InterBERT: An Effective Multi-Modal Pretraining Approach via Vision-and-Language Interaction

Junyang Lin[*1], **An Yang**[*1,2], **Yichang Zhang**[1], **Jie Liu**[1], **Jingren Zhou**[1], **Hongxia Yang**[†1]
[1] Alibaba Group
junyang.ljy, yichang.zyc, sanshuai.lj, jingren.zhou, yang.yhx@alibaba-inc.com
yangan@pku.edu.cn

Abstract

We propose a novel method for multi-modal pretraining, namely InterBERT (BERT for Interaction). The proposed architecture owns a strong capability of modeling interaction between the information flows of different modalities. The single-stream interaction module is capable of effectively processing information of multiple modalities, and the two-stream extraction module on top preserves the independence of each modality to avoid significant performance downgrade in single-modal tasks. The proposed pretraining task called masked group modeling (MGM) includes masked segment modeling and masked region modeling. It encourages the model to model a span or region instead of a single word or object, and it requires the model to learn from the general context. We pretrain the model with MGM and the conventional image-text matching, and finetune it on a series of vision-and-language downstream tasks, including caption-based image retrieval, zero-shot image retrieval, and visual commonsense reasoning. Experimental results demonstrate that InterBERT outperforms a series of strong baselines, including the most recent multi-modal pretraining methods. The analysis shows that the proposed MGM is effective for pretraining, and our method for multi-modal pretraining can adapt to single-modal tasks without significant performance decrease in comparison with the BERT-base model.

1 Introduction

Pretraining has raised much attention in the community due to its strong capability of generalization and efficient usage of large-scale data. The development of computer vision has been highly connected with pretraining, such as AlexNet [17], VGG [37] and ResNet [12], which are pretrained on the large-scale dataset ImageNet [9] for image classification. Recent years have witnessed the burst of pretraining in natural language processing. Pretrained models [29; 14; 10; 30; 22; 51; 11] have reached state-of-the-art performances in many downstream tasks of natural language processing (NLP), including question answering [31], natural language inference [45], and even natural language generation, such as neural machine translation [42; 4; 44] and abstractive summarization [33; 26].

Such significant progress in this field raised the concern of pretraining for task-agnostic multi-modal representation. A series of cross-modal pretraining methods were proposed, and the self-supervised learning provides the models with a strong ability to adapt to multiple multi-modal downstream tasks through finetuning [41; 20; 43; 24; 39; 7; 19]. However, these models are mostly pretrained by simple tasks such as masked language/object modeling (MLM/MOM) and image-text matching (ITM). Except for that, single-stream models [39; 7; 19] simply apply BERT and mix information

[*] Equal contribution. This work is done when An Yang is an intern at Alibaba Group.

[†] Corresponding author.

Preprint. Under review.

from two streams into one model, while two-stream models [24; 43] can only build interaction with co-attention, where there is no self attention to the self-context in each layer of co-attention.

Motivated by this observation, we propose a novel method for multi-modal pretraining, called **InterBERT**, which refers to ***BERT** for **Inter**action*. The proposed architecture consists of a single-stream interaction module for all the inputs from different modalities, as well as a two-stream extraction module that processes information from each modality separately. This architecture ensures sufficient interaction between modalities and generates contextualized mode representations. Besides, we pretrain the model with the conventional image-text matching and our proposed masked group modeling (MGM). The task is more challenging as it forces the model to predict a span or a region, which requires the model to build a stronger connection between modalities.

We pretrain our InterBERT on a series of large-scale datasets of image-text pairs, and we evaluate the effects of InterBERT on several multi-modal downstream tasks, including caption-based image retrieval [46], zero-shot caption-based image retrieval, and visual commonsense reasoning [53]. Experimental results demonstrate that our method can achieve significant improvements over the baseline models, and it outperforms or rivals the recent multi-modal pretrained models. We also evaluate the effects of our pretraining tasks and the model performance in single-modal tasks. The analysis demonstrates that our pretraining tasks can enhance model performance, and the model can adapt to single-modal tasks without significant performance decrease in comparison with the BERT-base model. We also find that the weight initialization for pretraining can make a difference in the finetuning of certain downstream tasks.

2 Approach

2.1 Background

We first introduce the background of NLP pretraining. Given an input text, a word (a character or a sub-word) sequence with a classification token and separation tokens $w = \{[\texttt{CLS}], w_1, w_2, \cdots, w_n, [\texttt{SEP}]\}$ of length $n + 2$, the model should learn to generate its high-level representations $h = \{h_{[\texttt{CLS}]}, h_1, h_2, \cdots, h_n, h_{[\texttt{SEP}]}\}$. The input can be a sequence of multiple sentences, where they are separated by "[SEP]". For the word embedding, except for the embedding layer, positional embedding and segment embedding are applied to denote the word positions and the segments they are from. For a BERT of l layers, the model can produce l sequences of representations $H = \{H^1, H^2, \cdots, H^l\}$. In most cases, we refer h to H^l. For further finetuning, the pretrained model except for the topmost layer for logits is applied as the backbone of the model for a specific downstream task.

Following this logic, such pretraining can be extended to learning multi-modal representations. In this work, we focus on the pretraining of vision and language. The dataset for multi-modal pretraining consists of paired image-text data, such as an image and its caption. To feed an image into BERT, a solution is to extract the object representations and bounding boxes with a detector, such as Faster-RCNN [32], and form a sequence of m object representations $o = \{o_1, o_2, \cdots, o_m\}$ together with their positions. Similar to the pretraining in NLP, we also add a representation $o_{[\texttt{CLS}]}$, which is the mean pooling of the original o in our implementation. The goal of the model is to learn the high-level representations of both image and text $h = \{h^i, h^t\}$, where $h^i = \{h^i_{[\texttt{CLS}]}, h^i_1, \cdots, h^i_m\}$ and $h^t = \{h^t_{[\texttt{CLS}]}, h^t_1, \cdots, h^t_n, h^t_{[\texttt{SEP}]}\}$.

2.2 Model overview

In this section, we illustrate the details of our proposed model InterBERT. The overview of the architecture is demonstrated in Figure 1. The simplest solution for multimodal pretraining is to pretrain a BERT-like model with the concatenation of image and text features. Lu et al. [24] pointed out that such a method of information fusing ignores the different requirements of processing for different modalities, and their experimental results show that the two-stream model outperforms the single-stream one in multiple tasks. We view that the effective interaction of modalities is the key to effective pretraining. Such interaction requires the gap bridging between image and text and the maintenance of the independence of each modality. Furthermore, an extra benefit of such independence enables transfer to both cross-modal downstream tasks and single-modal tasks. This enhances the robustness of the model and breaks the limitation of the form of pretraining data.

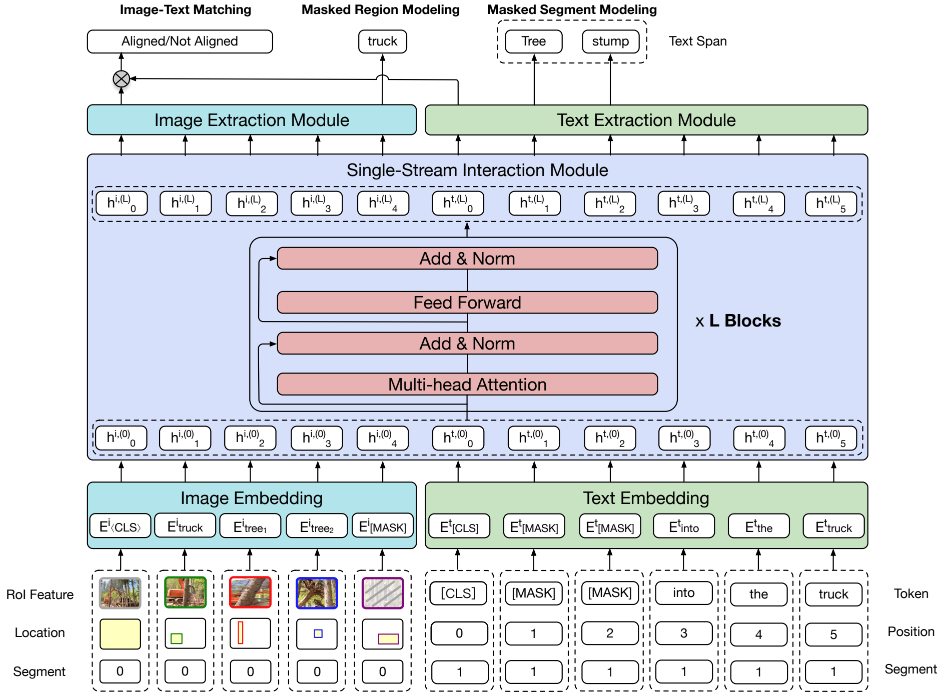
Image-Text Matching
Masked Region Modeling
Masked Segment Modeling
Aligned/Not Aligned
truck
Tree
stump
Text Span
Image Extraction Module
Text Extraction Module
Single-Stream Interaction Module
$h^{i,(L)}_0$ $h^{i,(L)}_1$ $h^{i,(L)}_2$ $h^{i,(L)}_3$ $h^{i,(L)}_4$ $h^{t,(L)}_0$ $h^{t,(L)}_1$ $h^{t,(L)}_2$ $h^{t,(L)}_3$ $h^{t,(L)}_4$ $h^{t,(L)}_5$
Add & Norm
Feed Forward
Add & Norm
Multi-head Attention
x L Blocks
$h^{i,(0)}_0$ $h^{i,(0)}_1$ $h^{i,(0)}_2$ $h^{i,(0)}_3$ $h^{i,(0)}_4$ $h^{t,(0)}_0$ $h^{t,(0)}_1$ $h^{t,(0)}_2$ $h^{t,(0)}_3$ $h^{t,(0)}_4$ $h^{t,(0)}_5$
Image Embedding
$E^i_{\langle CLS \rangle}$ E^i_{truck} $E^i_{tree_1}$ $E^i_{tree_2}$ $E^i_{[MASK]}$
Text Embedding
$E^t_{[CLS]}$ $E^t_{[MASK]}$ $E^t_{[MASK]}$ E^t_{into} E^t_{the} E^t_{truck}
RoI Feature
Location
Segment
0 0 0 0 0
[CLS] [MASK] [MASK] into the truck
0 1 2 3 4 5
1 1 1 1 1 1
Token
Position
Segment

Figure 1: An overview of the architecture of InterBERT. The model is built with an image embedding layer, a text embedding layer, a single-stream interaction module, and a two-stream extraction module.

Replacing co-attention with all-attention While co-attention demonstrates effects in VilBERT [24], we find that such a method of modal interaction limits the capability of the model. The representations of one modality can only attend to those of the other one, ignoring the self-context. The ideal attention should be one that attends to the whole context. Here we replace the co-attention with all-attention, which is a single-stream interaction module based on multi-head self attention (MHSA) and point-wise feed-forward neural network (FFN) [44; 10]. The input of the single-stream interaction module is the concatenation of image and text embeddings, and thus the attention can attend to the whole context of both modalities. The layer includes:

$$h^l = \mathtt{MHSA}(\mathbf{W_q} x^{l-1}, \mathbf{W_k} x^{l-1}, \mathbf{W_v} x^{l-1}), \tag{1}$$

$$\tilde{h^l} = \mathtt{LN}(x^{l-1} + h^l), \tag{2}$$

$$\hat{h^l} = \mathbf{W_2}[\mathtt{GeLU}(\mathbf{W_1} \tilde{h^l} + b_1)] + b_2, \tag{3}$$

$$x^l = \mathtt{LN}(\tilde{h^l} + \hat{h^l}), \tag{4}$$

where x_{l-1} is the whole context of image and text representations, instead of representations of a single modality. For multi-head attention, the model first transforms the inputs to query, key and value representations with weight matrices $\mathbf{W_q}$, $\mathbf{W_k}$, and $\mathbf{W_v}$, and split them into multiple heads and compute the attention scores of query and key as well as the weighted sum of the value. Layer normalization and residual connection are applied, and the activation function is GeLU [13].

This architecture enables strong interaction between modalities with the attention mechanism. Compared with the two-stream co-attention layer [24] which can only attend to the representations of the other modality, this architecture enables a combination of self attention and co-attention, and therefore the model can generate more contextualized representations. Furthermore, another advantage is that the architecture is identical to BERT and thus its weights can be initialized with the pretrained BERT's weights, which improves the availability of the previous pretrained models.

Extraction module for mode representations An ideal situation is that the model's outputs consist of visual and linguistic representations as well as the visual-linguistic ones. Also, a robust multimodal pretrained model should have the capability to transfer to single-modal tasks. As mentioned above, the single-stream interaction module fuses the visual and linguistic representations and make them more contextualized. In concern of the extraction of representations of each modality, we should develop a module to respectively generate representations to separate the fused information.

We implement a two-stream extraction module, which consists of an image extractor and a text extractor. Each extractor is based on self attention and FFN. The module is responsible for generating high-level object representations and text representations. Except for these, the model generates a general image representation and text representation for finetuning. The image and text representations are transformed into a cross-modal representation by a multi-layer feed-forward network. To validate our hypothesis, we analyze by finetuning our pretrained model and the single-stream multi-modal pretrained model (a simple BERT architecture) on natural language processing tasks to evaluate their performances on single-modal tasks. The analysis demonstrates that our architecture can achieve similar performance compared with the original BERT-base model, while the single-stream model without the two-stream extraction module performs much worse. This shows our model's advantage in preserving modal independence. More details are described in Section 3.6.

Text embedding and image embedding Following Devlin et al. [10], we tokenize the input text with WordPiece [50] and embed each word with an embedding layer. Positional embedding is required for the self-attention-based model to obtain the positional information, and segment embedding is required for the model to distinguish image and text.

A solution to adapt the image to Transformer is to obtain the object representations and their locations with a detector. Following Lu et al. [24], we apply a commonly used object detector Faster-RCNN [32] trained on Visual Genome [16; 2]. We extract the bounding boxes and the RoI (Region of Interest) features as the object representations. Similar to the aforementioned process, we apply embedding, positional embedding, and segment embedding to the extracted features.

2.3 Pretraining tasks

In this section, we introduce the pretraining tasks for our multi-modal pretraining, namely masked group modeling (MGM) and image-text matching (ITM).

Masked group modeling We propose MGM, which encourages the model to predict the masked groups of images and texts. We name the masked group modeling on text "masked segment modeling (MSM)", and the masked group modeling on image "masked region modeling (MRM)". Similar to MLM, MSM also replaces the selected words with the same strategy (replacing with the token "`[MASK]`", a random word or the original word). However, MSM masks a continuous segment of text instead of random words. Different from Joshi et al. [15], we mask multiple segments for each sample. As to MRM, it masks selected objects with zero vectors as MOM does. Yet, it endeavors to mask objects that are immediate to avoid information leakage due to the overlapping between objects. MRM masks objects which have a high proportion of mutual intersection.

For MSM, we randomly choose words as masking anchors by the probability of 10%, and we randomly mask the anchors and 0 to 2 words after the anchors by the probability of uniform distribution. For MRM, we also randomly choose objects as masking anchors by the probability of 10%, and we mask the objects whose IoUs with the anchors are larger than 0.4. The objective of the model is to predict the masked words and the categories of the masked objects. The training minimizes the loss:

$$\mathcal{L}_{\mathbf{MSM}} = -\mathbb{E}_{x\sim D} \log p\left(\overline{x}|\hat{x}\right) \approx -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathbf{m_t}(x^n, t) \log p_\theta\left(x_t^n|\hat{x}^n\right), \tag{5}$$

$$\mathcal{L}_{\mathbf{MRM}} = -\mathbb{E}_{x\sim D} \log p\left(\overline{x}|\hat{x}\right) \approx -\frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T} \mathbf{m_i}(x^n, t) \log p_\theta\left(x_t^n|\hat{x}^n\right), \tag{6}$$

where x is a random sample of image-text pair from the training set D, and $\overline{x}$ refers to the masked segment or the masked region, and $\hat{x}$ refers the whole masked sequence x. $\mathbf{m_i}$ and $\mathbf{m_t}$ refer to the masking functions for image and text. The objective functions encourage the model to predict the masked groups of words or the class of the masked groups of objects.

Image-text matching For learning the relation between image and text, we regard the image-text pairs in the dataset as positive samples, and we pair the images with randomly selected texts and regard the pairs as negative samples. Both the positive and negative samples share the same proportion in the training set. We add a simple MLP on top of the main architecture for computing the matching score between inputs of two modalities. Specifically, we first element-wisely multiply the image and text representations (the output representations at the position of "[CLS]") and send the generated representation through the MLP for the matching score. The training minimizes the cross-entropy loss:

$$\mathcal{L}_{\mathbf{ITM}} = -\mathbb{E}_{x,y\sim D}\left[y \log p\left(y|\hat{x}\right) + (1-y)\log\left(1-p\left(y|\hat{x}\right)\right)\right], \tag{7}$$

where x is a random sample from the training set D and $y \in \{0,1\}$ denotes whether x is positive or negative. $\hat{x}$ refers to the masked x.

The overall objective function is the weighted sum of the aforementioned terms, as shown below:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\mathbf{MSM}} + \lambda_2\mathcal{L}_{\mathbf{MRM}} + \lambda_3\mathcal{L}_{\mathbf{ITM}}, \tag{8}$$

where λ refers to the hyperparameter for the weights for each term.

2.4 Finetuning

We use the pretrained InterBERT as the backbone for the downstream tasks. We apply the pretrained model to three downstream tasks, including caption-based image retrieval, zero-shot caption-based image retrieval, and visual commonsense reasoning. Finetuning is simple as we can add simple MLP layers based on the requirements of the corresponding downstream tasks.[1]

3 Experiments

In this section, we provide an introduction to our experimental details, and we demonstrate the results as well as the analysis.

3.1 Pretraining datasets

For Flickr30K and VCR, we pretrain our model on the combination of three datasets, including Conceptual Caption (CC) [36], SBU Captions [27], and COCO captions [21].[2]

3.2 Downstream tasks

Caption-based image retrieval Caption-based image retrieval requires the model to retrieve an image from a large pool of images based on a given caption. We conduct experiments on Flickr30K [52], whose images are extracted from Flickr.[3] In Flickr30K, each image is paired with five captions, which are of relatively high quality. Following Lu et al. [24], in the stage of training, we change the task to 4-way multiple choice by adding three negative images for each image-caption pair. The training set contains 29K images, and the validation and test set contain 1K images respectively. The evaluation metrics are R@1, R@5, and R@10 (recall at 1, 5, and 10).

Zero-shot caption-based image retrieval This is the zero-shot setting for caption-based image retrieval. The model performs caption-based image retrieval without finetuning on the training data. This challenges the capability of the pretrained model to understand the relations between image and text. We use the same splits of the dataset and the same evaluation metrics as those of caption-based image retrieval.

Visual commonsense reasoning Visual commonsense reasoning (VCR) is a task connected with cognition and requires visual understanding [53]. There are three sub-tasks in VCR, including Q→A, QA→R, and Q→AR. Q→A refers to providing the answer based on the given image and question, QA→R refers to providing the rationale based on the given image, question, and answer. In Q→AR, provided an image and a question, the model should not only answer the question but also give the

[1] We demonstrate more implementation details about finetuning in Appendix 6.2.

[2] More details of pretraining data are in Appendix 6.1.

[3] `https://www.flickr.com`

Table 1: Results of the models on the three downstream tasks. The results of the baselines are those reported in their original papers. "-" denotes that the model was not implemented on the task in the original work. "w/o pt" refers to "without pretraining".

Models	IR			Zero-shot			VCR		
	R@1	R@5	R@10	R@1	R@5	R@10	Q→A	QA→R	Q→AR
SCAN	48.6	77.7	85.2	-	-	-	-	-	-
R2C	-	-	-	-	-	-	63.8	67.2	43.1
VisualBERT	-	-	-	-	-	-	70.8	73.2	52.2
VilBERT	58.2	84.9	91.5	31.9	61.1	72.8	72.4	74.5	54.0
VL-BERT	-	-	-	-	-	-	**73.8**	74.4	54.2
InterBERT (w/o pt)	53.1	80.6	87.9	-	-	-	63.6	63.1	40.3
InterBERT	**61.9**	**87.1**	**92.7**	**49.2**	**77.6**	**86.0**	73.1	**74.8**	**54.9**

correct rationale for the choice. For each question, there are 4 candidate answers and 4 candidate rationales. The training set contains 80K images and 213K questions, the validation set contains 10K images and 27K questions, and the test set contains 10K images and 25K questions. We apply accuracy score as the evaluation metric.

3.3 Baselines

For the comparison with the previous methods, we mainly compare our InterBERT with the previous models that achieved outstanding performances on the downstream tasks as well as the recent multi-modal pretrained models.

Previous methods For image retrieval, we compare InterBERT with SCAN [18], which is an architecture based on stacked cross-attention. For VCR, we compare InterBERT with R2C (Recognition to Cognition) [53], which contains modules for grounding, contextualizing, and reasoning.

Multi-modal pretrained models We compare our model with some recent multi-modal pretrained models. Specifically, we focus on the comparison of our model with VilBERT and VL-BERT for the reason that our implementation details are similar to theirs, including pretraining datasets and the number of object features. Moreover, VilBERT and VL-BERT are regarded as powerful baselines of two-stream and single-stream multi-modal pretrained models, respectively[4].

3.4 Implementation details

For pretraining, we first extract the object representations of the images with a trained object detector. Specifically, the object representations and their bounding boxes are generated by an object detector based on Faster R-CNN [32] with a backbone of ResNet-101 [12], which is trained on Visual Genome [16]. We pretrain the model with AdamW [23] with an initial learning rate of $1e - 4$, $\beta_1 = 0.9$, $\beta_2 = 0.9999$, $e = 1 \times 10^{-6}$ and a weight decay of 0.01. For the finetuning on Flickr30K image retrieval, the maximum number of objects is 100 and the actual numbers are between 90 and 100. The model reuses the output layer of the pretraining ITM task to compute the matching scores. We finetune the model on 8 Nvidia V100 for 20 epochs with AdamW optimizer with an initial learning rate of 4×10^{-5} and apply a linear decay learning rate scheduler with a warm-up period of 10000 steps. For the finetuning on VCR, we use a smaller learning rate 2×10^{-5} and train the model for only 5 epochs.[5]

3.5 Results

Table 1 demonstrates the experimental results of our proposed model InterBERT as well as the compared baselines on the downstream tasks. In the experiment of image retrieval, InterBERT outperforms SCAN by a large margin (+13.3 (27.4%) in R@1, +9.4 (12.1%) in R@5, and +7.5 (8.8%)

[4] They have released the codes for pretraining and finetuning. Please refer to https://github.com/jiasenlu/vilbert_beta and https://github.com/jackroos/VL-BERT.

[5] For more information about the implementation details, please refer to Appendix 6.2.

Table 2: An ablation study of the training techniques conducted on the validation set of VCR.

Tasks	VCR Q→A	 QA→R	 Q→AR
MLM+MOM+ITM	72.3	74.3	54.0
MSM+MRM+ITM	73.1	74.8	54.9

in R@10), and it also outperforms VilBERT by +3.7 (6.4%) in R@1, +2.2 (2.6%) in R@5, and +1.2 (1.3%) in R@10. As for zero-shot image retrieval, the advantage is significantly larger. It outperforms VilBERT by +17.3 (54.2%) in R@1, +16.5 (27.0%) in R@5, and +13.2 (18.1%) in R@10). In the experiment of VCR, InterBERT also significantly outperforms the baseline R2C by +9.3 (14.6%) in Q→A, +7.6 (11.3%) in QA→R, and +11.8 (27.4%) in Q→AR, and it also outperforms VilBERT by +0.7 (1.0%) in Q→A, +0.3 (0.4%) in QA→R, and +0.9 (1.7%) in Q→AR. Compared with VL-BERT, InterBERT also outperforms by +0.7 (1.3%) on the overall Q→AR accuracy.

InterBERT has advantages over the baselines in the tasks, especially in zero-shot image retrieval. Also, compared with VilBERT, InterBERT has an advantage in the number of parameters (173M vs 221M), which reflects the effects of single-stream interaction. The significant advantage in zero-shot learning demonstrates that our model has a strong capability of modeling image-text relations and transferring to downstream tasks without finetuning.

Also, we directly train our InterBERT without multi-modal pretraining to evaluate the effects of pretraining for the downstream tasks. To be more specific, as our pretrained model is initialized with the weights of BERT-base, we also train the InterBERT without pretraining with the BERT initialization. From Table 1, the model without pretraining suffers from the performance degrade (IR: -8.8 (-14.2%) in R@1, -6.5 (-7.5%) in R@5, and -4.8 (-5.2%) in R@10; VCR: -9.5 (-13.0%) in Q→A, -11.7 (-15.6%) in QA→R, and -14.6 (-26.6%) in Q→AR). The effective multi-modal pretraining can significantly impact the model performance in downstream tasks.

3.6 Analysis

In this section, we conduct a series of analyses to evaluate the effects of MGM, the performance on single-modal downstream tasks, and the effects of weight initialization for pretraining.

The effects of MGM We conduct an ablation study on the validation set of VCR to evaluate the effects of MGM, which includes MSM and MRM. Specifically, we pretrain two models with different pretraining tasks, including MLM+MOM+ITM and MSM+MRM+ITM. Table 2 demonstrates the results of the evaluation. It can be found that our proposed MSM and MRM are beneficial to the pretraining effects. The model trained with MSM and MRM can outperform the baseline by $+0.8$ in Q→A, $+0.5$ in QA→R, and $+0.9$ in Q→AR. The model trained with our tasks gains a stronger ability of modeling image and text by understanding contexts and building a stronger connection so that it can reach better performance in a task that requires reasoning over image and text.

Performances in the single-modal tasks While multi-modal pretraining demonstrates effects in the aforementioned downstream tasks, it is still a question whether it still preserve the knowledge of single-modal representation and whether it can still achieve comparable performances in the single-modal tasks. To evaluate the model's robustness, we conduct an experiment on 8 tasks of GLUE [45], including QNLI, CoLA, SST-2, STS-B, RTE, MNLI, QQP, and MRPC.[6] We compare InterBERT with BERT-base and the single-stream multi-modal pretrained model (a simple BERT architecture). From Table 3, it can be found that InterBERT can achieve similar performances compared with BERT-base (Avg: 82.0 vs 81.8), and it significantly outperforms the single-stream model without the two-stream extraction module (Avg: 81.8 vs 80.0). This indicates that InterBERT with the two-stream extraction module preserves the ability to model single-modal representations and it can adapt to single-modal downstream tasks without significant performance decrease.

The effects of initialization In our experiments, we surprisingly find that the different weight initialization for pretraining has different impacts on the finetuning on different downstream tasks.

[6] We provide the details of the GLUE datasets in Appendix 6.3.

Table 3: Results on the GLUE dev set. We evaluate the performance of the BERT-base model, a single-stream model, and InterBERT on 8 tasks of GLUE. The results show that InterBERT can rival the BERT-base model in the tasks of natural language understanding. We report F1 scores for QQP and MRPC, Spearman correlations for STS-B, and accuracy scores for the rest.

Model	QNLI	CoLA	SST-2	STS-B	RTE	MNLI (m/mm)	QQP	MRPC	Avg.
BERT-base	**91.5**	56.7	**93.2**	88.2	**65.0**	83.7 / **84.1**	87.9	**89.6**	**82.0**
Single Stream	90.8	52.3	91.6	88.6	59.2	82.6 / **84.1**	87.8	86.4	80.0
InterBERT	91.1	**57.3**	92.3	**88.9**	64.3	**84.1** / 83.7	**88.1**	88.6	81.8

As mentioned above, we initialize a part of our model with the weights of the pretrained BERT-base model. Here we compare the models with or without BERT initialization on the performances on the downstream tasks. While we find that such initialization has little impacts on the retrieval tasks, we also find that the initialization is surprisingly significant to the finetuning on VCR. The model without BERT initialization suffers from a severe performance downgrade. Its performances in Q→A and QA→R are 65.3 (-10.7%) and 64.4 (-13.9%), and VilBERT without BERT initialization performs worse (61.7 in Q→A and 59.7 QA→R). This demonstrates the importance of the pretrained NLP model on the multi-modal tasks concerned with reasoning as it can improve the effects of text processing and thus enhance its language understanding capability. It also shows that sufficient interaction between modals through all-attention can alleviate the problem. However, this can be a starting point for the research in the effect of initialization on multi-modal pretraining.

4 Related work

The success of pretraining in NLP raised the attention in multi-modal pretraining. VideoBERT [41] is regarded as the first work in multi-modal pretraining. It is a model pretrained on the extracted video frame features and texts. One of the following studies is CBT [40], which is also pretrained on video-text pairs. Inspired by the starting work in multi-modal pretraining, more researchers have turned their focus to visual-linguistic pretraining. There are mainly two streams of model architectures for this task. One is the single-stream model [1; 7; 19; 20; 39]. Li et al. [19] processed the concatenation of objects and words with a BERT model and pretrain it with the three conventional tasks. Chen et al. [7] proposed a similar method but with more pretraining tasks. Su et al. [39] used identical architecture but they pretrained the object detector and added single-modal data. The other is the two-stream model [43; 24; 25]. Tan and Bansal [43] proposed a two-stream model with co-attention and pretrained the model only with the in-domain data. Lu et al. [24] proposed a similar architecture with a more complex co-attention, and pretrained the model with the out-of-domain data, and Lu et al. [25] further improved VilBERT with multi-task learning. In this work, we simply focus on the design of architecture and pretraining tasks. The single-stream models mostly apply BERT to multi-modal pretraining in a straightforward fashion, while the two-stream models have respective encoders for modalities and a co-attention module for the cross-modal interaction. These models either lack the independence of each modality or lack sufficient interaction across modalities. Furthermore, there is still room for setting training tasks for more effective pretraining. Compared with the previous work, our proposed method has several significant differences. Our proposed model architecture is effective in capturing modal interaction with an all-attention-based module and obtaining modal independence with the two-stream extraction module. Besides, our proposed masked group modeling improves the model's ability to predict a span or a region, so that the model can be more effective.

5 Conclusion

In this paper, we propose a new approach for multi-modal pretraining, InterBERT. The model architecture consists of a single-stream interaction module for sufficient interaction between the information of different modes, and a two-stream extraction module for the separation of modal information to preserve the ability to transfer to single-modal tasks. Furthermore, to strengthen its ability of modeling image and language, we pretrain the model with the tasks of the conventional image-text matching as well as our proposed masked group modeling. Experimental results demonstrate that our InterBERT can outperform the baselines and rival the recent multi-modal pretrained models in the downstream tasks. The analyses show that the pretraining tasks can enhance the model performance, and InterBERT can adapt to single-modal tasks without significant performance downgrade. Also, we find out that the weight initialization for pretraining makes a difference to some downstream tasks, specifically VCR. We hope this study can provide some insights into multi-modal pretraining, and in the future, we will endeavor to figure out better model architecture and training tasks for the improvement in multi-modal representation learning.

Broader Impact

This work of multi-modal pretraining can be adaptive to a great number of downstream tasks concerning multiple modalities, and therefore it can produce significant positive outcomes. For example, it has the potential in highly improving the effects of detecting terrorism, pornography, anti-society, hate speech, etc. Also, it can alleviate the Matthew's effect on the items in today's recommender system. Due to the learning of statistical information, it is easy for the recommender system based on collaborative filtering to suffer from the Matthew's effect. However, the learning of multi-modal information can alleviate this problem though it still has potential in biasing on certain patterns of cross-modality. Since our work leverages the pretrained BERT to initialize the InterBERT model, the gender and moral bias lurked in BERT should be an issue to be considered. Recent works [34; 8] presented encouraging analysis results, which show that BERT representations may be less biased compared with the previous embedding techniques on several metrics. Meanwhile, in the future, we will further consider the existing debiasing approaches on pretrained language models [54] to better circumvent the problem.

References

[1] C. Alberti, J. Ling, M. Collins, and D. Reitter. Fusion of detected objects in text for visual question answering. 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 2019, Hong Kong, China, November 3-7, 2019, pages 2131–2140, 2019.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pages 6077–6086, 2018.

[3] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, pages 6077–6086, 2018.

[4] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In 3rd International Conference on Learning Representations, ICLR 2015, 2015.

[5] L. Bentivogli, B. Magnini, I. Dagan, H. T. Dang, and D. Giampiccolo. The fifth PASCAL recognizing textual entailment challenge. In Proceedings of the Second Text Analysis Conference, TAC 2009, 2009.

[6] D. M. Cer, M. T. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, pages 1–14, 2017.

[7] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. UNITER: learning universal image-text representations. CoRR, abs/1909.11740, 2019.

[8] C. B. M. R. Costa-juss and N. Casas. Evaluating the underlying gender bias in contextualized word embeddings. GeBNLP 2019, page 33, 2019.

[9] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), pages 248–255, 2009.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova. 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, NAACL-HLT 2019, pages 4171–4186, 2019.

[11] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H. Hon. Unified language model pre-training for natural language understanding and generation. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, pages 13042–13054, 2019.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pages 770–778, 2016.

[13] D. Hendrycks and K. Gimpel. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR, abs/1606.08415, 2016.

[14] J. Howard and S. Ruder. Universal language model fine-tuning for text classification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, pages 328–339, 2018.

[15] M. Joshi, D. Chen, Y. Liu, D. S. Weld, L. Zettlemoyer, and O. Levy. Spanbert: Improving pre-training by representing and predicting spans. CoRR, abs/1907.10529, 2019.

[16] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1):32–73, 2017.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012., pages 1106–1114, 2012.

[18] K. Lee, X. Chen, G. Hua, H. Hu, and X. He. Stacked cross attention for image-text matching. In Computer Vision - ECCV 2018 - 15th European Conference, pages 212–228, 2018.

[19] G. Li, N. Duan, Y. Fang, D. Jiang, and M. Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. CoRR, abs/1908.06066, 2019.

[20] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang. Visualbert: A simple and performant baseline for vision and language. CoRR, abs/1908.03557, 2019.

[21] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. In Computer Vision - ECCV 2014, pages 740–755, 2014.

[22] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. CoRR, abs/1907.11692, 2019.

[23] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019.

[24] J. Lu, D. Batra, D. Parikh, and S. Lee. 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, pages 13–23, 2019.

[25] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee. 12-in-1: Multi-task vision and language representation learning. CoRR, abs/1912.02315, 2019.

[26] R. Nallapati, B. Zhou, C. N. dos Santos, Ç. Gülçehre, and B. Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, pages 280–290, 2016.

[27] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, pages 1143–1151, 2011.

[28] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[29] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, pages 2227–2237, 2018.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9, 2019.

[31] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100, 000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, pages 2383–2392, 2016.

[32] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, pages 91–99, 2015.

[33] A. M. Rush, S. Chopra, and J. Weston. A neural attention model for abstractive sentence summarization. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, pages 379–389, 2015.

[34] P. Schramowski, C. Turan, S. Jentzsch, C. Rothkopf, and K. Kersting. Bert has a moral compass: Improvements of ethical and moral values of machines. arXiv preprint arXiv:1912.05238, 2019.

[35] L. Sharma, L. Graesser, N. Nangia, and U. Evci. Natural language understanding with the quora question pairs dataset. CoRR, abs/1907.01041, 2019.

[36] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, pages 2556–2565, 2018.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In 3rd International Conference on Learning Representations, ICLR 2015, 2015.

[38] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, pages 1631–1642, 2013.

[39] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai. VL-BERT: pre-training of generic visual-linguistic representations. CoRR, abs/1908.08530, 2019.

[40] C. Sun, F. Baradel, K. Murphy, and C. Schmid. Contrastive bidirectional transformer for temporal representation learning. CoRR, abs/1906.05743, 2019.

[41] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid. Videobert: A joint model for video and language representation learning. CoRR, abs/1904.01766, 2019.

[42] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, pages 3104–3112, 2014.

[43] H. Tan and M. Bansal. 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 2019, pages 5099–5110, 2019.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, pages 5998–6008, 2017.

[45] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In 7th International Conference on Learning Representations, ICLR 2019, 2019.

[46] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, pages 5005–5013, 2016.

[47] A. Warstadt, A. Singh, and S. R. Bowman. Neural network acceptability judgments. TACL, 7: 625–641, 2019.

[48] A. Williams, N. Nangia, and S. R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, pages 1112–1122, 2018.

[49] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew. Huggingface's transformers: State-of-the-art natural language processing. CoRR, abs/1910.03771, 2019.

[50] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR, abs/1609.08144, 2016.

[51] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, pages 5754–5764, 2019.

[52] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. TACL, 2:67–78, 2014.

[53] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, pages 6720–6731, 2019.

[54] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang. Gender bias in contextualized word embeddings. arXiv preprint arXiv:1904.03310, 2019.

Table 4: Data statistics of the datasets for pretraining. The numbers in the parentheses refer to the numbers of images.

Datasets	Training	Validation
Conceptual Caption	3.3M	14K
SBU	890K	10K
COCO	587K (117K)	15K (3K)

Table 5: Data statistics of the datasets of the downstream tasks. "i" refers to the number of images, and "t" refers to the number of texts.

Datasets	Training	Validation	Testing
Flickr30K	i:29K, t:145K	i:1K, t:5K	i:1K, t:5K
VCR	i:80K, t:213K	i:10K, t:27K	i:10K, t:25K

6 Appendix

6.1 Data statistics

Pretraining datasets The image caption datasets for pretraining are Conceptual Caption (CC) [36], SBU Captions [27] and COCO Captions [21]. The detail data statistics are demonstrated in Table 4. In CC and SBU, each image is paired with a text as its description, while in COCO, there are around 5 texts that describe the same text. We also provide the number of images in COCO in the table.

Downstream datasets We demonstrate the detail data statistics of the datasets for finetuning in Table 5. The numbers of image and text of each dataset are provided.

6.2 Implementation details

In the following, we introduce the details of our implementation in pretraining and finetuning on each downstream task, including the model architecture, optimizer, hyperparameters, etc.

Pretraining Here we provide the experimental details about our implementation for pretraining. The object representations of the images as well as their bounding boxes are generated by an object detector based on Faster R-CNN [32] with a backbone of ResNet-101 [12], which is trained on Visual Genome [16]. This detector is applied for the bottom-up top-down attention model for image captioning [3], and we downloaded the pretrained detector from their provided link.[7] For the text processing, we tokenize the texts with BERT's tokenizer and directly use BERT-base's embedding layer for word embedding. The vocabulary size is 30522 and the embedding size is 768. For the consistency between word embedding and object representation, we transform the object representations of 2048 dimensions to 768 through MLP.

The hidden size of multi-head attention is also set to 768. The number of attention head is 12. For the FFN, both the input and output sizes are 768 for stacking layers, and the intermediate size is 3072. As to the LN layer inside each layer, we use BERT's LN with $e = 1e - 12$. The single-stream interaction module consists of 12 layers of Transformer layer, and its weight parameters are initialized with the pretrained BERT-base model. The two-stream independence module contains two Transformers for both modalities on top of the single-stream interaction module. Each has 6 layers of Transformer layer. The new weight parameters are randomly initialized based on the Gaussian distribution of zero mean and standard deviation of 0.02, following [10]. We pretrain the model with AdamW whose initial learning rate of $1e - 4$, $\beta_1 = 0.9$, $\beta_2 = 0.9999$, $e = 1 \times 10^{-6}$ and a weight decay of 0.01. We apply the linear decay learning rate scheduler with a warm-up period of 10000 steps. The batch size for training is 512. For the pretraining dataset of image-caption pairs, we pretrain our InterBERT on 8 V100 GPUs for 20 epochs.

[7] https://github.com/peteanderson80/bottom-up-attention

Finetuning For the finetuning on Flickr30K image retrieval, the maximum number of objects is 100 and the actual numbers are between 90 and 100. The model reuses the output layer of the pretraining ITM task to compute the matching scores. We finetune the model with a batch size of 32 and train it on 8 V100 for 20 epochs. We use AdamW optimizer with an initial learning rate of 4×10^{-5} and apply a linear decay learning rate scheduler with a warmup period of 10000 steps. We finetune the model for 20 epochs. For the finetuning on VCR, we use similar hyperparameters with those in the finetuning on Flickr30K, but we use a smaller learning rate 2×10^{-5} and a smaller batch size 32, and we only finetune the model for 5 epochs. Furthermore, we apply exponential moving average with a rate of 0.9999 on the finetuned models for the final model, so that it can be more robust and reach better performance in testing.

Hardware configuration The experiments are conducted on a Linux server equipped with an Intel(R) Xeon(R) Platinum 8163 CPU @ 2.50GHz, 512GB RAM and 8 NVIDIA V100-SXM2-16GB GPUs.

Software The experiments are implemented in python 3.6 and PyTorch 1.1.0 [28]. The code is based on Transformers [49].[8]

6.3 Details of the GLUE tasks

The GLUE benchmark [45] consists of a series of NLP tasks, including QNLI, CoLA, SST-2, STS-B, RTE, MNLI, QQP, and MRPC,. We use them to evaluate the robustness of InterBERT in single-modal downstream tasks.

QNLI Question Natural Language Inference is a binary classification task of SQuAD (Stanford Question Answering Dataset) [31; 45]. It requires the model to judge whether the given answer is a correct one of the given question in a sentence pair.

CoLA The Corpus of Linguistic Acceptability [47] is a task of binary sentence classification. It requires the algorithms to check whether an English sentence is linguistically acceptable (grammatical and consistent with the world knowledge).

SST-2 The Stanford Sentiment Treebank [38] is a task of binary sentiment classification. It requires the algorithms to check whether a sentence is positive or negative. The sentences are extracted from movie reviews with human annotations.

STS-B The Semantic Textual Similarity Benchmark [6] is a task of classification of semantic similarity. The sentences are extracted from news headlines and other sources. The algorithms should learn to score two sentences from 1 to 5 for their semantic similarity.

RTE Recognizing Textual Entailment [5] is a task of natural language inference. This task provides a sentence pair and requires the algorithms to check the relations between the sentences, including "entailment", "contraction" and "neutral".

MNLI Multi-Genre Natural Language Inference [48] is a task of entailment with a large dataset. It requires the algorithm to figure out the relation of a pair of sentences. The relations include "entailment", "contradiction", and "neutral".

QQP Quora Question Pairs [35] is a task to check if two questions are semantically identical. The questions are extracted from Quora[9].

MRPC Microsoft Research Paraphrase Corpus is a dataset of sentence pairs from news websites. The task is to check whether two sentences are semantically identical.

[8] https://github.com/huggingface/transformers

[9] `http://quora.com/`

We truncate the input texts to ensure that the maximum length is 128. The input texts are all lowercased. We use a batch size of 128 and a learning rate of 2×10^{-5}. We finetune the model on 8 Nvidia V100 GPUs with gradient accumulation for 3 epochs.