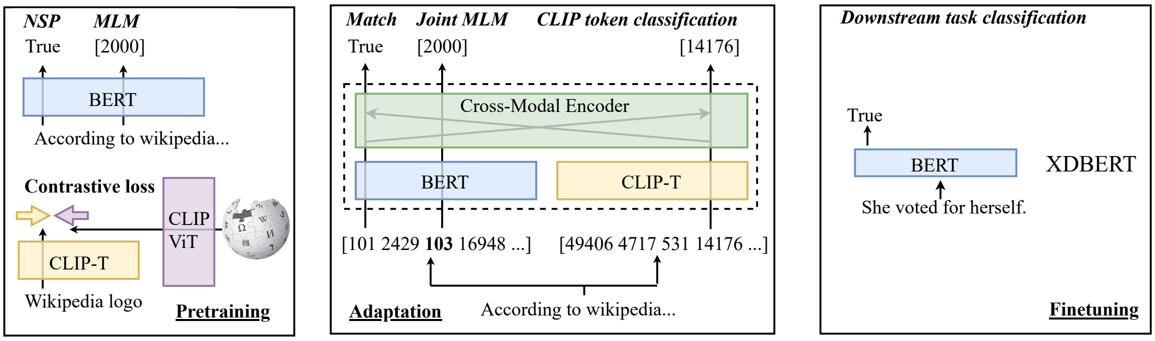

Figure 2: In our experimental setting, the transformers go through three phases of the training processes from left to right. The pretraining phase pretrains BERT and CLIP-T, both of which are then used in the adaptation phase and concatenated with a cross-modal encoder. Finetuning is performed on the language encoder only (XDBERT); in this case, a positive CoLA example is being processed to determine its linguistic acceptability. ViT stands for Vision Transformer (Dosovitskiy et al., 2021), and the input id 103 is the [MASK] token in BERT.

2.3 Adaptation

We define three adapting tasks that can be learned in a self-supervised manner, which is visualized in Figure 2. In these tasks, BERT and CLIP-T takes sentences A and B respectively as input, and losses are calculated from both BERT output and CLIP-T output. Our adapting tasks closely follow BERT text pretraining strategies to retain linguistic competence. Unlike pretraining, the adaptation is computationally inexpensive, as we found that training 1 epoch on wiki103 was already effective. Further training details can be found in Appendix C.

2.3.1 Joint Masked Language Modeling (MLM)

The MLM objective teaches the model to reconstruct masked tokens. The masked ratio and masked token replacement probabilities follow Devlin et al. (2019). Since there is no equivalent of a [MASK] token in CLIP, we leave the sentence as is.

2.3.2 Same sentence prediction (MATCH)

The Image-Text Matching (ITM) objective is widely used in multimodal learning (Tan and Bansal, 2020; Radford et al., 2021). We modify this objective to same sentence prediction as both streams of our model takes text as input. When choosing the input sentences for BERT and CLIP-T, we make the inputs nonidentical 50% of the time. A binary classifier over [CLS] differentiates between the two cases. This motivates the [CLS] output to encode sentence related information, and trains the cross-attention weights.

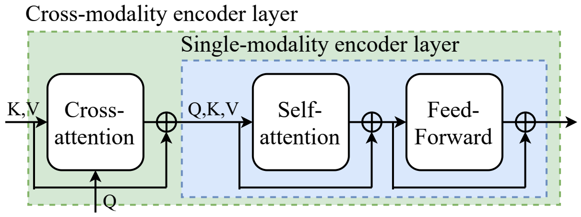

Figure 3: Single-modality encoder layer (blue) and cross-modal encoder layer (green)

2.3.3 CLIP Token Classification

This is the MLM objective done on the CLIP-T side of the full model, omitting the masking part because CLIP has no mask token. Same as MLM, 15% of the tokens are randomly selected for reconstruction. We address concerns on trivial solutions learned by the model in Section 5 and 9 in the appendix.

2.4 Finetuning

Finetuning follows the methods described in Devlin et al. (2019), and is applied to the language encoder only (XDBERT), therefore the number of parameters are kept equal to pretrained-BERT.

3 Experimental Results

We evaluated our model on three NLU benchmarks, namely GLUE, SWAG and READ. We tested our adaptation strategy on three different language encoders coupled with CLIP-T, including BERT-base, ELECTRA-base, and ELECTRA-large. We fix the finetuning parameters between models where comparison is intended, and select the median result of

	RTE	MPRC	STSB	CoLA	SST2	QNLI	QQP	MNLI	SWAG	READ↓
CLIP-T	51.62	76.20	22.07	25.41	–	–	–	–	–	–
BERT-b	66.43	87.38	88.64	56.52	92.46	90.92	89.51	84.35	81.0	–
XDBERT-b	**69.31**	**88.02**	**89.32**	**57.55**	**92.78**	**91.52**	**89.57**	**84.75**	**81.35**	–
ELECTRA-b	78.70	89.49	90.77	66.09	94.5	92.69	90.29	88.23	88.60	–
XDELECTRA-b	**80.51**	**90.55**	**91.04**	**66.76**	**95.20**	**93.03**	**90.4**	**88.75**	**88.73**	–
ELECTRA-l	86.64	91.53	91.88	69.27	96.90	94.78	**91.34**	90.99	92.46	0.685
XDELECTRA-l	**87.73**	**92.12**	**91.97**	**70.98**	**97.36**	**94.93**	91.29	**91.02**	**92.59**	**0.635**

Table 1: NLU task results on the test set (READ) and the dev set (GLUE,SWAG). The results are the median value of 5 runs using different random seeds (9 runs on RTE). BERT-b is the BERT-base-uncased model from Devlin et al. (2019), while XDBERT-b is the proposed models shown in the right part of Figure 2. ELECTRA-b and ELECTRA-l refer to the ELECTRA-base model and the ELECTRA-large model from Clark et al. (2020) respectively. READ (readability benchmark) uses RMSE loss as the evaluation metric.

multiple runs. Details of finetuning are provided in Appendix C.

Table 1 shows experimental results. Each of our XD-model constantly outperforms the original encoder (For fair comparison, we train the original encoder with one more epoch of wiki103). We found that performance gains are more significant on smaller datasets (RTE, MRPC, STSB, CoLA), indicating that visual features help increase generalization when the amount of training data is limited. The gains are also significant on the readability benchmark (READ).

We show that the results of finetuning CLIP-T alone on GLUE does not perform well. Since the language capability of the CLIP-T model is weak, the distilled information obtained by XDBERT/XDELECTRA is predominantly visual.

It is also possible to finetune the entire cross-modal transformer after adaptation. The performance further increases but the model has more parameters. The results are in Appendix C.3.

4 Analysis

To justify the use of a cross-modal encoder, we first conducted a pairwise projection weighted canonical correlation analysis (PWCCA) on word embeddings. The PWCCA is a good measure to determine how close the distributions of two vector groups are to each other. The PWCCA results in Table 2 show low scores on both BERT/CLIP and ELECTRA/CLIP before co-training, so the cross-modal encoder is useful in learning from both distributions.

We inspect RTE, MRPC, and CoLA results of 5 runs in detail to show that the improvements are likely from visual information of CLIP-T. Over the 5 runs, XDBERT-b has accumulated +38 more correct classifications than BERT-b, or +2.74%(38/5/277) gain in performance. MPRC and CoLA show +0.3% and +0.9% gains in accuracy respectively, and translates to a larger gain in performance with their original metric (MRPC F1: +0.83%, CoLA Corr: +2.2%). We then separate each of the glue datasets entries into two categories: entries that XDBERT-b improves classification over BERT-b, and entries of the opposite. Entries where both models obtain the same performance are set aside. Analyzing the separated entries as a whole, we discovered that the better-performing entries have a larger visually grounded ratio (Figure 4), as the quartile, median and mean values are generally higher for improved samples. The enhancement of visually grounded token rep-

Systems	PWCCA
BERT/ELECTRA	0.5498
BERT/CLIP	0.4980
ELECTRA/CLIP	0.4645
BERT/RANDOM	0.3569

Table 2: PWCCA results for different combinations of systems. RANDOM denotes embeddings generated from a uniform distribution.

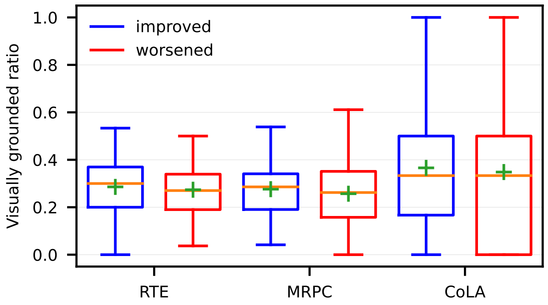

Figure 4: Characteristic analysis of RTE, MRPC, and CoLA entries categorized by performance difference between XDBERT-b and BERT-b. The Green plus symbol denotes the mean value. The visually grounded ratio estimation follows Tan and Bansal (2020).

	RTE	**MPRC**	**STSB**	**CoLA**
MLM+MATCH+CLIPTC(proposed)	69.31	88.02	89.32	56.27
MLM+MATCH	70.04	86.93	88.8	54.62
MLM	68.23	87.25	89.29	54.78
1 cross attention layer	66.79	87.66	89.32	53.62
2 Epochs (2x)	69.31	88.04	89.31	55.91
20 Epochs (20x)	57.4	87.74	-	-
wiki(14G), same steps as above	65.3	87.78	89.1	-

Table 3: Ablation study results. The results are the median value of 5 runs using a learning rate of 1e-4 on XDBERT-b. The CoLA learning rate differs from that in the main paper.

resentations is a rough indicator that XDBERT has obtained distilled visual information from CLIP-T. We show examples of each category in Appendix D.

5 Ablation study

We tried various combinations of adaptation tasks and found out that using all three yielded the best results. We also tried to reduce the number of cross-modal encoder layers to one; however, no further improvements were made upon the visually grounded language encoder. Other experiments include changing the number of layers in the cross-modal encoder, training for longer, and swapping to a much larger wiki (14G). Swapping to wiki reduces potential overfitting from the 20 Epochs setting trained on wiki103, as training for the same amount of steps on wiki is less than 1 epoch. We tested these changes on RTE, MPRC, STSB, and CoLA on 5 random seeds, and the results are shown in Table 3, where MLM refers to the joint MLM objective, MATCH refers to the cross-modal matching objective, and CLIPTC refers to the CLIP token classification objective.

Besides experimental evidence, we also justify the CLIPTC loss via further analysis, as the CLIPTC objective can theoretically be trivially solved by identity mapping. Despite this possibility, we find that the loss is crucial to cross attention learning. Since we do not impose negative hard samples from sampled sentences, the MATCH objective can be solved sufficiently simply by guiding the cross attention to focus on common trivial words. With the CLIPTC objective, the diversity of the input embeddings corresponding to different tokens must be retained in the cross-modal encoder, leading to more robust cross-modal attention. We show comparisons of the attention maps generated from the cross-modal encoders with a random sequence from RTE in Table 9 in the Appendix to verify this claim.

6 Conclusion

In this study, we explored using cross-modal encoders to distill visual information to BERT. We adapted the model with multiple objectives, and we were able to achieve improved performance on NLU tasks. Our adaptation techniques are computationally inexpensive and straightforward. Furthermore, our method is language encoder agnostic, as we show similar performance gains on XDELECTRA.

Acknowledgements

This work was supported in part by Ministry of Science and Technology (MOST), Taiwan, under Contract 110-2223-E-002 -007-MY3.

References

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Tanmay Gupta, Alexander G. Schwing, and Derek Hoiem. 2019. Vico: Word embeddings from visual co-occurrences. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 7424–7433. IEEE.

Po-Yao Huang, Junjie Hu, Xiaojun Chang, and Alexander Hauptmann. 2020. Unsupervised multimodal neural machine translation with pseudo visual pivoting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8226–8237, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pretraining for vision-language tasks. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pretraining of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2021. Zero-shot image-to-text generation for visual-semantic arithmetic. *CoRR*, abs/2111.14447.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Visual-Text Transformers Results on NLU

We show the result of Visual-Text Transformers on GLUE, reported by Tan and Bansal (2020) in Table 7. All of the listed methods (except LXMERT) have their text-transformers initialized from BERT. The results show that multi-modal training for solving vision-language tasks does not improve the performance of the models on natural language understanding tasks.

	BERT-b	BERT-l	CLIP
dim	768	1024	512
max_len	512	512	77
#layers	12	24	12

Table 4: BERT and CLIP configurations. ELECTRA has a structure identical to that of BERT. The tokenizers of BERT and CLIP are also different.

B Modeling sequences on CLIP

While BERT and CLIP have similar forwarding mechanisms, the specifications of the transformer architecture are different, resulting in challenges to jointly model both models (Table 4).

Mismatching dimensions pose a problem in cross-attention. We use a linear transformation to generate **Q**, **K**, and **V** of matching dimensions, but clarify that this linear transformation layer exists in the original LXMERT setting where hidden representations have unified dimensions.

We modify the input to address the mismatched max_len of the two systems. In the joint MLM, we used a fixed sequence length of 512 for the BERT. However, the same cannot be done for CLIP as the maxmum model sequence length is 77 for CLIP. We found that most BERT sequences (>99%) of length 512 encode into CLIP sequences of length less than 693, so we pad the CLIP sequence to length 693, and then split the CLIP sequence into 9 sub-sequences of length 77. Therefore, a batch of inputs will contain BERT inputs of size (batch_size, 512) and CLIP inputs of size (batch_size, 9, 77). The output was resized to (batch_size, 693) in the cross-modal encoder. The issue is also present in the finetuning phase, and the maximum sequence length of GLUE and SWAG is 128; therefore we used 2 blocks of CLIP sub-sequences to model it. For bi-sequence classification tasks such as RTE and MRPC, we ensure that separate sentences do not use the same block in the CLIP encoder. Therefore, uni-sequence classification tasks will have a CLIP input size of (batch_size, 2, 77) and the bi-sequence classification task will have a CLIP input size of (batch_size, 4, 77).

C Further Training Details

C.1 Adaptation

We use publicly available wiki103 and preprocessing methods similar to Tan and Bansal (2020) [3]. Wiki103 (500MB) is a subset of the Wikipedia corpus consisting of only good and featured articles. The adaptation of 1 epoch on wiki103 finished in 35 minutes on 8 V100s (BERT-base). We trained for at most 20 epochs(16k steps) and found that further adaptation steps did not increase scores in early epochs, and significantly decreased performance in late epochs. We used the following parameters for adaptation : learning rate = 1e-4, max_epoch = 40 (although we stopped early due to plummeting performance), warmup ratio = 0.05

C.2 Finetuning

The learning rates are listed in Table 5.

	base-sized	large-sized
RTE,MRPC,STSB	1e-4	5e-5
others	2e-5	1e-5

Table 5: Finetuning configurations for NLU tasks. The full model uses the same learning rate as its language encoder

We used a warmup ratio of 0.1, with a learning rate decay of 0.9, and trained the model for 3 epochs. We report the median results of 5 runs on different random seeds, except for RTE, which is unstable; therefore, we report the median results of 9 runs instead. The reproduce results of ELECTRA on RTE and STSB are lower than values reported by Clark et al. (2020) because we did not start from an MNLI checkpoint.

C.3 Finetuning with Full Model

Since our cross-modal transformer itself is can also be viewed as a language encoder, finetuning can be done on the full model. This approach, however, adds extra parameters to pretrained-BERT, so comparison with pretrained-BERT is not intended, instead, we focus on showing the feasibility of this approach. The number of additional parameters is only a function of the hidden size in BERT/ELECTRA, so when the language encoder is large, the ratio of additional parameters is much more insignificant. To simplify notations, we use X-(language encoder) to represent the full model. The

[3] https://github.com/airsplay/vokenization

number of parameters of the full model is shown in Table 6 and the results on NLU tasks are shown in Table 8.

model	parameters
BERT-b / ELECTRA-b	109482240
XBERT-b / XELECTRA-b	202059009
ELECTRA-l	334092288
XELECTRA-l	442671617

Table 6: Number of paramters for each model.

D RTE Examples

We provide three RTE example of each type in Figure4, and we choose extreme examples where performance difference is huge over 5 runs for both "Improved" and "Worsened" categories. We follow Tan and Bansal (2020) to classify tokens as visually-grounded if it is not a stopword and has more than 100 occurrences in MSCOCO. In the following examples, **Bold** words are visually-grounded, while normal words are non-visually-grounded. Words in brackets are stopwords and does not count towards either category.

D.1 Improved : XDBERT outperforms BERT

Example1 :
Visually-grounded ratio : 11/(11+16) = 0.4074
BERT answered correctly : 0/5
XDBERT answered correctly : 5/5

hands across (the) divide (was) formed (in) march 2001 (,) (and) **one** (of) (its) immediate aims (was) (to) press (for) (more) freedom (of) contact (and) communication **right away** (between) (the) **two parts** (of) cyprus (,) (and) (for) early **progress towards** (a) solution (to) (') (the) cyprus problem (') (.)

cyprus (was) divided (into) **two parts** (in) march 2001 (.)

Example2 :
Visually-grounded ratio : 4/(10+4) = 0.2857
BERT answered correctly : 0/5
XDBERT answered correctly : 5/5

(it) (is) hoped (that) **women** (,) (who) constitute (more) (than) **half** (of) (the) population (,) (will) vote (for) (other) **women** (and) ensure (that) (their) issues (are) represented (in) parliament (.)

women (are) poorly represented (in) parliament (.)

Example3 :
Visually-grounded ratio : 13/(13+17) = 0.4333
BERT answered correctly : 0/5
XDBERT answered correctly : 5/5

ho ##dler claimed (there) (were) **also** irregularities (in) (the) campaigns **organized** (by) atlanta (for) (the) 1996 summer **games** (,) sydney (for) (the) summer olympics (in) 2000 (and) salt **lake city** (for) (the) 2002 **winter games** (.)

(before) salt **lake city** (,) **winter** olympic **games** took **place** (in) naga ##no (.)

D.2 On Par : XDBERT and BERT perform equally

Example1 :
Visually-grounded ratio : 6/(6+32) = 0.1375
BERT answered correctly : 0/5
XDBERT answered correctly : 0/5

(on) october 1 2001 (,) eu (and) (other) countries introduced (the) option (for) domestic **animal** owners (to) apply (for) **pet** passports (under) (the) pets **travel** scheme (() pets (for) **short** ()) (,) (for) pets **returning** (from) abroad (to) (the) **united** kingdom (.) (this) replaced (the) **old system**(of) 6 months compulsory qu ##aran ##tine (for) (all) domestic pets (.)

(in) 2001 (,) (the) eu introduced (a) passport(for) pets (.)

Example2 :
Visually-grounded ratio : 5/(5+16) = 0.2381
BERT answered correctly : 5/5
XDBERT answered correctly : 5/5

security forces (were) (on) **high** alert (after) (an) election campaign (in) (which) (more) (than) 1 (,) 000 **people** (,) **including seven** election candidates (,) (have) (been) killed (.)

security forces (were) (on) **high** alert (after) (a) campaign marred (by) violence (.)

Example3 :
Visually-grounded ratio : 8/(8+16) = 0.3333
BERT answered correctly : 5/5
XDBERT answered correctly : 5/5

(in) 1979 (,) (the) leaders signed (the) egypt (-) israel peace treaty (on) (the) **white house lawn** (.) (both) president begin (and) **sad ##at** received (the) nobel peace prize (for) (their)

work (.) (the) **two** nations (have) enjoyed peaceful relations (to) (this) **day** (.)

(the) israel (-) egypt peace agreement (was) signed (in) 1979 (.)

D.3 Worsened : XDBERT underperforms BERT

Example1 :
Visually-grounded ratio : 11/(11+29) = 0.2750
BERT answered correctly : 5/5
XDBERT answered correctly : 0/5

jean (-) claude tri ##chet (,) (the) **european** central **bank** president (,) **made** (it) **clear** (,) (on) wednesday (,) (that) (he) would oppose **un** ##war ##rant **##ed** political **attempts** (to) remove antonio **fa** ##zio (:) (the) **bank** (of) italy governor (,) engulfed (in) controversy (over) (his) handling (of) **bank** takeover bids (.)

antonio **fa** ##zio (is) subordinate (to) jean (-) claude tri ##chet (.)

Example2 :
Visually-grounded ratio : 11/(11+29) = 0.4167
BERT answered correctly : 5/5
XDBERT answered correctly : 0/5

(about) **half** (were) **along** (a) 20 (-) mile stretch (of) **santa** monica **bay** (from) **top** anga canyon boulevard (to) (the) palo **s** verde **s** peninsula (.)

(the) coastline (of) **santa** monica **bay** (is) 50 miles **long** (.)

Example3 :
Visually-grounded ratio : 32/(32+55) = 0.3678
BERT answered correctly : 5/5
XDBERT answered correctly : 0/5

cairo (is) (now) **home** (to) (some) 15 million **people** (-) (a) **bu** ##rgeon **##ing** population (that) produces approximately 10 (,) 000 tonnes (of) rubbish per **day** (,) **putting** (an) enormous strain (on) **public** services (.) (in) (the) **past** 10 years (,) (the) government (has) tried **hard** (to) encourage private investment (in) (the) refuse sector (,) (but) (some) estimate **4** (,) 000 tonnes (of) waste (is) **left behind** every **day** (,) fest ##ering (in) (the) heat (as) (it) **waits** (for) **someone** (to) **clear** (it) (up) (.) (it) (is) often (the) **people** (in) (the) poor ##est neighbourhoods (that) (are) worst affected (.) (but) (in) (some) areas (they) (are) **fighting** **back** (.) (in) shu ##bra (,) **one** (of) (the) northern districts (of) (the) **city** (,) (the) residents (have) **taken** (to) (the) **streets** armed (with) **dust** ##pan **##s** (and) **brushes** (to) **clean** (up) **public** areas (which) (have) (been) **used** (as) **public dump ##s** (.)

15 million tonnes (of) rubbish (are) produced daily (in) cairo (.)

	Diff. to BERT weight	**SST-2**	**QNLI**	**QQP**	**MNLI**
VL-BERT	6.4e-3	90.1	89.5	88.6	82.9
VisualBERT	6.5e-3	90.3	88.9	88.4	82.4
Oscar	41.6e-3	87.3	50.5	86.6	77.3
LXMERT	42.0e-3	82.4	50.5	79.8	31.8
BERT/ViLBERT	–	90.3	89.6	88.4	82.4

Table 7: Results of using Visual-Text Transformers on Natural Language Understanding reported by Tan and Bansal (2020). ViLBERT is identical to BERT because its weights are frozen during multimodal finetuning.

	RTE	**MPRC**	**STSB**	**CoLA**	**SST2**	**QNLI**	**QQP**	**MNLI**	**SWAG**	**READ↓**
XBERT-b	69.31	88.46	89.59	59.05	92.89	91.47	89.37	84.62	81.34	–
XELECTRA-b	79.78	91.06	91.46	66.8	95.06	93.04	90.62	88.97	88.91	–
XELECTRA-l	88.45	92.33	92.04	70.51	97.36	94.97	91.4	91.03	92.83	0.565

Table 8: NLU task results using the full model.

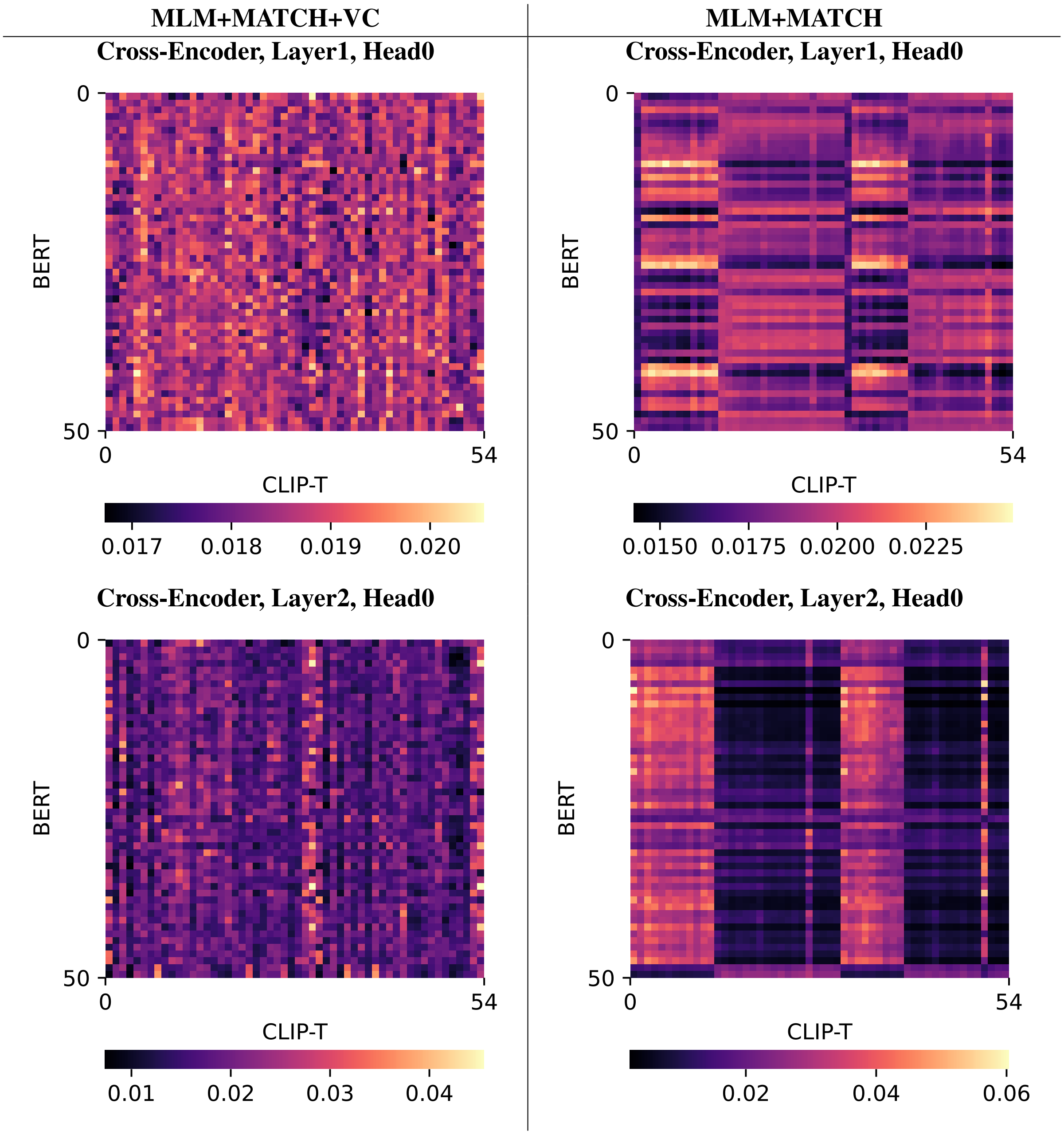

Table 9: Attention map of the cross-attention layers.different experiments. Left: Trained with visual classification loss, Right : Trained without visual classification loss. When trained with VC loss, the different tokens of BERT attends to the different tokens of CLIP-T more diversely.

BERT sequence : [’[CLS]’, ’scientists’, ’had’, ’observed’, ’that’, ’mice’, ’with’, ’a’, ’defective’, ’k’, ’##lot’, ’##ho’, ’gene’, ’aged’,’prematurely’, ’and’, ’wondered’, ’if’, ’an’, ’enhanced’, ’gene’, ’would’, ’have’, ’an’, ’opposite’, ’effect’, ’.’, ’[SEP]’, ’scientists’, ’have’, ’discovered’, ’a’, ’gene’, ’that’, ’produces’, ’a’, ’hormone’, ’that’, ’raises’, ’the’, ’life’, ’expect’, ’##ancy’, ’in’, ’mice’, ’by’, ’30’, ’percent’, ’.’, ’[SEP]’]

CLIP-T sequence : [’<|startoftext|>’, ’scientists’, ’had’, ’observed’, ’that’, ’mice’, ’with’, ’a’, ’defe’, ’ctive’, ’klo’, ’tho’, ’gene’, ’aged’, ’pre’, ’matu’, ’rely’, ’and’, ’wondered’, ’if’, ’an’, ’enhanced’, ’gene’, ’would’, ’have’, ’an’, ’opposite’, ’effect’, ’.’, ’<|endoftext|>’,’<|startoftext|>’, ’scientists’, ’have’, ’discovered’, ’a’, ’gene’, ’that’, ’produces’, ’a’, ’hormone’, ’that’, ’raises’, ’the’, ’life’, ’expect’, ’ancy’, ’in’, ’mice’, ’by’, ’3’, ’0’, ’percent’, ’.’, ’<|endoftext|>’]