GROUNDING LANGUAGE REPRESENTATION WITH VISUAL OBJECT INFORMATION VIA CROSS MODAL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Previous studies of visual grounded language learning use a convolutional neural network (CNN) to extract features from the whole image for grounding with the sentence description. However, this approach has two main drawbacks: (i) the whole image usually contains more objects and backgrounds than the sentence itself; thus, matching them together will confuse the grounded model; (ii) CNN only extracts the features of the image but not the relationship between objects inside that, limiting the grounded model to learn complicated contexts. To overcome such shortcomings, we propose a novel object-level grounded language learning framework that empowers the language representation with visual object-grounded information. The framework is comprised of three main components: (i) ObjectGroundedBERT captures the visual-object relations and literary portrayals by cross-modal pretraining via a Text-grounding mechanism, (ii) Visual encoder represents a visual relation between objects and (iii) Cross-modal Transformer helps the Visual encoder and ObjectGroundedBERT learn the alignment and representation of image-text context. Experimental results show that our proposed framework consistently outperforms the baseline language models on various language tasks of GLUE and SQuAD datasets. [1]

1 INTRODUCTION

Grounded language learning is related to learning the meaning of language as it is applied to the present reality. People, particularly youngsters, acquire knowledge from not only pure textual information but also other modalities such as vision and sound, which contain rich data that cannot be captured by text alone. Most current pretrained language models (Devlin et al., 2019; Brown et al., 2020) are trained distinctly from text-based corpora, thus having the constraint in learning complex semantics that requires the blend of sign-in information through cross-referencing and synthesis.

Recently, there are many studies trying to improve the language representation with visual information (Lazaridou et al., 2015; Collell et al., 2017; Kiela et al., 2018; Bordes et al., 2019; Tan & Bansal, 2020). In those attempts, the whole image is usually encoded into feature maps using a CNN and then matched to the corresponding sentence or token representations. However, there are two main drawbacks to such an approach. First, while CNN can capture both low and high-level features of the individual objects in the image, it cannot apprehend the semantic relationships between the objects. As an example given in Fig. 1, the feature maps (red rectangles) can successfully capture the features of individual objects such as *"cat", "computer desk", "monitor"*, and *"chair"*, but it can not represent the relation: *"cat is standing on the computer desk"* as described in the caption. As such, the current grounded language models which use CNN features for learning will fail to map the implicit relations in the image to the facts given in the description sentence, limiting the grounded language to learn complex semantics.

Second, in most grounded language learning datasets, the whole image usually contains more information than the description sentence. For example, in Fig. 1, the image contains many objects such as *"monitor', "chair", "window"*, and *"lamp"* that are not mentioned in the description sentence.

[1] Our codes and data are available at https://github.com/... (the link is hidden now due to double-blind review)

Therefore, direct grounding the whole image information to the sentence representation will trigger many noises in the contextual representation of the language, causing the grounded model to learn incorrect contexts.

In this paper, we propose a novel grounded language learning framework that enriches the language representation with visual object-level grounded information such as object features, attributes, and positional information. In particular, the proposed framework consists of three main components, namely Text Encoder, Visual Encoder, and Cross-modal Transformer. Text Encoder is stretched out with a Text-grounding module to learn the visual-grounded representations of language from text-image pairs. Instead of using CNN feature maps, we propose to use features of detected objects as images representation and employ Transformer architecture ((Vaswani et al., 2017)) to learn the relationships between objects. A Cross-modal Transformer is then put on top to connect the language and visual modalities. By applying a pretrained learning strategy, our method not only precisely maps the information between text and image at the object-level but also learns the implicit relationships between the objects in the image, which CNN-based feature extractor can not handle. Finally, we join two components together and optimize the whole model as multi-tasks of Masked Language Modeling, Masked Visual Feature Prediction, and Image-Text Matching.

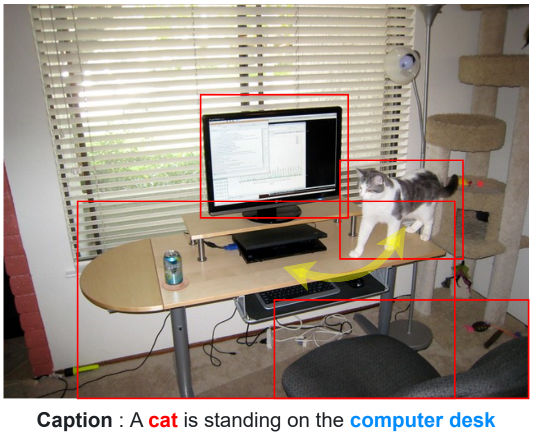

Figure 1: A grounded language learning example.

Our contributions can be summarized as:

- To the best of our knowledge, this study is the first to investigate the grounded language learning at the object-level with the rich grounded information containing object features, attributes, and positions. By doing so, we can enhance the ability of grounded language to capture more complex relations and avoid the confusion during the learning process.
- To this end, we propose a novel grounded language framework that enhances language representation with visual-objected-grounded information. Instead of using CNN to encode the whole image, we embed the features of objects from an off-the-shelf object detector into the encoder and connect them with the language modality via a cross-modal Transformer. A Text-grounding mechanism is also proposed to capture the visual object information and their relations found from the semantic correlation of words and images via multi-task pretraining strategy.
- We conduct extensive experiments on various language downstream tasks in GLUE and SQuAD datasets, and significantly outperforms the baselines on these tasks.

2 Related Work

Over the previous many years, many approaches have been proposed to learn language representation. Skip-gram (Mikolov et al., 2013), GLOVE (Pennington et al., 2014) were proposed to learn word representations. On the other hand, FastSent (Hill et al., 2016), QuickThought (Logeswaran & Lee, 2018), SkipThought (Kiros et al., 2015), Sentence-BERT (Reimers & Gurevych, 2019) or Le & Mikolov (2014); Conneau et al. (2017) tried to learn the sentence representations. Recently, many language models such as ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), GPT (Brown et al., 2020), ELECTRA (Clark et al., 2019), ALBERT (Lan et al., 2019) were proposed to learn the contextual relationships between words in a text. Those studies, however, only train the language representation with only textual corpora.

In recent years, many vision-and-language pre-trained models have been proposed to build joint cross-modal representations and focus on vision-and-language tasks such as visual question answering and natural language for visual reasoning. While Li et al. (2019); Chen et al. (2019); Su et al. (2020); Zhou et al. (2020); Li et al. (2020) use only one cross-modal Transformer for learning, Lu et al. (2019); Tan

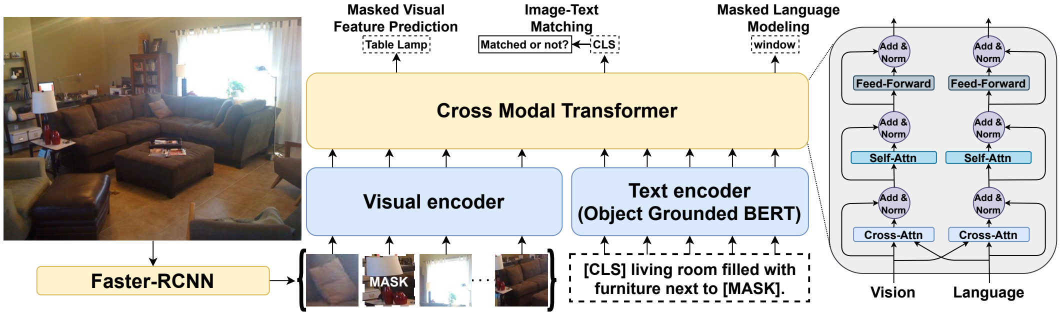

Figure 2: The architecture of our proposed framework, which consists of three components: Visual encoder, Cross Modal Transformer and Text encoder (ObjectGroundedBERT). The pretraining tasks are Masked Visual Feature Prediction, Image-Text Matching and Masked Language Modeling.

& Bansal (2019) proposed to use two single-modal Transformers and one cross-modal Transformer. Pretraining tasks such as masked language model and masked visual-feature classification were used in those studies to learn the vision-and-language representation.

On the other hand, there are a few attempts to improve language representation with visual information. Lazaridou et al. (2015) introduces multimodal skip-gram models (MMSKIP-GRAM) taking visual information into account. Collell et al. (2017) proposes IMAGINET which consists of GRU networks and tries to predict its visual feature map and the next word in the sentence. Kiela et al. (2018) uses bi-directional LSTM for sentence encoder; moreover, it aims to predict both the visual feature map and the other caption giving one caption. Bordes et al. (2019) proposes an intermediate space - grounded space and learns the visual and text representation with cluster information and perceptual information. Tan & Bansal (2020) introduces the vokenization process and pre-trains the language model with additional voken-classification task along with masked language modeling. These methods, however, only ground language models at the image level, i.e. mapping the whole image with its description sentence instead of between corresponding objects inside them.

3 Methodology

The overall architecture of our proposed framework is shown in Fig. 2, which comprises three main components: (i) Text encoder (ObjectGroundedBERT) to learn the object-level visual-grounded information from text-image pairs, (ii) Visual encoder to represent the visual representation, and (iii) Cross-modal Transformer to connect the visual and language modalities. The whole framework will be trained with a pretraining learning strategy.

3.1 Visual encoder

Object detection

We take the feature of objects as the embeddings of images. The off-the-shelf object detection model detects m objects $o_1, \ldots, o_m$ from the image. Each object contains region-of-interest (ROI) feature f_j and the region geometric feature p_j. The visual objected embedding is the sum of outputs of two FC layers of ROI feature and positional feature:

$$v_j = \left(FC_f(f_j)) + FC_p(p_j)\right)/2 \tag{1}$$

where v_j is visual objected embedding of the object j in the image, FC_f, FC_p are the FC layers for ROI feature and positional feature respectively.

Visual encoder

To learn the relationships between objects in the images, we employ Transformer encoder to visual objected embedding to obtain the object-relationship representations:

$$\hat{hv}_1, \hat{hv}_2, \ldots, \hat{hv}_m = VE(v_1, v_2, \ldots, v_m) \tag{2}$$

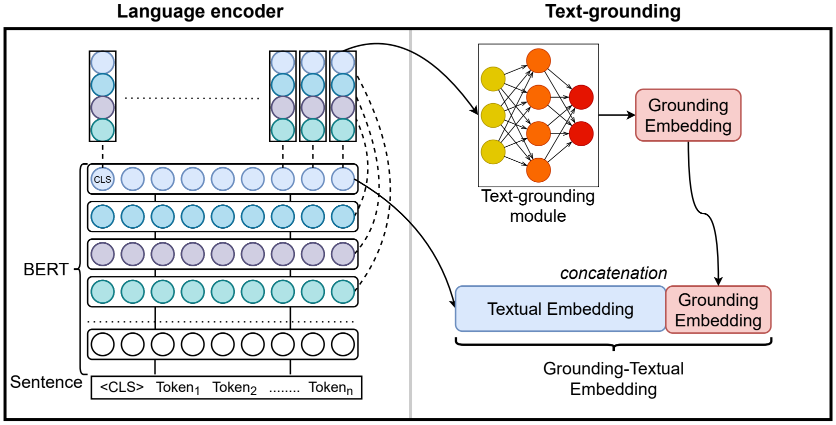

Figure 3: ObjectGroundedBERT comprises two components: Language encoder and Text-grounding. The final representation of the language model comprises textual and grounding embeddings.

where $v_j, \hat{hv_j}$ are the visual objected embedding and the object-relationship representation of the object j in the image respectively, m is the number of objects receives from the Object Detection model, VE stand for Visual encoder.

3.2 Text encoder (ObjectGroundedBERT)

Text encoder (ObjectGroundedBERT) is used to learn text representation of the input sentence together with the empowered grounding embedding. As shown in Fig. 3, ObjectGroundedBERT consists of a Language encoder and a Text-grounding. While we do not directly fine-tune the language model in order to keep the original knowledge learned from the huge language corpora, the final representation of the text encoder will be enhanced with the visual-grounded information learned during pretraining.

Language encoder We use BERT (Devlin et al., 2019) as the language encoder. Given an input sentence $s = \{w_1, \dots, w_n\}$, we use pretrained BERT model to contextually embed the discrete token w_i into hidden-output vector h_i:

$$h_1, h_2, \dots, h_n = BERT(w_1, w_2, \dots, w_n) \tag{3}$$

where $h_i = \{h_i^1, h_i^2, \dots, h_i^L\}$, h_i^l is the hidden-output vector of token w_i at layer l.

Text-grounding Module We concatenate hidden outputs from k final Transformer layers to form the input for the Text-grounding module:

$$\tilde{h}_i = [h_i^{L-k+1}, h_i^{L-k+2}, \dots, h_i^L] \tag{4}$$

The Text-grounding module uses a multi-layer perceptron with an activation function to transform the contextual representation of each token in the sentence into the grounding embedding:

$$g_i = MLP(\tilde{h}_i) \tag{5}$$

where g_i is grounding embedding of token i.

The grounding embedding and textual embedding, final hidden outputs of the language encoder, are concatenated to form a unified Grounding-textual representation of the token w_i:

$$\hat{hg}_i = [h_i^L, g_i] \tag{6}$$

3.3 Cross Modal Transformer

The Cross-model Transformer takes into account the outputs of Visual encoder and ObjectGrounded-BERT to connect the language and visual modalities based on the pretrained learning strategy, as shown in Fig. 2. Each cross-modal layer (the right block in Fig. 2) in the cross-modal Transformer consists of Cross-Attention layer, Self-Attention layer, and Feed-Forward layer. The Cross-Attention layer is shared for both modalities, while there are two Self-Attention layers and Feed-Forward layers for each modality. Cross-Modal Transformer is implemented by stacking these cross-modality layers. First, the cross-attention layer is applied to the vision (attention to language) and language (attention to vision):

$$\dot{hv}_1^k, \dot{hv}_2^k, \ldots, \dot{hv}_m^k = \text{Cross-Attn}\left(hv_1^{k-1}, hv_2^{k-1}, \ldots, hv_m^{k-1} | hg_1^{k-1}, hg_2^{k-1}, \ldots, hg_n^{k-1}\right) \tag{7}$$

$$\dot{hg}_1^k, \dot{hg}_2^k, \ldots, \dot{hg}_n^k = \text{Cross-Attn}\left(hg_1^{k-1}, hg_2^{k-1}, \ldots, hg_n^{k-1} | hv_1^{k-1}, hv_2^{k-1}, \ldots, hv_m^{k-1}\right) \tag{8}$$

where hg_i^k is the language representation of token i at layer k, hv_j^k is the visual objected representation of object j at layer k, ”Cross-Attn” stands for Cross-Attention layer.

The Cross-Attention layer is utilized to exchange the information and align the elements between the two modalities to learn cross representation. Then, Self-Attention layers are applied to the outputs of the Cross-Attention layer to earn their relationships:

$$\ddot{hv}_1^k, \ddot{hv}_2^k, \ldots, \ddot{hv}_m^k = \text{Self-Attn}_{vis}\left(\dot{hv}_1^k, \dot{hv}_2^k, \ldots, \dot{hv}_m^k\right) \tag{9}$$

$$\ddot{hg}_1^k, \ddot{hg}_2^k, \ldots, \ddot{hg}_n^k = \text{Self-Attn}_{lang}\left(\dot{hg}_1^k, \dot{hg}_2^k, \ldots, \dot{hg}_n^k\right) \tag{10}$$

where Self-Attn$_{vis}$ is the Self-Attention layer for vision modality, Self-Attn$_{lang}$ is the Self-Attention layer for language modality.

Finally, the Feed Forward layers are fed with each output of Self-Attention layers:

$$hv_1^k, hv_2^k, \ldots, hv_m^k = \text{FF}_{vis}\left(\ddot{hv}_1^k, \ddot{hv}_2^k, \ldots, \ddot{hv}_m^k\right) \tag{11}$$

$$hg_1^k, hg_2^k, \ldots, hg_n^k = \text{FF}_{lang}\left(\ddot{hg}_1^k, \ddot{hg}_2^k, \ldots, \ddot{hg}_n^k\right) \tag{12}$$

where FF$_{vis}$ is the Feed-Forward layer for vision modality, FF$_{lang}$ is the Feed-Forward layer for language modality. The Cross-Modal Transformer takes the output of the Visual encoder and ObjectGroundedBERT as an input. The output of the Cross-Modal Transformer is the vision and language representation of an image-text pair, that contains the alignment of two modalities:

$$hv_1, hv_2, \ldots, hv_m, hg_1, hg_2, \ldots, hg_n = \text{CMT}\left(\hat{hv}_1, \hat{hv}_2, \ldots, \hat{hv}_m, \hat{hg}_1, \hat{hg}_2, \ldots, \hat{hg}_n\right) \tag{13}$$

where CMT stands for Cross-Modal Transformer, $\hat{hv}_j, \hat{hg}_i$ are the output representation from Visual encoder and Object Grounded BERT respectively.

3.4 Training

In this section, we introduce the pretraining tasks for our framework. We denote that the image contains m objects $\mathbf{o} = \{o_1, ..., o_m\}$, the input words tokenized from the sentence having n tokens $\mathbf{w} = \{w_1, ..., w_n\}$ and θ is the parameters of the model need to be optimized.

Masked Language Modeling The task setup is similar to BERT pretraining where words are randomly masked with a probability of 0.15 and the model is asked to predict these masked words. Different from BERT, the masked words are predicted not only from the non-masked words but also from the visual information. Therefore, it helps to build the relationship between the visual information and the language representation, also the Text-grounding module can learn the visual object representation of the language. The model aims to predict the masked words $\mathbf{w}_{masked}$ based on their surrounding words $\mathbf{w}$ and the objects from the image $\mathbf{o}$, by optimizing the objective function:

$$\mathcal{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(\mathbf{w},\mathbf{o})\sim D} \log P_\theta(\mathbf{w}_{mask}|\mathbf{w}, \mathbf{o}) \tag{14}$$

Masked Visual Feature Prediction Same as the Masked Language Modeling task, we train the model by randomly masking objects with a probability of 0.15 (masking feature vector with zero vector) and asking the model to predict the class of the masked objects so that the model can infer masked regions from the visible objects and from the sentences. The goal of this task is to predict the masked object class $c(\mathbf{o}_{mask})$ based on the observation of their textual caption and visible objects. Specifically, the Text-grounding module will learn how to transform the contextual representation into visual information and eventually help the Cross-Modal Transformer know which object is missing. In the end, the model will learn object relationships and the visual grounding of the language. The objective minimizes the cross-entropy loss:

$$\mathcal{L}_{\text{MVFP}}(\theta) = \mathbb{E}_{(\mathbf{w},\mathbf{o})\sim D} \sum_{\mathbf{o}_{masked}} \text{CE}(c(\mathbf{o}_{mask}), \mathbf{o}^{gt}_{mask}), . \tag{15}$$

where CE is Cross-entropy loss function, , $\mathbf{o}^{gt}_{mask}$ is the object detection output from Faster R-CNN as the label of the masked object.

Image-Text Matching In this task, the model learns an instance-level alignment between the whole image and the caption that describes the image. This objective aims to learn the global representation of the image and the sentence. The model has to learn the object relationship in the image also the contextual meaning of the sentence to know whether they are matched together. Moreover, with the negative samples, we teach the model to find negative information for classifier training by comparing the linguistic descriptions of those objects in the image. We take the output representation of a special token [CLS] and feed it into an FC layer with a sigmoid function to predict an alignment score between 0 and 1:

$$\mathcal{L}_{\text{ITM}}(\theta) = -\mathbb{E}_{(\mathbf{o},\mathbf{v})\sim D}[y \log \text{fc}(\mathbf{o}, \mathbf{v}) + (1 - y) \log(1 - \text{fc}(\mathbf{o}, \mathbf{v}))]). \tag{16}$$

where fc is the output score of the text-image pair from an FC layer, y is the label for image-sentence pair, 0 is unmatched and 1 is matched.

4 Experimental Setup

4.1 Datasets

Training We use MSCOCO[2] dataset (Lin et al., 2014) as the training data for image projection and Text-grounding module. This dataset consists of 118K/5K/41K (train/val/test) images, each with five captions describe the image.

Evaluation After training process, we finetune and evaluate our model on GLUE[3] (Wang et al., 2018), SQuAD 1.1 (Rajpurkar et al., 2016), and SQuAD 2.0[4] datasets (Rajpurkar et al., 2018). In GLUE dataset, we evaluate our model on various tasks over 8 corpora: CoLA (Warstadt et al., 2018), MNLI (Williams et al., 2018), MRPC (Dolan & Brockett, 2005), QNLI (Rajpurkar et al., 2016), QQP (Iyer et al., 2017), RTE (Dagan et al., 2006; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), SST-2 (Socher et al., 2013), STS-B (Cer et al., 2017). The statistics of datasets are given in table 1.

Table 1: Task descriptions and statistics.

Corpus	**Train**	**Test**	**Metrics**
GLUE			
CoLA	8.5k	1k	Matthews corr
MNLI	393k	20k	acc.
MRPC	3.7k	1.7k	acc./F1
QNLI	105k	5.4k	acc.
QQP	364k	391k	acc./F1
RTE	2.5k	3k	acc.
SST-2	67k	1.8k	acc.
STS-B	7k	1.4k	Pearson corr.
SQUAD			
SQUAD V1.1	87K	10K	exact match/F1
SQUAD V2.0	130K	11K	exact match/F1

[2] https://cocodataset.org/#home

[3] https://gluebenchmark.com/

[4] https://rajpurkar.github.io/SQuAD-explorer/

Table 2: Downstream task results of BERT, V&L pretrained models and our ObjectGroundedBERT (OGBERT) on BERT-base architectures. MRPC and QQP results are F1 score, STS-B results are Pearson correlation, SQuAD v1.1 and SQuAD v2.0 results are exact matching and F1 score respectively. The Δ_{base} and Δ_{voken} column show the difference between our model and the baselines.

Task	OGBERT	BERT-base	$\Delta_{base} \uparrow$	Vokenization	$\Delta_{voken} \uparrow$	LXMERT	VisualBERT	VL-BERT	ViLBERT
CoLA	**59.08**	54.68	4.41	_	_	15.76	45.14	57.01	56.05
MNLI	**85.21**	83.48	1.84	82.6	2.6	35.44	80.68	81.18	81.29
MNLI-MM	**86.13**	84.05	1.9	_	_	35.22	80.96	81.38	81.02
MRPC	**89.31**	88.82	1.00	_	_	80.64	87.36	87.76	86.95
QNLI	**92.76**	91.37	1.27	88.6	4.1	50.54	87.39	89.20	86.95
QQP	**89.15**	87.33	1.82	88.6	0.5	79.80	86.01	85.94	85.41
RTE	**73.65**	67.87	2.89	_	_	52.71	66.43	62.09	70.40
SST-2	**93.82**	92.43	1.51	92.2	1.6	82.11	88.88	88.88	90.14
STS-B	**90.30**	89.00	0.30	_	_	42.23	90.03	89.48	89.98
SQuADv1.1	**78.81/86.98**	78.10/86.31	0.63/0.68	78.8/86.7	0.0/0.2	9.39/17.65	68.51/77.71	72.62/81.30	72.95/81.35
SQuADv2.0	**68.94/72.13**	67.92/71.08	1.04/1.03	68.1/71.2	0.8/0.9	46.52/47.04	59.17/62.53	62.38/65.63	63.36/66.56

Table 3: Downstream task results of our method with different dimension of grounding embedding.

Dimension	CoLA	MNLI	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	STS-B	SQuAD V1.1	SQuAD V2.0
72	**59.08**	84.41	**86.13**	**89.31**	92.31	**89.15**	**73.65**	**93.82**	**90.3**	**78.81/86.98**	68.85/71.92
108	58.95	**85.21**	85.76	88.8	**92.76**	88.87	72.2	93.23	90.25	78.71/86.72	68.77/71.72
144	58.25	84.97	85.73	**89.3**	92.35	88.92	72.56	93.7	90.21	78.6/86.6	**68.94/72.13**

4.2 Evaluation tasks and metrics

All tasks are classification while STS-B is a regression task. MNLI has three classes whereas all other classification tasks are binary classification. The evaluation tasks are also various: question answering (QNLI, SQuAD), acceptability (CoLA), sentiment (SST-2), paraphrase (MRPC, QQP), inference (MNLI, RTE, QNLI). The metric of each task is shown in table 1.

For MRPC and QQP, we report F1 score. For STS-B, we report Pearson correlation. For both SQuAD v1.1 and SQuAD v2.0, we report exact matching and F1 score respectively.

4.3 Implementation

We use BERT-base-uncased as the language model for the baseline also as the language model of ObjectGroundedBERT. We load the BERT weight trained on Bookcorpus and Wikipedia from Pytorch framework Huggingface [5]. We stack 5 transformer encoder layers for Visual Encoder and 5 Cross-modal layers for Cross-modal Transformer. The language base model is freeze and also the object detector. We train the Visual Encoder, Cross-modal Transformer and the Text-grounding part based on the contextual representation and objects from the image.

Text-grounding module is a multi-layer perceptron with 1 hidden layer and applies relu activation. We set the MLP final output dimension in set $\{72, 108, 144\}$ for evaluating how visual information impact on the textual-visual representation in Sec 6.1. The hidden state dimension of the Visual Encoder and Cross-modal Transformer is the sum of the hidden state dimension of the language encoder and the Text-grounding output dimension, we set 12 heads of each layer. We set $k = 4$ final hidden output layers, $J = 36$ fixed number of objects in the image. Our model is trained with a learning rate $l_r = 1e^{-4}$ in 8 epochs using AdamW (Loshchilov & Hutter, 2018) as an optimizer, we set batch size of 256 and the model is trained on one A100 GPU for approximately 2.5 days.

5 Experimental Results

5.1 Compared to the baseline models

In this experiment, we compare our method with two baselines:

- BERT (Devlin et al., 2019): BERT-base-uncased which was pretrained on BooksCorpus and English Wikipedia.

[5] https://huggingface.co/

Table 4: Downstream task results of our ObjectGroundedBERT without training the Text-grounding Module. The first three rows report the fine-tuned results of our model without training with the visual grounded datasets, the last three rows show the difference to the results reported in Table 3.

Dimension	CoLA	MNLI	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	STS-B	SQuAD V1.1	SQuAD V2.0
72_{wo}	55.62	83.66	84.45	87.74	90.85	86.79	70.4	91.25	89.97	77.99/86.51	67.85/70.98
108_{wo}	55.86	83.55	84.05	88.28	90.88	87.31	71.12	91.71	89.78	78.17/86.47	68.34/71.55
144_{wo}	54.88	83.33	83.95	86.96	90.89	87.54	68.95	91.82	89.47	78.24/86.4	66.86/69.93
Δ_{72}	-3.46	-0.75	-1.68	-1.57	-1.46	-2.36	-3.25	-2.57	-0.33	-0.82/-0.47	-1/-0.94
Δ_{108}	-3.09	-1.66	-1.71	-0.52	-1.88	-1.56	-1.08	-1.52	-0.47	-0.54/-0.25	-0.43/-0.17
Δ_{144}	-3.37	-1.64	-1.78	-2.34	-1.46	-1.38	-3.61	-1.88	-0.74	-0.36/-0.2	-2.08/-2.2

Table 5: Downstream task results on different pretraining tasks.

Approach	CoLA	MNLI	MNLI-MM	MRPC	QNLI	QQP	RTE	SST-2	STS-B	SQuAD V1.1	SQuAD V2.0
MLM	56.78	84.92	85.54	88.26	91.94	87.77	71.84	92.77	89.85	78.35/86.52	68.18/71.26
MLM + ITM	58.09	84.25	85.7	88.52	91.93	88.6	72.56	92.89	89.81	78.63/86.55	68.44/71.56
MLM + MVFP	57.37	84.54	85.72	88.43	92.01	88.28	**73.29**	92.77	89.93	78.64/86.66	68.67/71.69
All tasks	**58.95**	**85.21**	**85.76**	**88.8**	**92.76**	**88.87**	72.2	**93.23**	90.25	**78.71/86.72**	**68.77/71.72**

- Vokenizer (Tan & Bansal, 2020): This is the closest related work and state-of-the-art grounded language learning method applying on BERT. Note that while Vokenizer directly grounds visual knowledge into the language model, we freeze the language model and expand the text representation with the grounding embedding containing visual information learned from the pretraining process.

The fine-tune results on 10 different natural-language tasks are reported in Table 2. Our ObjectGroundedBERT outperforms the baselines on all downstream tasks. Specifically, we achieve an improvement from 0.30 to 4.41 score compared to BERT-base and up to 4.1 scores compared to Vokenization. This shows that our grounded language model representation can capture more useful information for language understanding.

Our method has the significant improvement on the tasks of **CoLA, MNLI, RTE, QQP, QNLI** and also question answering tasks **QNLI, SQuAD V2.0**. It may due to the fact that in these tasks, the visual information and relation between objects are significantly important. As such, our approach can ground the language representation with useful visual-objected information that empowers the textual modality with the external knowledge from the visual modality.

5.2 COMPARED TO OTHER VISION-AND-LANGUAGE PRETRAINED MODELS

To show the effectiveness of our proposed grounded language learning approach, we compare it with the following state-of-the-art vision-and-language pretrained models which also train with the image-text pairs:

- **LXMERT** (Tan & Bansal, 2019) consists of two single-modal and one cross-modal Transformer to connect vision and language semantics via pre-training tasks.
- **ViLBERT** (Lu et al., 2019) extends the BERT architecture to a multi-modal two-stream model and process both visual and textual inputs.
- **VisualBERT** (Li et al., 2019) consists of a stack of Transformer layers that implicitly align elements of an input text and regions in an associated input image with self-attention.
- **VL-BERT** (Su et al., 2020) uses powerful Transformer model as the backbone, and extends it to input both visual and linguistic embedded features.

We also fine-tune all models on 10 different natural-language tasks of GLUE and SQuAD datasets. Following their works, all models before pretrained with image captioning datasets are initialized with the pretrained BERT weights, except LXMERT. As shown on the right in Table 2, the finetuning results on our model consistently outperform other Vision Language pretrained models in all tasks. The results show that pretraining the Transformer with image-text pairs after initialized the pretrained BERT weight leads to bad performance on pure language tasks.

6 ANALYSIS

6.1 IMPACT OF VISUAL DIMENSION

We study the impact of the contribution visual grounding with different grounding embedding dimensions. Table 3 reports the evaluation of our model on GLUE and SQUAD on 3 different dimensions $\{72, 108, 144\}$.

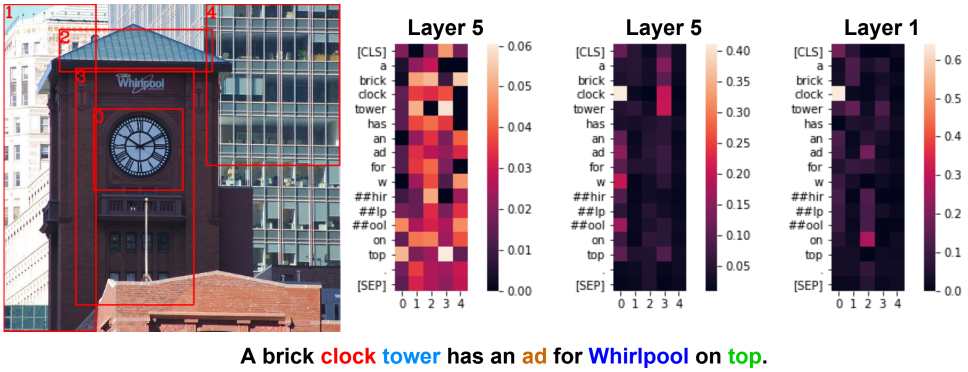

Figure 4: Illustration of the Attention map of Cross modal layers.

We train all 3 models with the same hyper-parameters (number of epochs, batch size, learning rate, model hyper-parameters). The output dimension of textual embedding is fixed the same as the language model and we do not fine-tune it on visual data so that the dimension of grounding embedding means how much visual information we want to add to the language model.

The results show that most of the downstream tasks achieve the best performance at the dimensions of 72, except MNLI (108), QNLI (108), and SQuAD V2.0 (144). The experiments show that too much grounded visual information adding to the language representation is not good for downstream language tasks as it leads to bias in visual information, unbalance the contribution of textual meaning and visual grounding.

6.2 The impact of visual grounding

The authors backing and promote the standards of open science and reproducible examination. The algorithms and models proposed in this work will be publicly released in a free and public code online with simple to-utilize content. In fact, the tables included in the paper mention results on the number of layers and the total number of model parameters that are trained. Similarly, the visualization and illustrations presented in the main paper contain the exact details on the dataset used, the model architectures, and the counterexample.

6.3 Visualization of alignment between tokens and objects

In Fig. 4, we visualize the attention in the cross-modality transformer to show the connections between objects and tokens. We find that the attention of words **”clock, tower, ad, top”** is focused on the right regions of the image on the shallow layers (Layer 1 and 3). While on the deep layer (Layer 5), the attention map is more complex when the relationship between objects in the image is learned during pretraining.

6.4 Evaluation on Pre-training Tasks

We analyze the effectiveness of different pre-training settings on the downstream tasks by conducting experiments on ObjectGroundedBERT pretrained with different combinations of pretraining tasks. Table 5 reports evaluations of our model on GLUE and SQUAD on 4 different approaches, i.e., only Masked Language Modeling (MLM), Masked Language Modeling and Image-Text Matching (MLM + ITM), Masked Language Modeling and Masked Visual Feature Prediction (MLM + MVFP), and our full approach (MLM + MVFP + ITM). The results show that our full approach achieves the highest score in all tasks except RTE, and both (MLM + MVFP) and (MLM + ITM) achieve better performance than the only (MLM).

7 Conclusion

In this paper, we propose a novel grounded language framework that enhances visual-objected-grounded into language representation. Instead of using the features of the whole image encoded from CNN, we use the output from an off-the-shelf object detector as the input of the visual modality. Via the multi-task pretraining strategy, the Cross-modal Transformer can connect the visual and language modality together and teaches the Text-grounding part to have the ability to capture the visual object information and their relations from the contextual meaning of language. Our model significantly outperforms the baseline language models on various language tasks of GLUE and SQuAD datasets.

Ethics Statement

In this study, we introduce a grounding language learning method by adding a text-grounding module to empower the text representation with the visual representation. Our main work is instead of extracting features from the whole image using CNN networks, we propose a novel object-level grounded language framework to connect the objects in the image and the sentence together. This framework learns the alignment between the visual scene and the tokens in the caption, thus the text-grounding module will have the ability to transform the context into the visual information. Moreover, we do not ground the visual content directly into the language model, thus the language model will not forget the knowledge from the textual corpora. In consequence, one beneficial impact of freezing the language model is that its use can reduce computational resources when training a deep learning model.

Reproducibility Statement

The authors support and advocate the principles of open science and reproducible research. The algorithms and architectures proposed in this work shall be open-sourced in a free and public code repository with easy-to-use scripts to reproduce several experiments and evaluations presented. In fact, the tables included in the paper mention results on the number of layers and the total number of model parameters that are trained. Similarly, the visualization and illustrations presented in the main paper contain the exact details on the dataset used, the model architectures, and the counterexample.

References

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *TAC*, 2009.

Patrick Bordes, Éloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. Incorporating visual semantics into sentence representations within a grounded space. In *EMNLP*, 2019.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *CoRR*, abs/2005.14165, 2020. URL `https://arxiv.org/abs/2005.14165`.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 1–14, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2001. URL `https://www.aclweb.org/anthology/S17-2001`.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2019.

Guillem Collell, Ted Zhang, and Marie-Francine Moens. Imagined visual representations as multimodal embeddings. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 670–680, 2017.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alché Buc (eds.), *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pp. 177–190, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-33428-6.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

William B. Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005. URL `https://www.aclweb.org/anthology/I05-5002`.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and William B Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pp. 1–9, 2007.

R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, 2006.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 1367–1377, 2016. URL `http://aclweb.org/anthology/N/N16/N16-1162.pdf`.

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. First quora dataset release: Question pairs. *data. quora. com*, 2017.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 408–418, 2018.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3294–3302, 2015. URL `http://papers.nips.cc/paper/5950-skip-thought-vectors`.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019.

Angeliki Lazaridou, Marco Baroni, et al. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 153–163, 2015.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pp. 1188–1196, 2014.

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

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*, 2020.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL `https://doi.org/10.1007/978-3-319-10602-1_48`.

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

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL `https://openreview.net/forum?id=rJvJXZb0W`.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke 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, Volume 1 (Long Papers)*, pp. 2227–2237, 2018.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 784–789, 2018.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084, 2019. URL `http://arxiv.org/abs/1908.10084`.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/D13-1170`.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*, 2020.

Hao Tan and Mohit 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)*, pp. 5103–5114, 2019.

Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *CoRR*, abs/2010.06775, 2020. URL `https://arxiv.org/abs/2010.06775`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018. URL `http://arxiv.org/abs/1804.07461`.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. Neural network acceptability judgments. *CoRR*, abs/1805.12471, 2018. URL `http://arxiv.org/abs/1805.12471`.

Adina Williams, Nikita Nangia, and Samuel 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, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. URL `http://aclweb.org/anthology/N18-1101`.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pp. 5754–5764, 2019.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, 2020.