ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks
- Jiasen Lu et al. (NeurIPS 2019)

Presented by - Chinmoy Samant, cs59688
Overview

- Introduction
- Motivation
- Approach
  - BERT
  - ViLBERT
- Implementation details
- Results
  - Quantitative
  - Qualitative
- Critique
- Follow-up work
- Concurrent work
INTRODUCTION AND MOTIVATION
Vision and Language Tasks - Introduction

**VQA**
Visual Question Answering

- "Is there something to cut the vegetables with?"

**Image Captioning**

- "construction worker in orange safety vest is working on road."
- "two young girls are playing with lego toy."

**VCR Q→A**
Visual Commonsense Reasoning

- "Why is person A pointing at person B?"

**VCR QA→R**

- "Rationale: a) is correct because..."
- "Person A is pointing at the person in front of him."
- "Person B is facing everyone's order and asked for clarification."
- "Person C is looking at the person in front of him and person E is smiling slightly."
- "Person D is delivering food to the table, and the right side wears white shirt to wheelchair."
Vision and Language Tasks - Common Approach

- Visual Question Answering
- Image Captioning
- Visual Commonsense Reasoning
- Referring Expression
Q: What type of plant is this?  
A: Banana

C: A bunch of red and yellow flowers on a branch.

**Failure in Visual Grounding!**

(Common model for visual grounding and leverage them on a wide array of vision-and-language tasks)
Motivation for Pretrain->Transfer

Step 1 - Dataset

- Image Classification

Step 2 - Pretrain

- Object Detection

Step 3 - Transfer

- Semantic Segmentation
- Question Answering
- Sentiment Analysis
Motivation for Pretrain->Transfer

Step 1 - Dataset

Step 2 - Pretrain

Step 3 - Transfer

Image Classification

Object Detection

Semantic Segmentation

Question Answering

Sentiment Analysis
Dataset - Conceptual Captions

- ~3.3 million image/caption pairs
- created by automatically extracting and filtering image caption annotations from web pages
- Measured by human raters to have ~90% accuracy
- Wider variety of image-caption styles as the captions are extracted from web
APPROACH
Proposed Vision and Language BERT (ViLBERT), a joint model for learning task-agnostic visual grounding from paired visio-linguistic data.

Based on top of BERT architecture.

- **Key technical innovation?**
  - Separate streams for vision and language processing that communicate through co-attentional transformer layers.

- **Why?**
  - Separate streams can accommodate the differing processing needs of each modality
  - Co-attentional layers provide interaction between modalities at varying representation depths.

- **Result?**
  - Demonstrated that this structure outperforms a single-stream unified model across multiple tasks.
First we BERT, then we ViLBERT!

To have a better understanding of ViLBERT architecture, let’s first understand how BERT and more generally how transformers work.
BERT (Bidirectional Encoder Representations from Transformers)

- BERT is an attention-based bidirectional language model.
- Pretrained on a large language corpus, BERT can learn effective and generalizable language models.
- Proven to be very effective for transfer learning to multiple NLP tasks.
- Composed of multiple transformer encoders as building blocks.
Transformer

Transformer encoder
Transformer

Transformer encoder

- Multi-headed self attention
  - Models context
Transformer

Transformer encoder

- Multi-headed self attention
  - Models context
- Feed-forward layers
  - Computes nonlinear hierarchical features
Transformer

Transformer encoder

- **Multi-headed self attention**
  - Models context
- **Feed-forward layers**
  - Computes nonlinear hierarchical features
- **Layer norm and residuals**
  - Makes training easier and more stable
Transformer encoder

- Multi-headed self attention
  - Models context
- Feed-forward layers
  - Computes nonlinear hierarchical features
- Layer norm and residuals
  - Makes training easier and more stable
- Positional embeddings
  - Allows model to learn the relative positioning
Like the transformer encoder, BERT takes a sequence of words as input. Each layer applies self-attention, passes it through a feed-forward network, and sends it to the next encoder. Each position outputs a vector of size = hidden_size (768 in BERT\textsubscript{Base}). Can use all or a set of these outputs to perform different NLP tasks.
Let's look at spam detection task as an example.

For this task, we focus on the output of only the first position.

That output vector can now be used as the input for any spam detection classifier.

Papers have achieved great results by just using a single-layer neural network as the classifier.
BERT Training

● Next important aspect - How to train BERT?

● Choosing pretraining tasks crucial to ensure that it learns a good language model.

● BERT is pretrained on the following two tasks:
  ○ Masked Language Modeling (MLM)
  ○ Next Sentence Prediction (NSP)

Let's look at these two tasks as well as how they inspired the pre-training tasks for ViLBERT model.
Masked Language Modeling (MLM)

- Randomly divide input tokens into masked $X_M$ and observed $X_O$ tokens (approximately 15% of tokens being masked).
**Masked Language Modeling (MLM)**

- Masked tokens replaced with a special MASK token 80% of the time, a random word 10%, and unaltered 10%.
- BERT model then trained to reconstruct these masked tokens given the observed set.
MLM-inspired masked multi-modal learning for visiolinguistic tasks

ViLBERT model must reconstruct image region categories or words for masked inputs given the observed inputs.
In next sentence prediction task, BERT model is passed two text segments A and B following the format shown and is trained to predict whether or not B follows A in the source text.

In a sense, this is equivalent to modeling if Sentence B aligns with Sentence A or not.
NSP-inspired pretraining for visiolinguistic tasks

ViLBERT model must predict whether or not the caption describes the image content.
BERT v/s ViLBERT

- One may ask:
  - Why do we need ViLBERT with two separate streams for vision and language?
  - Why can't we use same BERT architecture with image as additional inputs?
- Because different modalities may require different level of abstractions.

Linguistic stream:

Visual stream:

```
Original Alt-Text: “Ray Manzarek and Robbie Krieger perform on stage.”
Conceptual Caption: “hard rock artists perform on stage.”
```
Solution - ViLBERT

Two-stream model which process visual and linguistic separately.

Different number of layers in each stream, $k$ in vision, $l$ in language.
Fusing different modalities

- Problem solved till now -
  - Multi-stream BERT architecture that can model visual as well as language information effectively.

- Problem remaining -
  - Learning visual grounding by fusing information from these two modalities

- Solution -
  - Use co-attention - [proposed by Lu et al. 2016] to fuse information between different sources.
Co-Transform (Co-TRM) layer
Co-Transform (Co-TRM) layer
Co-Attentional Transformer

- Same transformer encoder-like architecture but separate weights for visual and linguistic stream.
- Transformer encoder with query from another modality. Visual stream has query from Language and Linguistic stream has query from vision.
- Aggregate information with residual add operation.
Pre-training objectives

Masked multi-modal modelling

- Follows masked LM in BERT.
- 15% of the words or image regions to predict.
- Linguistic stream:
  - 80% of the time, replace with [MASK].
  - 10% of the time, replace random word.
  - 10% of the time, keep same.
- Visual stream:
  - 80% of the time, replace with zero vector.

Multi-modal alignment prediction

- Predict whether image and caption is aligned or not.
Image Representation

- Faster R-CNN with Res101 backbone.
- Trained on Visual Genome dataset with 1600 detection classes.
- Select regions where class detection probability exceeds a confidence threshold.
- Keep between 10 to 36 high-scoring boxes.
- Output = Sum of region embeddings and location embeddings.
- Transformer and co-attentional transformer blocks in the visual stream have hidden state size of 1024 and 8 attention heads.
Text Representation

- BERT language model pretrained on BookCorpus and English Wikipedia.
- BERT\textsubscript{BASE} model - 12 layers of transformer blocks, each block’s hidden state size - 762 and 12 attention heads.
- Output is sum of three embeddings: Token embeddings + Segment embeddings + Position embeddings.
Training details

- 8 TitanX GPUs - total batch size of 512 for 10 epochs.
- Adam optimizer with initial LR of $10^{-4}$. Linear decay LR scheduler with warm up to train the model.
- Both training task losses are weighed equally.
Experiments - Vision-and-Language Transfer Tasks

**Visual Question Answering (VQA)**

**Visual Commonsense Reasoning (VCR)**

**Caption-Based Image Retrieval**

**Referring Expressions**

- A large bus sitting next to a very tall building.
- Guy in yellow dribbling ball.

Example questions and image descriptions:

- Is there something to cut the vegetables with?
- VQA: Visual Question Answering
- VCR Q→A: Visual Commonsense Reasoning
- VCR QA→R: Referring Expressions
Transfer learning details

● **Common Fine Tuning strategy** -
  ○ Modify the pretrained base model to perform new task, then train entire model end-to-end.
  ○ In all cases, the modification is trivial – typically learning a classification layer.

● **Task specific details** -
  ○ Visual Question Answering (VQA) -
    ■ VQA 2.0 dataset, 2-layer MLP on top, multi-class classification task.
  ○ Visual Commonsense Reasoning (VCR) -
    ■ VCR dataset, linear layer to predict score for question-response pair, then softmax.
  ○ Grounding Referring Expressions -
    ■ RefCOCO+ dataset, rerank a set of image region proposals using referring expression.
  ○ Caption-Based Image Retrieval -
    ■ Fine-tuned on Flickr30k dataset, 4-way prediction task.
  ○ ‘Zero-shot’ Caption-Based Image Retrieval -
    ■ Perform Caption-Based Image Retrieval on Flickr30k, without fine-tuning on Flickr30k dataset.
Models considered

- ViLBERT - Main model
- Baselines
  - Single-Stream -
    - Single BERT architecture that processes both modality inputs through the same set of transformer blocks – sharing parameters and processing stacks for both visual and linguistic inputs.
    - This baseline establishes the impact of two-stream architecture.
  - ViLBERT† -
    - ViLBERT architecture that has not undergone their pre training tasks. Still has BERT initialization for linguistic stream and represents image regions with the same Faster R-CNN model as the full ViLBERT model.
    - This baseline helps isolate gains over task-specific baseline models that might be due to the architecture, language initialization, or visual features as opposed to the pretraining process on Conceptual Captions.
RESULTS
QUANTITATIVE RESULTS
## Results - Table

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>test-dev (test-std)</td>
<td>Q→A</td>
<td>QA→R</td>
<td>Q→AR</td>
<td>val</td>
</tr>
<tr>
<td>SOTA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DFAF 36</td>
<td>70.22 (70.34)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>R2C 25</td>
<td>-</td>
<td>63.8 (65.1)</td>
<td>67.2 (67.3)</td>
<td>43.1 (44.0)</td>
<td>-</td>
</tr>
<tr>
<td>MAAttNet 33</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>65.33</td>
</tr>
<tr>
<td>SCAN 35</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Single-Stream†</td>
<td>65.90</td>
<td>68.15</td>
<td>68.89</td>
<td>47.27</td>
<td>65.64</td>
</tr>
<tr>
<td>Single-Stream</td>
<td>68.85</td>
<td>71.09</td>
<td>73.93</td>
<td>52.73</td>
<td>69.21</td>
</tr>
<tr>
<td>ViLBERT†</td>
<td>68.93</td>
<td>69.26</td>
<td>71.01</td>
<td>49.48</td>
<td>68.61</td>
</tr>
<tr>
<td>ViLBERT</td>
<td><strong>70.55 (70.92)</strong></td>
<td><strong>72.42 (73.3)</strong></td>
<td><strong>74.47 (74.6)</strong></td>
<td><strong>54.04 (54.8)</strong></td>
<td><strong>72.34</strong></td>
</tr>
</tbody>
</table>

Transfer task results for ViLBERT model compared with existing state-of-the-art and sensible architectural ablations.
Results - Plot

Full ViLBERT model outperforms task-specific state-of-the-art models across all tasks.

Key Findings -

- Proposed architecture improves performance over a single-stream model.
- Proposed Pretraining tasks result in improved visiolinguistic representations.
- Finetuning from ViLBERT is a powerful strategy for vision-and-language tasks.
QUALITATIVE RESULTS

(NOT INCLUDED IN MAIN PAPER)
A boat covered in flowers near the market.
Co-Attention Visualization - Text to Vision
Co-Attention Visualization - Text to Vision
Co-Attention Visualization - Vision to Text
Co-Attention Visualization - Vision to Text

Layer 0

Layer 5
CRITIQUE
The good

- Learns more consistent visual grounding than models trained for separate tasks.
- ViLBERT Model performance:
  - SOTA results on nearly all experiment tasks using single model, beating many task-specific models.
- Many novel methods proposed:
  - Use of Multiple transformers to deal with differing processing needs of multi-modal information.
  - Use of Co-attention to provide interaction between modalities at various representation depths.
- Dataset selected is large-scale, varied because it's extracted from web and highly accurate as verified by human experts.
- Proposed a common model for visual grounding with exceptional performance on a wide array of vision-and-language tasks with simple fine-tuning schemes.
- Pretraining tasks generalizable to other model architectures as shown in the results.
- Provided insights into the selection of pre-training tasks and verified its effectiveness by achieving superior performance on multiple model architectures.
- Detailed and relevant ablation studies:
  - Ablation studies performed justify all their newly proposed methods.
The not so good

- ViLBERT can still learn inconsistent grounding during task-specific finetuning?
  - A possible solution can be training multiple vision and language tasks together.
- Use language information to help guide vision model extract region features?
  - Current model uses high-scoring region proposals. Extract region proposals for objects in the text?
- Can this idea of using multiple transformer streams extend to other tasks as well?
  - Can we have more than two modalities? Can be used to jointly model text-vision-audio.
  - Can we change each stream’s modality? Can we have text input from two different languages as input to two different streams and learn a joint language model?
- Should have included qualitative results and co-attention visualizations?
  - Provides more insight, helps understand what the co-attention layers are learning.
  - Maybe because of conference-paper length restrictions?
- How to extend it to videos + text rather than static images + text?
  - Possible issues include getting video segments, extracting representations for each segment.
- Both training task losses are weighed equally? Should have explored a weighted approach.
- Experiment with design decisions for co-TRM layers? Do we need them alternating with TRM?
- Authors used BERT$_{\text{BASE}}$ instead of BERT$_{\text{LARGE}}$?
  - Maybe due to training time? Proposed model is huge and slow to train even with lots of GPU resources.
- Improve automatic data collection?
  - Affects model performance? Design automated data checking to remove noisy, less-specific captions.
FOLLOW UP AND CONCURRENT WORK
Follow Up - 12-in-1: Multi-Task Vision and Language Representation Learning

- Follow-up work by the authors of this paper. They test the ViLBERT model on 4 different tasks and 12 different datasets as described below:
- Vocab-based VQA - VQA v2, GQA and Visual Genome (VG) QA datasets.
- Image Retrieval - COCO and Flickr30K captioning datasets.
- Referring Expressions - RefCOCO(+/g), Pointing questions in Visual7W and dialog sequences in the GuessWhat datasets.
- Multi-modal Verification - NLVR$^2$ and SNLI-VE datasets.

<table>
<thead>
<tr>
<th>Clean</th>
<th>Vocab-based VQA (G1)</th>
<th>Image Retrieval (G2)</th>
<th>Referring Expression (G3)</th>
<th>Verification (G4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>VQA</strong></td>
<td><strong>GQA</strong></td>
<td><strong>VG QA</strong></td>
<td><strong>COCO</strong></td>
</tr>
<tr>
<td>1</td>
<td>Single-Task (ST)</td>
<td>71.82</td>
<td>58.19</td>
<td>34.38</td>
</tr>
<tr>
<td>2</td>
<td>Single-Task (ST)</td>
<td>✓ 71.24</td>
<td>59.09</td>
<td>34.10</td>
</tr>
<tr>
<td>3</td>
<td>Group-Tasks (GT)</td>
<td>✓ 72.03</td>
<td>59.60</td>
<td>36.18</td>
</tr>
<tr>
<td>4</td>
<td>All-Tasks (AT)</td>
<td>✓ 72.57</td>
<td>60.12</td>
<td>36.36</td>
</tr>
<tr>
<td>5</td>
<td>All-Tasks w/o G4</td>
<td>✓ 72.62</td>
<td>59.55</td>
<td>36.76</td>
</tr>
<tr>
<td>6</td>
<td>GT finetune→ST</td>
<td>✓ 72.61</td>
<td>59.96</td>
<td>35.81</td>
</tr>
<tr>
<td>7</td>
<td>AT finetune→ST</td>
<td>✓ 72.92</td>
<td>60.48</td>
<td>36.56</td>
</tr>
<tr>
<td>8</td>
<td>AT finetune→ST</td>
<td>✓ 73.15</td>
<td>60.65</td>
<td>36.64</td>
</tr>
</tbody>
</table>
UNITER: Learning Universal Image-Text Representations

Figure 1: Overview of the proposed UNITER model (best viewed in color), consisting of an Image Embedder, a Text Embedder and a multi-layer self-attention Transformer, learned through three pre-training tasks.

(From Microsoft Dynamics 365 AI Research https://arxiv.org/abs/1909.11740)
Figure 2: 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 (Objective 1), and sentence-image prediction task (Objective 2), on caption data and then fine-tuned for different tasks. See §3.3 for more details. (Li et al. 2019) [arXiv:1908.03557]
And many others..

VL-BERT: Pre-training of Generic Visual-Linguistic Representations

Wenjie Su1,2,*, Xizhou Zhu1,2,*, Yue Cao2, Bin Li1, Lewei Lu2, Furu Wei2, Jifeng Dai2
1 University of Science and Technology of China
2 Microsoft Research Asia
{jackroos, ezra0408}@mail.ustc.edu.cn, binli@ustc.edu.cn
{yuecao, lewlu, furuwei, jifdai}@microsoft.com

In short, this is a very hot topic right now!

LXMERT: Learning Cross-Modality Encoder Representations from Transformers

Hao Tan  Mohit Bansal
UNC Chapel Hill
{haotan, mbansal}@cs.unc.edu

Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training

Gen Li1,*, Nan Duan5, Yuejian Fang1, Ming Gong1, Daxin Jiang1, Ming Zhou5
1 School of Software & Microelectronics, Peking University, Beijing, China
5 STCA NLP Group, Microsoft, Beijing, China
{genli, nan.duan, fangyj}@s.pku.edu.cn
{nanduan, milin, djiang, mingzhou}@microsoft.com

Unified Vision-Language Pre-Training for Image Captioning and VQA

Luowei Zhou1,*, Hamid Palangi2, Lei Zhang3, Houdong Hu4, Jason J.C. Corso1, Jianfeng Gao2
1 University of Michigan 2 Microsoft Research 3 Microsoft Cloud & AI 4 Microsoft AI & Research
{luowei.zhou, jjcorso}@umich.edu  {hpalangi, leizhang, henhua, jfgao}@microsoft.com
THANK YOU!

ANY QUESTIONS?
References

- http://ai.google.com/research/ConceptualCaptions