How Good are these Visio-Linguistic models? An Evaluation of Pre-trained Multimodals on Challenging Datasets

Anonymous ACL submission

Abstract

Numerous recent works have proposed pre-training generic visio-linguistic representations and then finetuning them for downstream vision and language tasks. VilBert and VisualBert are such two popular multimodals that push new state-of-the-art (SOTA) performances on various vision and language tasks. In this work, we evaluate VilBert and VisualBert on three challenging vision and language datasets and compare them with unimodally pre-trained multimodals. We conclude that visio-linguistic representations are generalizable across different types of vision language tasks and outperform previous multimodals that are not jointly pre-trained.

1 Introduction

Pre-trained models like BERT have been proven effective in learning representations of natural language sentences and achieved new state-of-the-art performance on various natural language understanding tasks like question answering and language inference (Devlin et al., 2018). Seeing the impact of BERT has made in the NLP community, researchers have extended BERT to design frameworks that learn visio-linguistic representations for modeling a broad range of vision-and-language tasks. tasks.

VilBert and VisualBert are two recently proposed models that learn task-agnostic joint representations of image content and natural language (Lu et al., 2019; Li et al., 2019). The authors pretrained them on large vision-and-language datasets and transferred them to multiple established vision-and-language tasks – visual question answering, visual commonsense reasoning, referring expressions, and caption-based image retrieval – by making only minor additions to the base architecture. These models proved significant improvements on question answering, vision-language reasoning, and referring expressions tasks.

Although VilBert and VisualBert have shown distinct improvements in some vision-language tasks, some of the tasks are well-studied and relatively easier comparing to other more challenging tasks that, for instance, require reasoning and inference. Will pretrained multimodals still achieve SOTA performance on other more challenging vision-language tasks? Besides, jointly pretrained visio-linguistic models like VilBert and VisualBert outperform unimodally pretrained multimodals in lots of vision-language tasks like VQA (Agrawal et al., 2015). However, it is more computationally expensive and requires a large dataset of parallel data of texts and images to train these models, whereas two separate datasets of texts and images are often available. Will jointly pretrained multimodals also perform better than multimodals that are unimodally pretrained on these new datasets?

In this work, we evaluate VilBert and VisualBert on three recent vision-language tasks: visual entailment, multilabel movie genre prediction, and hateful memes detection. We specifically selected these three tasks and datasets as they represented different aspects of difficulties a vision-language system may face in real life: visual entailment requires inference, movie genre prediction is a multilabel task, and hateful memes detection require subtle reasoning that rely on multimodal signals. We finetuned pre-trained VilBert and VisualBert on these three datasets and compared them with the best results reported in these dataset papers that are achieved using unimodally pretrained multimodals. We found that VilBert and VisualBert outperform the previous best models even on these more challenging datasets.

In summary, our contributions of this paper are two-fold:

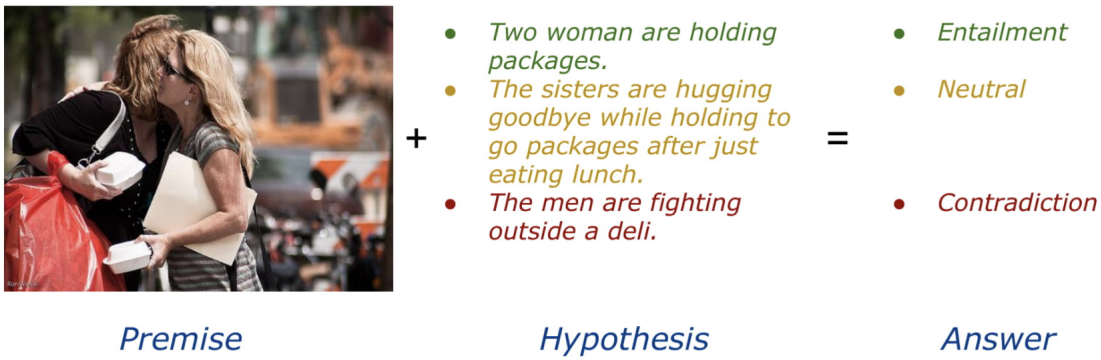


Figure 1: A sample data of the SNLI-VE dataset taken from (Xie et al., 2019).

- We evaluate VilBert and VisualBert on three recently proposed vision and language datasets and conclude that these models are generalizable and achieve new SOTA for these datasets.
- We compare these results with previous best results achieved by unimodally trained multimodals and believe that joint pretraining yields better results than pretraining visual and language representations separately.

2 Related Work

Jointly Pre-trained Multimodals Numerous recent works have proposed pretraining generic visio-linguistic representations and then finetuning them for downstream vision and language tasks. Besides VilBert and VisualBert, one other popular framework is LXMERT, short for Learning Cross-Modality Encoder Representations from Transformers (Tan and Bansal, 2019). In LXMERT, they build a large-scale Transformer model that consists of three encoders: an object relationship encoder, a language encoder, and a cross-modality encoder. Next, to endow their model with the capability of connecting vision and language semantics, they pre-train the model with large amounts of image-and-sentence pairs, via five diverse representative pre-training tasks. After fine-tuning, their model shows new state-of-the-art results on two visual question answering datasets (i.e., VQA and GQA). They also show the generalizability of their pre-trained cross-modality model by adapting it to a challenging visual-reasoning task, NLVR2 , and improve the previous best result by 22% absolute (54% to 76%). Another example of visio-linguistic model is DeVLBert, in which the authors propose a Deconfounded Visio-Linguistic Bert framework to study the problem of out-of-domain visio-linguistic pretraining (Zhang et al., 2020).

Unimodally Pre-trained Multimodals Explainable Visual Entailment System (EVE) is an architecture based on the Attention Top-Down/BottomUp model. Similar to the Attention Top-Down/Bottom-Up, their EVE architecture is composed of a text and an image branch. The text branch extracts features from the input text hypothesis H_{text} through an RNN. The image branch generates image features from P_{image}. The features produced from the two branches are then fused and projected through fully-connected (FC) layers towards predicting the final conclusion. EVE-Image, which is a variant of EVE architecture and achieves the previous best results on the SNLI-VE task, is such multimodal that the two branches, text and image, are pre-trained separately using different corpus. The image features are configured to take the feature maps from a pre-trained convolutional neural network, where as the text representations are using the pre-trained Glove embeddings (Xie et al., 2019).

3 Datasets

SNLI-VE: Visual Entailment Dataset Visual Entailment (VE) is a new inference task consisting of image-sentence pairs whereby a premise is defined by an image, rather than a natural language sentence as in traditional Textual Entailment tasks. The goal of a trained VE model is to predict whether the image semantically entails the text, which requires finegrained reasoning in real-world settings. Figure 1 shows an example of instances in the SNLI-VE dataset.

The SNLI-VE dataset is generated based on SNLI and Flickr30k datasets. Given an image and

Figure 2: A sample data of the MMIMDB dataset taken from (Arevalo et al., 2017)

Figure 3: A sample data of the Hateful Memes dataset taken from (Kiela et al., 2020)

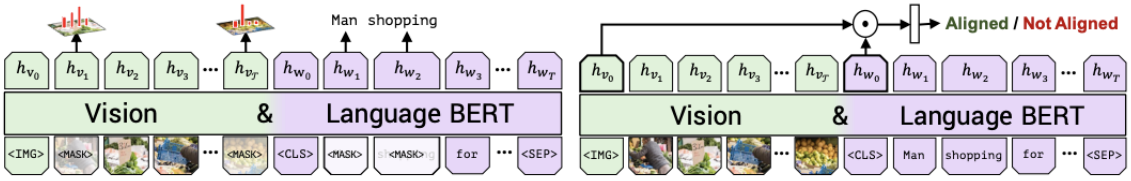
a natural language statement, the visual entailment task involves classifying whether the statement is true (entailment), false (contradiction) or neutral w.r.t. to the image. The dataset contains 550K image/statement pairs and evaluation is done using classification accuracy (Xie et al., 2019).

MMIMDB: Multimodal IMDB Dataset This dataset proposes a task of multilabel movie genre prediction based on its plot and image poster. Figure 2 shows a sample data point from the MMIMDB dataset.

The MM-IMDB(Multi Modal IMDB) dataset consists of 26K movie plot outlines and movie posters. Each plot contains on average 92.5 words, while the longest one contains 1, 431 words and the average of genres per movie is 2.48. This is a multilabel prediction problem, i.e., one movie can have multiple genres and we use micro-F1 and macro-F1 as evaluation metrics following (Arevalo et al., 2017).

Hateful Memes Dataset This dataset proposes a hate speech detection task in multimodal memes. Memes pose an interesting multimodal fusion problem: Consider a sentence like "love the way you smell today" or "look how many people love you". Unimodally, these sentences are harmless, but combine them with an equally harmless image of a skunk or a tumbleweed, and suddenly they become mean. That is to say, memes are often subtle and while their true underlying meaning may be easy for humans to detect, they can be very challenging for AI systems. Figure 3 shows a data sample from the Hateful memes dataset.

The dataset contains exactly 10k memes. The dataset comprises five different types of memes: multimodal hate, where benign confounders were found for both modalities, unimodal hate where one or both modalities were already hateful on their own, benign image and benign text confounders and finally random not-hateful examples. The objective of the task is, given an image and the pre-extracted—i.e., it does not require OCR—text, to classify memes according to their hatefulness. The results are evaluated using ROC AUC and classification accuracy



(a) Masked multi-modal learning (b) Multi-modal alignment prediction

Figure 4: They train ViLBERT on the Conceptual Captions dataset (Sharma et al., 2018) under two training tasks to learn visual grounding. In masked multi-modal learning, the model must reconstruct image region categories or words for masked inputs given the observed inputs. In multi-modal alignment prediction, the model must predict whether or not the caption describes the image content. This image is taken from (Lu et al., 2019)

4 Models

In this paper, we evaluate two similar but different visio-linguistic models that are pretrained on large amounts of vision-language parallel data, VilBert and VisualBert.

VilBert Inspired by BERT's success at language modeling, researchers started to develop analogous models and training tasks to learn joint representations of language and visual content from paired data. Specifically, they consider jointly representing static images and corresponding descriptive text (Lu et al., 2019).

One straightforward approach is to make minimal changes to BERT – simply discretizing the space of visual inputs via clustering, treat these visual 'tokens' exactly like text inputs, and start from a pretrained BERT model (Devlin et al., 2018). This architecture suffers from a number of drawbacks. First, initial clustering may result in discretization error and lose important visual details. Second, it treats inputs from both modalities identically, ignoring that they may need different levels of processing due to either their inherent complexity or the initial level of abstraction of their input representations. For instance, image regions may have weaker relations than words in a sentence and visual features are themselves often already the output of a very deep network. Finally, forcing the pretrained weights to accommodate the large set of additional visual 'tokens' may damage the learned BERT language model. Instead, authors of VilBert develop a two-stream architecture modelling each modality separately and then fusing them through a small set of attention-based interactions. This approach allows for variable network depth for each modality and enables cross-modal connections at different depths.

The architecture of ViLBERT is shown in Fig 4 and consists of two parallel BERT-style models operating over image regions and text segments. Each stream is a series of transformer blocks (TRM) and novel co-attentional transformer layers (Co-TRM) which they introduce to enable information exchange between modalities. Given an image I represented as a set of region features $v_1, ..., v_T$ and a text input $w_0, ..., w_T$, VilBert outputs final representations $h_{v0}, ..., h_{vT}$ and $h_{w0}, ..., h_{wT}$. Notice that exchange between the two streams is restricted to be between specific layers and that the text stream has significantly more processing before interacting with visual features – matching their intuitions that their chosen visual features are already fairly high-level and require limited context-aggregation compared to words in a sentence (Lu et al., 2019).

VisualBert is a single stream BERT model (Devlin et al., 2018) with multiple transformer blocks (Li et al., 2019). The image regions's embeddings concatenated with textual embeddings are the input to the model in a similar fashion as BERT but twice as wide. The image embeddings are computed by adding image region embeddings, image positional embeddings and a specific embedding which distinguishes it from the text embeddings. Figure 5 shows the architecture of VisualBert.

The key of their idea is to re-use the self-attention mechanism within the Transformer to align elements of the input text and regions in the input image. In addition to BERT, they introduce a set of visual embeddings, F, to model an image. Each $f \in F$ corresponds to a bounding region in the image, derived from an object detector (Li et al., 2019).

Each embedding in F is computed by summing three embeddings: (1) a visual feature representation of the bounding region of f, computed by a convolutional neural network, (2) a segment em-

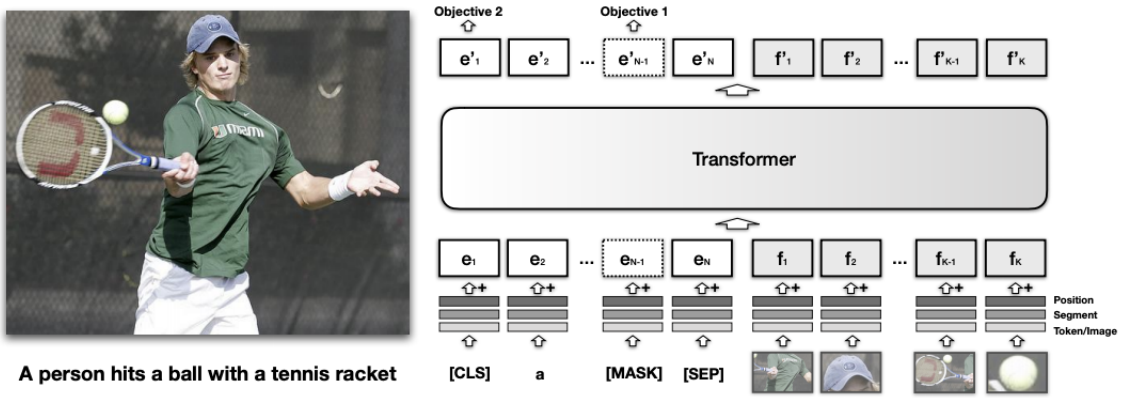


Figure 5: The architecture of VisualBERT. Image regions and language are combined with a Transformer to allow the self-attention to discover implicit alignments between language and vision. It is pre-trained with a masked language modeling, and sentence-image prediction task, on image captioning dataset and then fine-tuned for different tasks. Image taken from (Li et al., 2019).

bedding indicating it is an image embedding as opposed to a text embedding, and (3) a position embedding, which is used when alignments between words and bounding regions are provided as part of the input, and set to the sum of the position embeddings corresponding to the aligned words. The visual embeddings are then passed to the multi-layer Transformer along with the original set of text embeddings, allowing the model to implicitly discover useful alignments between both sets of inputs, and build up a new joint representation (Li et al., 2019).

5 Experiment

5.1 Implementation Details

We use the mmf framework[1] to implement and fine-tune the two visio-linguistic models on the three datasets. MMF is a modular framework for vision and language multimodal research from Facebook AI Research. MMF contains reference implementations of state-of-the-art vision and language models and has powered multiple research projects at Facebook AI Research.

We use VilBert pre-trained on the Conceptual Captions (CC) dataset and VisualBert pre-trained on the COCO dataset as our models for evaluations. Both models are then finetuned under the default settings provided by the framework, which are consistent with the architecture from the two models original papers. No hyper-parameters tuning process was performed. However, some training configurations were adjusted due to the limited computing resources. Specifically, both VilBert and VisualBert are trained under batch size of 16 (8 for MMIMDB dataset due to CUDA memory limit), 0 num workers, and 180000 max updates. We decrease the batch size from the original 480 so it can fit into the CUDA memory our GPU has, and increase the max updates to keep the training quality. The models are all trained on GTX 1080 Ti GPUs with ~12 GB of memory. We used an adamw optimizer with learning rate 5e-5. The average training time for finetuning one VisualBert model on a dataset is ~20 hours and for finetuning a VilBert model on a dataset is ~30 hours.

5.2 Results

The evaluation results on local validation set are reported in Table 1. The previous state-of-the-art (SOTA) results are the highest scores taken from these three dataset papers that reported them (except for hateful memes dataset since the model that reached the highest score is VisualBert COCO which we also evaluated so we took the second best model). The three models that reached these previous best scores are EVE-Image, Gated Multimodal Units (GMU), and VilBert without pretraining on CC. These are all multimodals that are unimodally pre-trained as we described before in section 2.

5.3 Discussion

We can see that both jointly pre-trained multimodals VilBert and VisualBert outperform the pre-

[1] https://github.com/facebookresearch/mmf

	SNLI-VE	MMIMDB		HATEFUL MEMES	
Metric	Acc	Macro F1	Micro F1	Acc	AUROC
PREV SOTA	71.40	54.10	63.00	62.20	71.13
VILBERT CC	74.75	**57.80**	**66.16**	67.22	**72.06**
VISUALBERT COCO	**77.43**	57.43	65.90	**69.44**	70.74

Table 1: Evaluation Results

vious SOTA models. It shows that VilBert and VisualBert not only can achieve SOTA performance on previous well-studied tasks like VQA, but also can reach SOTA on recently proposed tasks and datasets that are more challenging. It demonstrates the generalizability of visio-linguistic representations across different types of vision-language tasks.

Besides, comparing to the previous best results reported in these datasets' paper, which are all achieved by multimodals pre-trained separately, jointly pre-trained multimodals indeed yield better results for these more challenging dataset. It shows that jointly pre-trained multimodals are generally better choices over unimodally pre-trained multimodals, given that a pre-trained model is available or there's enough parallel data to pretrain the models.

Another side product of these evaluations is that, the finetuning process of VilBert is less stable and more expensive than of VisualBert. During our first training with VilBert on the SNLI-VE dataset, the local validation accuracy score is only ~64 %. We believe that this could be due to the stochasticity nature of neural networks and finetuned it again with exactly the same parameters. It then reaches comparable performance with VisualBert as we reported in the table. Besides, the finetuning process of VilBert requires more computing resources, i.e. more memory for the GPU, and also takes about ~1.5x longer than VisualBert when training on the same data. So we recommend trying VisualBert first before training VilBert as their performances are comparable but one is computationally more expensive.

6 Conclusion

To conclude, pre-trained visio-linguistic models are often a good starting point for most of the vision-language tasks. They are generalizable and can produce SOTA results across different types of vision and language tasks. Jointly pretrained multimodals like VisualBert should be your go-to choice over unimodally pretrained multimodals whenever given a vision-language task if viable. We believe that visio-linguistic models is one possible key solution to vision and language tasks, and there will be more research focused on building upon them or designing variants of them.

Acknowledgments

Great thanks to Professor Mark Yatskar for his helpful discussions and suggestions.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. 2015. VQA: Visual Question Answering. *arXiv e-prints*, page arXiv:1505.00468.

John Arevalo, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. 2017. Gated Multimodal Units for Information Fusion. *arXiv e-prints*, page arXiv:1702.01992.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv e-prints*, page arXiv:1810.04805.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *arXiv e-prints*, page arXiv:2005.04790.

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 e-prints*, page arXiv:1908.03557.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv e-prints*, page arXiv:1908.02265.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *arXiv e-prints*, page arXiv:1908.07490.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual Entailment: A Novel Task for Fine-Grained Image Understanding. *arXiv e-prints*, page arXiv:1901.06706.

Shengyu Zhang, Tan Jiang, Tan Wang, Kun Kuang, Zhou Zhao, Jianke Zhu, Jin Yu, Hongxia Yang, and Fei Wu. 2020. DeVLBert: Learning Deconfounded Visio-Linguistic Representations. *arXiv e-prints*, page arXiv:2008.06884.