Expand BERT Representation with Visual Information via Grounded Language Learning with Multimodal Partial Alignment

Cong-Duy Nguyen*
Nanyang Technological University
nguyentr003@ntu.edu.sg

The-Anh Vu-Le*
University of Illinois
Urbana-Champaign
vltanh@illinois.edu

Thong Nguyen
National University of Singapore
e0998147@u.nus.edu

Tho Quan
Ho Chi Minh City University of
Technology (HCMUT), VNU-HCM
qttho@hcmut.edu.vn

Anh-Tuan Luu
Nanyang Technological University
anhtuan.luu@ntu.edu.sg

ABSTRACT

Language models have been supervised with both language-only objective and visual grounding in existing studies of visual-grounded language learning. However, due to differences in the distribution and scale of visual-grounded datasets and language corpora, the language model tends to mix up the context of the tokens that occurred in the grounded data with those that do not. As a result, during representation learning, there is a mismatch between the visual information and the contextual meaning of the sentence. To overcome this limitation, we propose GroundedBERT - a grounded language learning method that enhances the BERT representation with visually grounded information. GroundedBERT comprises two components: (i) the original BERT which captures the contextual representation of words learned from the language corpora, and (ii) a visual grounding module which captures visual information learned from visual-grounded datasets. Moreover, we employ Optimal Transport (OT), specifically its partial variant, to solve the fractional alignment problem between the two modalities. Our proposed method significantly outperforms the baseline language models on various language tasks of the GLUE and SQuAD datasets.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Computer vision**; • **Theory of computation** → **Theory and algorithms for application domains**.

KEYWORDS

Grounded Language Learning, Optimal Transport

ACM Reference Format:
Cong-Duy Nguyen, The-Anh Vu-Le, Thong Nguyen, Tho Quan, and Anh-Tuan Luu. 2023. Expand BERT Representation with Visual Information via Grounded Language Learning with Multimodal Partial Alignment. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3612248

*Both authors contributed equally to this research.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10.
https://doi.org/10.1145/3581783.3612248

Table 1: Statistics of some common datasets used in visual grounded language learning task.

	Book Corpus	Wikipedia	MS COCO
# of words	985M	2471M	6M
# of sentences	74M	113M	616K
# of unique words	1M	8M	44K

1 INTRODUCTION

Grounded language learning is concerned with learning the meaning of language as it applies to the real world. Humans, especially children, learn language from not only pure textual information but also different modalities such as vision and audio, which contain rich information that cannot be captured by text alone [35, 44, 50]. However, many traditional language models are learned only from textual corpora [3, 11]. They have the limitation in learning complex semantics that requires the combination of signals in data through cross-referencing and synthesis.

Recently, there are many studies trying to improve the language representation with visual information [2, 9, 16, 21, 49]. In those attempts, they update the weights of the language encoder using the visual objective together with the pure language-based objective during pretraining. However, there is usually a huge gap in the distribution and quantity of word tokens between visual datasets and language corpora. For example, in Table 1, the Book Corpus and Wikipedia, two conventional language corpora, contain billions of words with millions of unique tokens, while MS COCO, a common visual-grounded dataset, contains only 6 million words and 44 thousand unique tokens. Therefore, during visual-grounded learning, only the tokens from the visual datasets are updated while the majority of the tokens are not equipped with visual information. However, during pretraining, those tokens with and without information from the images will be mixed up in the same context of the sentence, confusing the contextual learning process.

Moreover, previous attempts compressed the entire image into one vector as a global representation and then matched it to the paired caption. However, as shown by the samples picked from the Visual Genome [19] dataset in Figure 1, many of the captions

Figure 1: Example image-caption pairs in Visual Genome dataset [19]

only describe local regions in the corresponding image. Thus, using a global representation vector can distract the encoder from capturing local information, making it difficult for the model to align between modalities. As a solution to this issue, we use the Vision Transformer [13] model as the visual encoder to store local information in patch embeddings.

Additionally, aligning information from different modalities is a crucial phase in vision-language representation learning because it is how two sources of information are combined. There are existing researches that use Optimal Transport to solve this alignment problem. Uniter [7] and ViLT [17] used the OT-based distance as a pretraining objective, while Graph Optimal Transport [5] considered two OT distance: Wasserstein distance (WD) and Gromov-Wasserstein distance (GWD) for cross-domain alignment in Visual Question Answering. Nevertheless, the classical optimal transportation problem seeks a transportation map that satisfies marginal constraints, requiring masses from all sources to be moved to all destinations. In some cases, we want only a fraction of masses to be carried, making this requirement restrictive. For instance, as stated above, the caption only describes a part of the image. To get a more flexible alignment, we propose to adapt the Partial Optimal Transport variant to align between the modalities.

Our contributions can be summarized as:

- We propose GroundedBERT - a grounded language representation that extends the BERT representation with visual information. The visual-grounded representation is first learned from the text-image pairs and then concatenated with the original BERT representation to form a unified visual-textual representation.
- We use patch embeddings from Vision Transformer to maintain local information of the image instead of a single global representation. We also adapt Partial Optimal Transport to align between the two modalities.
- We conduct extensive experiments on various language downstream tasks on the GLUE and SQuAD datasets. Empirical result shows that we significantly outperforms the baselines on these tasks.

2 RELATED WORK

Over the past decades, many approaches have been proposed to learn language representation. Skip-gram [30], GLOVE [38] were proposed to learn word representations. On the other hand, FastSent [14], QuickThought [27], SkipThought [18], Sentence-BERT [42], or [10, 22] tried to learn the sentence representations. Recently, many language models such as ELMo [39], BERT [11], RoBERTa [26], XLNet [54], GPT [3], ELECTRA [8], ALBERT [20] were proposed to learn the contextual representation. However, these studies learn the language representation on only textual corpora.

In recent years, many vision-and-language pretrained 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 [7, 23, 47]. While [24, 55] used only one cross-modal Transformer for learning, [29, 48] 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.

Advanced machine learning algorithms such as the Contrastive Learning framework have been applied to the natural language processing and computer vision [31, 33, 34, 36, 37, 45]. Optimal Transport has also been extensively used in many natural language processing tasks and also the integration of vision and language fields, for example, Cross-Lingual Abstractive Summarization [32], machine translation [6], Vision and language pretraining [7, 17], Visual Question Answering [5], etc. Nevertheless, the application of the variants of OT has been less attractive in vision-and-language research.

There are many works on grounded language learning [1, 15, 43] having been introduced in the past few years. On the other hand, there are few attempts to improve language representation with visual information. [21] introduced multimodal skip-gram models (MMSKIP-GRAM) taking visual information into account. [9] proposed IMAGINET which consists of GRU networks and tried to predict the visual representation and the next word in the sentence. [16] was similar to IMAGINET but they used a bi-directional LSTM for sentence encoder. Moreover, it aimed to predict both the visual feature and the other captions given one caption. [2] proposed an intermediate space called the grounded space and learns the visual and textual representation with cluster information and perceptual information. [49] introduced the concept of vokenization and pretrained the language model with an additional voken-classification task.

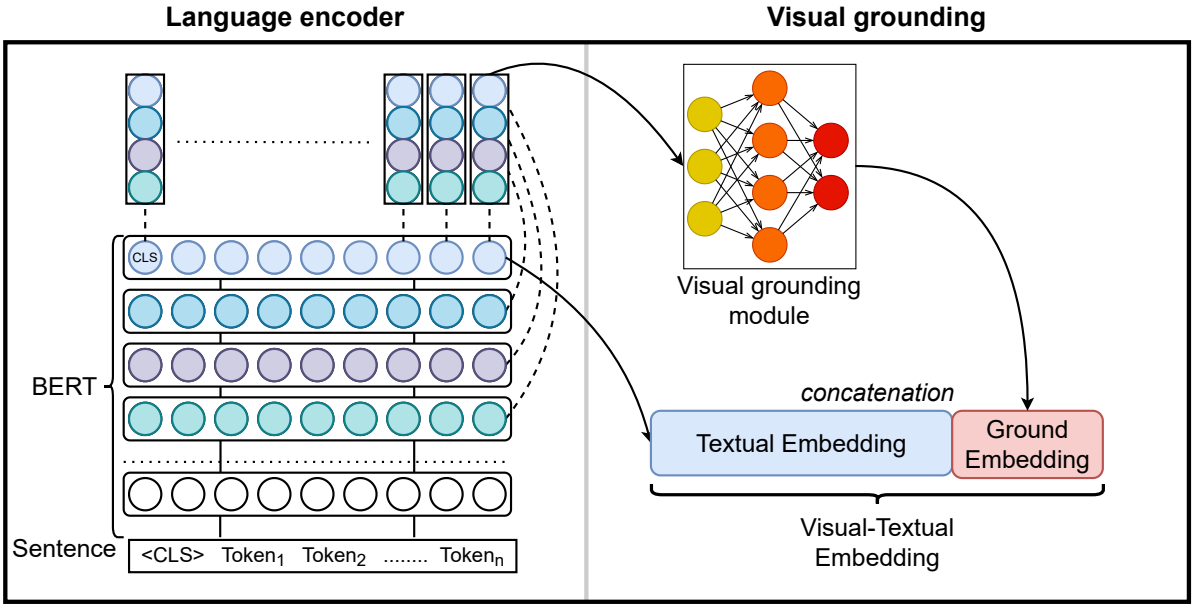

Figure 2: Implementation of our GroundedBERT. The model consists of two components, i.e. Language encoder and Visual grounding part. The new representation of language model combines of Textual embedding and Visual embedding.

3 METHODOLOGY

In this section, we introduce the details of our proposed GroundedBERT. As shown in Figure 2, our model consists of two components: a language encoder and a visual grounding module. The complete framework is illustrated in Figure 3 where two objectives are introduced.

3.1 Language encoder

We use BERT [11] as the language encoder. Given an input sentence $s = (w_1, \ldots, w_n)$, we use the pretrained BERT model to contextually embed the discrete tokens w_i's into hidden-output vector $\boldsymbol{h}_i$'s:

$$\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_l = BERT(w_1, w_2, \ldots, w_l) \tag{1}$$

where $\boldsymbol{h}_i = (h_i^1, h_i^2, \ldots, h_i^L)$, h_i^l is the hidden state of token i at layer l of the Transformer.

3.2 Visual grounding

Visual grounding module. The visual grounding module is a multi-layer perceptron to transform the contextual representation of each token in the sentence into the (visual) ground embedding.

We take the hidden states of k final Transformer layers h_i^{L-k+1}, $h_i^{L-k+2}, \ldots, h_i^L$ and concatenate them as the input for the visual grounding module.

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

$$g_i = MLP_{VG}(\tilde{h}_i) \tag{3}$$

where g_i is (visual) ground embedding of token i, $[h_i^{L-k+1}, h_i^{L-k+2}, \ldots, h_i^L]$ is the concatenation of hidden states of token i from layer $L-k+1$ to L, VG stands for Visual Grounding.

Visual-Textual Embedding. The textual embedding is the final hidden state of the language encoder. The ground embedding are concatenated to this textual embedding to form a unified visual-textual embedding of the token in the sentence.

$$t_i = [h_i^L, g_i] \tag{4}$$

where t_i is the visual-textual embedding vector of the i-th token, which we take as the final representation of the token using our GroundedBERT model, $[h_i^L, g_i]$ is the concatenation of the final hidden state h_i^L and the ground embedding g_i.

3.3 Visual encoder

Patch embedding. Instead of using traditional convolution-based architectures for visual feature extraction [2, 16, 49], we use Vision Transformer (ViT) [13]. Let img be the input image having size of (c, w, h) which stands for the number of channels, width, and height of the image. Image img goes through the ViT to get a global feature vector $\tilde{v}_{CLS}$ and m patch embeddings $\tilde{v}_1, \ldots, \tilde{v}_m$.

$$\tilde{v}_{CLS}, \tilde{v}_1, \ldots, \tilde{v}_m = ViT(img) \tag{5}$$

Image projection. We use a multi-layer perceptron to project the feature vector $\tilde{v}_i$ of each patch to the grounded space and represent visual context learned from visual features.

$$v_{CLS}, v_1, ..., v_m = MLP_{prj}(\tilde{v}_{CLS}, \tilde{v}_1, ...\tilde{v}_m) \tag{6}$$

where v_{CLS} is the global embedding of the input image, v_i's are the patch embeddings and prj stands for (image) projection.

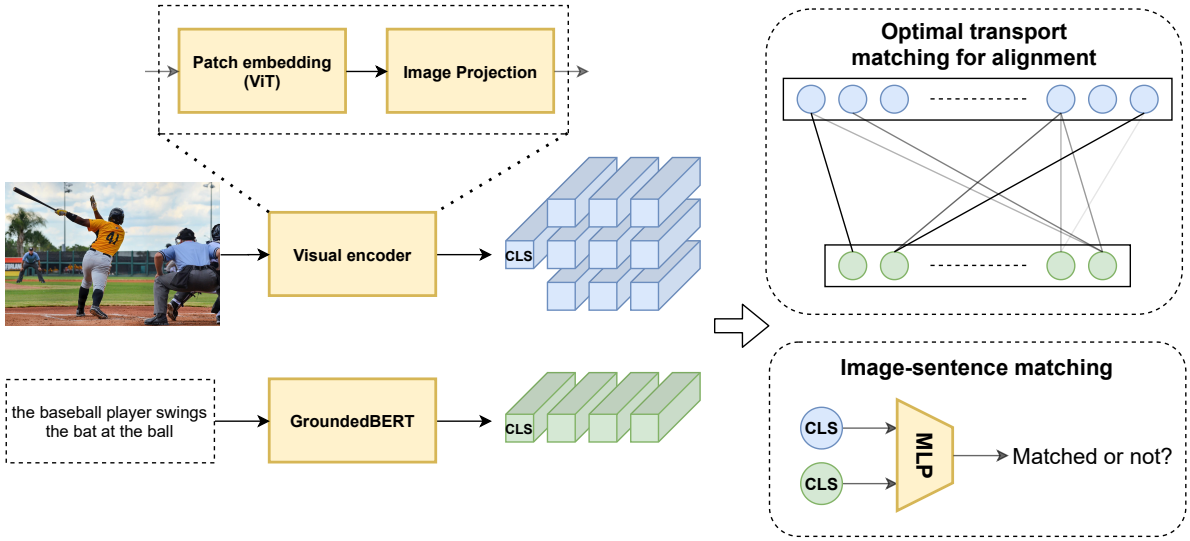

Figure 3: Implementation of our training framework. The framework consists of two parrallel pipeline for visual and text, finally, the whole model is train with two objectives: Image-sentence matching and Optimal transport matching for alignment.

Algorithm 1 Computing Optimal Transport.

Input: $\mathbf{C} \in \mathbb{R}^{m \times n}$, $\boldsymbol{a} \in \mathbb{R}^m$, $\boldsymbol{b} \in \mathbb{R}^n$, β, $iter$
$\boldsymbol{\sigma} = \mathbf{1}_n / n$, $\mathbf{T} = \mathbf{1}_m \mathbf{1}_n^\top$
$\mathbf{A} = \exp\left(-\frac{\mathbf{C}}{\beta}\right)$
for $t = 1, 2, 3, \ldots, iter$ **do**
  // all division operations are element-wise
  $\mathbf{Q} = \mathbf{A} \odot \mathbf{T}$ // $\odot$ is the Hadamard product
  $\boldsymbol{\delta} = \frac{\boldsymbol{a}}{\mathbf{Q}\boldsymbol{\sigma}}$, $\boldsymbol{\sigma} = \frac{\boldsymbol{b}}{\mathbf{Q}^\top \boldsymbol{\delta}}$
  $\mathbf{T} = \mathrm{diag}(\boldsymbol{\delta}) \mathbf{Q} \mathrm{diag}(\boldsymbol{\sigma})$
end for
$\mathcal{D} = \langle \mathbf{C}, \mathbf{T} \rangle$ // $\langle \cdot, \cdot \rangle$ is the Frobenius dot-product
Return $\mathbf{T}$, $\mathcal{D}$

Algorithm 2 Computing Partial Optimal Transport.

Input: $\mathbf{C} \in \mathbb{R}^{m \times n}$, $\boldsymbol{a} \in \mathbb{R}^m$, $\boldsymbol{b} \in \mathbb{R}^n$, β, s, $iter$
$\mathbf{T} = \exp\left(-\frac{\mathbf{C}}{\beta}\right)$
$\mathbf{T} = \frac{s}{\mathbf{1}_n^\top \mathbf{T} \mathbf{1}_m} \mathbf{T}$
for $t = 1, 2, 3, \ldots, iter$ **do**
  // all division operations are element-wise
  $\boldsymbol{k}_a = \min\left(\frac{\boldsymbol{a}}{\mathbf{T}\mathbf{1}_n}, \mathbf{1}_m\right)$
  $\mathbf{T}_a = \mathrm{diag}(\boldsymbol{k}_a)\mathbf{T}$
  $\boldsymbol{k}_b = \min\left(\frac{\boldsymbol{b}}{\mathbf{T}_a^\top \mathbf{1}_m}, \mathbf{1}_n\right)$
  $\mathbf{T}_b = \mathrm{diag}(\boldsymbol{k}_b)\mathbf{T}_a$
  $\mathbf{T} = \frac{s}{\mathbf{1}_n^\top \mathbf{T}_b \mathbf{1}_m} \mathbf{T}_b$
end for
$\mathcal{D} = \langle \mathbf{C}, \mathbf{T} \rangle$ // $\langle \cdot, \cdot \rangle$ is the Frobenius dot-product
Return $\mathbf{T}$, $\mathcal{D}$

3.4 Training

In this section, we introduce two different optimization objectives: Image-sentence matching for global matching and Optimal transport matching for alignment between local features.

Image-sentence Matching. The Image-sentence Matching task is inherited from the Image-text Matching task from many vision-and-language pretraining literatures mentioned in Section 2. Learning how to perform well on this task will encourage the model to better find the relationship between the textual information and the visual signal in a global sense.

From each modality, we take a vector as its global representation. For the vision side, we use the global feature vector v_{CLS} from ViT. For the language side, we use the visual-textual embedding of the CLS token. We concat these two vectors before feeding into a fully connected layer with sigmoid activation to make the binary prediction of whether the sentence describes the image.

$$\hat{y} = \sigma(FC([v_{CLS}, t_{CLS}])) \tag{7}$$

where $\hat{y}$ is predicted probability, $\sigma(x) = [1 + \exp(-x)]^{-1}$ is the sigmoid function, $[.,.]$ is the concatenation operation.

The negative pair is created by replacing the image with another randomly selected image from the training set. We apply the binary cross-entropy loss for optimization.

$$\mathcal{L}_{cls} = -y \log \hat{y} - (1 - y) \log(1 - \hat{y}) \tag{8}$$

where y is the binary indicator, $y = 1$ if the image matches the sentence and 0 otherwise.

Optimal transport for vision-language alignment. To solve the alignment between language and vision, we use Optimal Transport (OT), specifically the Partial Optimal Transport (POT) variant.

For each image, we have m patch embeddings $\boldsymbol{v} = (v_1, \ldots, v_m)$. For each sentence, we have n hidden representations of the words $\boldsymbol{t} = (t_1, \ldots, t_n\}$. We consider these two collections as the supports of two empirical distributions with uniform weights. We then use OT to estimate the distance between these two distributions.

Specifically, we compute the cost matrix $\mathbf{C}$ where $c_{ij} = 1 - \cos \angle(v_i, t_j)$, or the cosine distance between the corresponding patch and word embedding. We also let $\boldsymbol{a} = \mathbf{1}_m / m$ and $\boldsymbol{b} = \mathbf{1}_n / n$ be the two uniform weight vectors.

Table 2: Downstream task results of BERT, Vokenization [49] and our GroundedBERT, we conduct the experiments on BERT-base architecture. MRPC results are F1 score, STS-B results are Pearson correlation, SQuAD v1.1 and v2.0 results are exact matching and F1 score. The results, which outperform the other one are marked in bold, are all scale to range 0-100. The Δ_{base} and Δ_{Vok} columns show the difference between our model and the baseline, and the Vokenization method respectively.

Task	Our	BERT-base	$\Delta_{base}\uparrow$	Vokenization[1]	$\Delta_{Vok}\uparrow$
CoLA	**60.95**	54.68	2.41	_	_
MNLI	**84.15**	83.48	0.84	82.6	1.55
MNLI-MM	**84.54**	84.05	0.83	_	_
MRPC	**89.25**	88.82	0.74	_	_
QNLI	**91.43**	91.37	0.5	88.6	2.83
RTE	**72.56**	67.87	3.6	_	_
SST-2	**93.12**	92.43	0.57	92.2	0.92
STS-B	**89.88**	89.00	0.84	_	_
SQuADv1.1	78.49/86.62	78.10/86.31	0.39/0.32	**78.8/86.7**	-0.31/-0.08
SQuADv2.0	**70.69/73.92**	67.92/71.08	2.77/2.84	68.1/71.2	2.59/2.72

Table 3: Task descriptions and statistics.

Corpus	Train	Test	Metrics
GLUE			
CoLA	8.5k	1k	Matthews corr
MNLI	393k	20k	matched acc./mismatched acc.
MRPC	3.7k	1.7k	acc./F1
QNLI	105k	5.4k	acc.
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

The distance between the two modalities can be defined using the OT-based distance as

$$\begin{aligned} \mathcal{D}(\boldsymbol{v},\boldsymbol{t}) &= \min_{\mathbf{T}} \langle \mathbf{T}, \mathbf{C} \rangle_F \\ \text{s.t.} \quad \mathbf{T}\mathbf{1}_n &= \boldsymbol{a}, \mathbf{T}^\top \mathbf{1}_m = \boldsymbol{b}, \mathbf{T} \geq \mathbf{0}_{m\times n} \end{aligned} \tag{9}$$

This formulation places constraints that all the mass from one distribution must be transported to the other distribution. We find, however, that this constraint is restrictive for the problem at hand, where the sentence describes only partially the corresponding image. Therefore, it is intuitively more apt to use the POT variant, described as follows.

$$\begin{aligned} \mathcal{D}(\boldsymbol{v},\boldsymbol{t}) &= \min_{\mathbf{T}} \langle \mathbf{T}, \mathbf{C} \rangle_F \\ \text{s.t.} \quad \mathbf{T}\mathbf{1} \preceq \boldsymbol{a}, \mathbf{T}^\top \mathbf{1}_n &\preceq \boldsymbol{b}, \mathbf{T} \geq \mathbf{0}_{m\times n} \\ \mathbf{1}_m^\top \mathbf{T} \mathbf{1} &= s \end{aligned} \tag{10}$$

where s is the total amount of mass to be transported. In our implementation, s is set as the total uniform weight vector of text.

We use sinkhorn-based algorithms to calculate the transportation plan $\mathbf{T}$ and the OT-based distance. Algorithm 1 and Algorithm 2 are for OT and POT, respectively. The average distance $\mathcal{D}$ for every matching pair of sentence and image will be minimized, while non-matching pair distance will be maximized. Formally, the alignment loss will be:

$$\mathcal{L}_{align} = \sum_{t,\boldsymbol{v}^+,\boldsymbol{v}^-\in S} \left[\mathcal{D}(\boldsymbol{v}^+,\boldsymbol{t}) - \mathcal{D}(\boldsymbol{v}^-,\boldsymbol{t})\right] \tag{11}$$

where S is the given dataset, $\boldsymbol{v}^+$ and $\boldsymbol{v}^-$ are the matching and non-matching image respectively corresponding to the sentence t. The procedure to pick the negative image is similar to in the Image-Sentence Matching task.

4 EXPERIMENTAL SETUP

4.1 Datasets

Training. We use MS COCO [25] and Visual Genome [19] image captioning datasets as the training data for image projection and Visual grounding module.

Evaluation. After training process, we finetune and evaluate our model on GLUE [51], SQuAD 1.1 [41], and SQuAD 2.0 datasets [40]. In GLUE dataset, we evaluate our model on various tasks over 7 corpora: CoLA [52], MNLI [53], MRPC [12], QNLI [41], RTE, SST-2 [46], STS-B [4]. The statistics of datasets are given in table 3.

4.2 Evaluation tasks and metrics

All tasks are single sentence or sentence pair classification except STS-B, which is a regression task. MNLI has three classes, all other classification tasks are binary classification. The evaluation tasks are also various: question answering (QNLI, SQUAD), acceptability (CoLA), sentiment (SST-2), paraphrase (MRPC), inference (MNLI,

Table 4: Downstream task results of different vision and language pretrained model.

Task	LXMERT	VisualBERT	VL-BERT	ViLBERT	Oscar	GroundedBERT
CoLA	15.76	45.14	57.01	56.05	41.21	**57.09**
MNLI	35.44	80.68	81.18	81.29	76.64	**84.32**
MNLI-MM	35.22	80.96	81.38	81.02	76.67	**84.88**
MRPC	80.64	87.36	87.76	86.95	80.58	**89.56**
QNLI	50.54	87.39	89.20	86.95	50.54	**91.87**
RTE	52.71	66.43	62.09	70.40	55.96	**71.47**
SST-2	82.11	88.88	88.88	90.14	87.61	**93.00**
STS-B	42.23	90.03	89.48	89.98	71.45	**89.84**
SQuADv1.1	9.39/17.65	68.51/77.71	72.62/81.30	72.95/81.35	21.77/32.20	**78.49/86.62**
SQuADv2.0	46.52/47.04	59.17/62.53	62.38/65.63	63.36/66.56	45.31/46.77	**70.69/73.92**

Table 5: Downstream task results and comparison of our GroundedBERT without training the Visual grounding module. The first two rows report the fine-tuned results of our model without training with the visual grounded datasets, while the last 4 rows show the results of our approaches on both OT and POT.

Dimension	CoLA	MNLI	MNLI-MM	MRPC	QNLI	RTE	SST-2	STS-B	SQuAD V1.1	SQuAD V2.0
64_{wo}	54.37	83.27	84.47	88.11	90.78	69.18	91.71	88.97	77.87/86.2	67.78/70.97
128_{wo}	53.59	83.78	84.04	88.63	91.45	69.53	91.12	89.3	77.98/86.03	68.18/71.45
64_{OT}	58.30	84.35	84.64	88.71	91.58	70.40	92.43	89.32	78.14/86.42	69.00/72.24
128_{OT}	59.1	84.49	84.76	88.81	91.61	69.31	92.78	89.75	78.21/86.46	68.6/71.94
64_{POT}	60.95	84.15	84.54	89.25	91.43	72.56	93.12	89.88	78.49/86.62	70.69/73.92
128_{POT}	57.77	84.06	84.3	88.93	91.31	70.04	92.32	89.86	78.27/86.45	69.53/72.88

RTE, QNLI). The metric of each task is shown in table 3. For MRPC, we report F1 score. For STS-B, we report Pearson correlation. For both SQuAD, we report exact matching and F1 score.

4.3 Implementation

We use BERT-base-uncased as the language model and vit base patch16 224 for the visual encoder. We load the BERT weight pretrained on Bookcorpus and Wikipedia from Pytorch framework Huggingface, and load the ResNeXt weight pretrained on ImageNet. The Language encoder and Patch embedding extraction are frozen, we just train the Image projection and Visual grounding module based on the contextual representation and image feature map. Both modules are multi-layer perceptron with 1 hidden layers and apply relu activation. We set the MLP final output dimension in set 64, 128 for evaluating how visual information impact on the textual-visual representation in Sec 6.1. Our model is trained with a learning rate $l_r = 1e^{-4}$ in 12 epochs using AdamW [28] as optimizer, we set batch size of 512 on 1 V100 GPU and train for 3-4 days.

5 EXPERIMENTAL RESULTS

5.1 Compared to the baseline models

The fine-tune results on 9 different natural-language tasks are reported in Table 2. We compare our GroundedBERT with the BERT-base as the language encoder to the BERT-base and Vokenization baseline respectively. Our GroundedBERT outperforms the baselines on all down-stream tasks. Specifically, we achieve an improvement from 0.5 to 3.6 score on BERT-base. Compared to Vokenization, we also achieve higher on most tasks, except SQuADv1.1. This shows that our grounded language model representation can capture more useful information for language understanding without changing the original language model.

5.2 Compared to other vision-and-language pretrained models

To prove the effectiveness of our proposed grounded language learning approach, we compare it with the following state-of-the-art vision-and-language pretrained models.

- **LXMERT** [48] consists of two single-modal and one cross-modal Transformer to connect vision and language semantics.
- **VisualBERT** [23] 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** [47] uses Transformer model as the backbone, and extends it to input both visual and linguistic embedded features.
- **ViLBERT** [29] extends the BERT architecture to a multi-modal two-stream model and process both visual and textual inputs.
- **Oscar** [24] uses object tags detected in images as additional points to ease the learning of alignments between text and image.

Table 6: Downstream task results of BERT and our GroundedBERT with different learning rates on GLUE.

Model	LR	CoLA	MNLI	MNLI-MM	MRPC	QNLI	RTE	SST-2	STS-B
BERT-base	2e-5	53.13	83.48	83.68	86.78	91.37	64.62	92.43	88.74
	3e-5	54.68	82.98	84.05	87.25	90.83	66.79	92.43	88.44
	4e-5	53.80	83.25	83.76	88.82	90.76	67.87	92.09	89.00
	5e-5	52.85	82.52	82.72	88.72	90.06	67.15	92.09	88.60
Our OT + Classifier	2e-5	55.73	84.35	84.64	87.15	90.54	68.23	91.74	89.32
	3e-5	58.30	83.88	84.13	87.46	91.58	66.79	92.09	88.63
	4e-5	55.22	83.24	83.79	88.71	90.33	65.70	92.43	88.59
	5e-5	54.48	81.63	82.32	88.56	90.33	70.40	90.37	88.97
Our POT + Classifier	2e-5	58.44	84.15	84.52	88.13	91.43	66.79	93.12	89.88
	3e-5	60.95	83.43	84.54	89.25	90.96	70.04	92.20	88.75
	4e-5	58.07	82.36	83.14	87.69	90.76	72.56	92.09	89.11
	5e-5	52.50	82.59	83.08	88.51	90.32	68.59	91.51	88.71

Table 7: Downstream task results different approaches when training the Visual grounding module.

Optimal transport?	Classification	CoLA	MNLI	MNLI-MM	MRPC	QNLI	RTE	SST-2	STS-B	SQuAD V1.1	SQuAD V2.0
No	Yes	58.05	83.93	84.12	88.64	91.20	69.31	92.78	89.37	78.34/86.42	70.14/73.71
Classical	No	60.85	84.17	84.85	88.68	90.87	72.20	92.43	89.45	77.46/85.88	67.67/71.4
Partial	No	58.20	**84.38**	**84.88**	**89.25**	90.99	69.31	92.66	89.27	77.99/86.35	68.29/71.65
Classical	Yes	58.30	84.35	84.64	88.71	**91.58**	70.40	92.43	89.32	78.14/86.42	69.00/72.24
Partial	Yes	**60.95**	84.15	84.54	**89.25**	91.43	**72.56**	**93.12**	**89.88**	**78.49/86.62**	**70.69/73.92**

We also fine-tune all models on 9 different natural-language tasks of GLUE and SQuAD datasets. To have a fair comparison, all models are initialized with the pretrained BERT weights, except LXMERT that is pretrained from scratch. As shown in Table 4, the finetuning results on our model consistently outperform other pretrained models in all tasks. The results show that finetuning the BERT model will make it forget the original knowledge learned from a huge language corpus.

6 ANALYSIS

6.1 The impact of visual grounding

To understand the impact of visual grounding on text representation, we train GroundedBERT without using visual information and the weights of the Visual grounding module are randomly initialized. The results in Table 5 show that the visual information has the significant contribution in the language grounding and is beneficial to the textual representation. We also study the impact of the contribution visual grounding with different visual embedding dimensions. Since the dimension of the hidden representation of language encoder, which is BERT, is fixed depend on its configuration, we can setup the dimension of additional visual information flexibly.

6.2 Different learning rates on GLUE

Following the setting in BERT [11] on GLUE tasks, we also conduct additional experiments with more runs on different learning rates similar to the BERT paper. The learning rates are also similarly set to be $\{2, 3, 4, 5\}e-5$ to have a fair comparison. The results in Table 6 show that our model consistently outperforms the baseline in all datasets for all learning rates.

6.3 Different training strategy

We conduct the experiments on GroundedBERT trained with different settings: Optimal transport and Classifier. Table 7 reports evaluations of our model on GLUE and SQUAD on 5 different approaches, i.e., only Classifier, only Classical OT (OT), only Partial OT, Classifier + OT and Classifier + POT. The results show that the combination of Classifier and Partial OT achieves the highest score in most tasks, while Partial OT perform better than Classical OT in both combination with Classifier or not.

7 CONCLUSION

In this paper, we propose GroundedBERT as a grounded language learning model that incorporates visual information into BERT representation. We introduce the visual grounding module to capture the visual information which is later joined with the text representation to create a unified visual-textual representation. Our model significantly outperforms the baseline language models on various language tasks of the GLUE and SQuAD datasets.

8 ACKNOWLEDGEMENT

This research is supported by AI Singapore technology grant AISG2-TC-2022-005.

REFERENCES

[1] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph P. Turian. 2020. Experience Grounds Language. *CoRR* abs/2004.10151 (2020). arXiv:2004.10151 https://arxiv.org/abs/2004.10151

[2] Patrick Bordes, Éloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. Incorporating Visual Semantics into Sentence Representations within a Grounded Space. In *EMNLP*.

[3] 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. 2020. Language Models are Few-Shot Learners. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[4] Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Association for Computational Linguistics, Vancouver, Canada, 1–14. https://doi.org/10.18653/v1/S17-2001

[5] Liqun Chen, Zhe Gan, Yu Cheng, Linjie Li, Lawrence Carin, and Jingjing Liu. 2020. Graph Optimal Transport for Cross-Domain Alignment. *CoRR* abs/2006.14744 (2020). arXiv:2006.14744 https://arxiv.org/abs/2006.14744

[6] Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Improving sequence-to-sequence learning via optimal transport. *arXiv preprint arXiv:1901.06283* (2019).

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: Learning UNiversal Image-TExt Representations. *CoRR* abs/1909.11740 (2019). arXiv:1909.11740 http://arxiv.org/abs/1909.11740

[8] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*.

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

[10] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 670–680.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[12] William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*. https://www.aclweb.org/anthology/I05-5002

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *CoRR* abs/2010.11929 (2020). arXiv:2010.11929 https://arxiv.org/abs/2010.11929

[14] Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. 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*. 1367–1377. http://aclweb.org/anthology/N/N16/N16-1162.pdf

[15] Alexander G. Ororbia II, Ankur Arjun Mali, Matthew A. Kelly, and David Reitter. 2018. Visually Grounded, Situated Learning in Neural Models. *CoRR* abs/1805.11546 (2018). arXiv:1805.11546 http://arxiv.org/abs/1805.11546

[16] Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. 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)*. 408–418.

[17] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*. PMLR, 5583–5594.

[18] Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. 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*. 3294–3302. http://papers.nips.cc/paper/5950-skip-thought-vectors

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. https://arxiv.org/abs/1602.07332

[20] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *International Conference on Learning Representations*.

[21] Angeliki Lazaridou, Marco Baroni, et al. 2015. 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*. 153–163.

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

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

[24] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. *arXiv preprint arXiv:2004.06165* (2020).

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

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* (2019).

[27] Lajanugen Logeswaran and Honglak Lee. 2018. 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*. https://openreview.net/forum?id=rJvJXZb0W

[28] Ilya Loshchilov and Frank Hutter. 2018. Fixing weight decay regularization in adam. (2018).

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

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

[31] Thong Nguyen and Anh Tuan Luu. 2021. Contrastive Learning for Neural Topic Model. *CoRR* abs/2110.12764 (2021). arXiv:2110.12764 https://arxiv.org/abs/2110.12764

[32] Thong Nguyen and Luu Anh Tuan. 2021. Improving Neural Cross-Lingual Summarization via Employing Optimal Transport Distance for Knowledge Distillation. *CoRR* abs/2112.03473 (2021). arXiv:2112.03473 https://arxiv.org/abs/2112.03473

[33] Thong Nguyen, Xiaobao Wu, Xinshuai Dong, Anh Tuan Luu, Cong-Duy Nguyen, Zhen Hai, and Lidong Bing. 2023. Gradient-Boosted Decision Tree for Listwise Context Model in Multimodal Review Helpfulness Prediction. *arXiv preprint arXiv:2305.12678* (2023).

[34] Thong Nguyen, Xiaobao Wu, Anh-Tuan Luu, Cong-Duy Nguyen, Zhen Hai, and Lidong Bing. 2022. Adaptive Contrastive Learning on Multimodal Transformer for Review Helpfulness Predictions. *arXiv preprint arXiv:2211.03524* (2022).

[35] William O'Grady. 2005. *How Children Learn Language*. Cambridge University Press. https://www.cambridge.org/core/books/how-children-learn-language/04C336554C93315A5F78F4E03777A4E6

[36] Lin Pan, Chung-Wei Hang, Avirup Sil, Saloni Potdar, and Mo Yu. 2021. Improved Text Classification via Contrastive Adversarial Training. *arXiv preprint arXiv:2107.10137* (2021).

[37] Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501* (2021).

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

[39] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. 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)*. 2227–2237.

[40] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. 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)*. 784–789.

[41] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.

[42] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *CoRR* abs/1908.10084 (2019). arXiv:1908.10084 http://arxiv.org/abs/1908.10084

[43] Brett D. Roads and Bradley C. Love. 2019. Learning as the Unsupervised Alignment of Conceptual Systems. *CoRR* abs/1906.09012 (2019). arXiv:1906.09012 http://arxiv.org/abs/1906.09012

[44] Jacqueline Sachs, Barbara Bard, and Marie L Johnson. 1981. Language learning with restricted input: Case studies of two hearing children of deaf parents. *Applied Psycholinguistics* 2, 1 (1981), 33–54. https://www.cambridge.org/core/journals/applied-psycholinguistics/article/language-learning-with-restricted-input-case-studies-of-two-hearing-children-of-deaf-parents/4F5BF799996DCD5977A94BC5F1233578

[45] Lei Shi, Kai Shuang, Shijie Geng, Peng Su, Zhengkai Jiang, Peng Gao, Zuohui Fu, Gerard de Melo, and Sen Su. 2020. Contrastive Visual-Linguistic Pretraining. *CoRR* abs/2007.13135 (2020). arXiv:2007.13135 https://arxiv.org/abs/2007.13135

[46] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://www.aclweb.org/anthology/D13-1170

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

[48] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5103–5114.

[49] Hao Tan and Mohit Bansal. 2020. Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision. *CoRR* abs/2010.06775 (2020). arXiv:2010.06775 https://arxiv.org/abs/2010.06775

[50] Gabriella Vigliocco, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: implications for language learning, processing and evolution. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4123671/

[51] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. *CoRR* abs/1804.07461 (2018). arXiv:1804.07461 http://arxiv.org/abs/1804.07461

[52] Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2018. Neural Network Acceptability Judgments. *CoRR* abs/1805.12471 (2018). arXiv:1805.12471 http://arxiv.org/abs/1805.12471

[53] Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. 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)* (New Orleans, Louisiana). Association for Computational Linguistics, 1112–1122. http://aclweb.org/anthology/N18-1101

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

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