ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks

- Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee

Mitodru Niyogi

Faculty of Mathematics and Computer Science, Heidelberg University, and SAP SE

January 15, 2020

Objective

- To align natural language and visual stimuli, i.e, to perform visual grounding
- **Common approach**: Develop separate language and vision models pretrained for other large-scale tasks and then perform transfer learning to learn grounding as part of task training
- **Aim:**
  - To develop a joint model for learning task-agnostic visual grounding from paired visiolinguistic data
  - Use pretrain for visual grounding instead of pretrain-then-transfer learning approach

Author's key contributions

- Introduces separate streams for vision and language processing that communicate through co-attentional transformer layers
- The model architecture provides interaction between modalities at varying representation depths
- The model reports improvements of 2 to 10 percentage points over the latest state-of-the-art in vision-language task-specific baselines using separately pretrained vision and language models
- **How?**
  - Trained model on Conceptual Captions on two proxy tasks: predicting the semantics of masked words and image regions given the unmasked inputs, and
  - predicting whether an image and text segment correspond

Bidirectional Encoder Representations from Transformers (BERT)

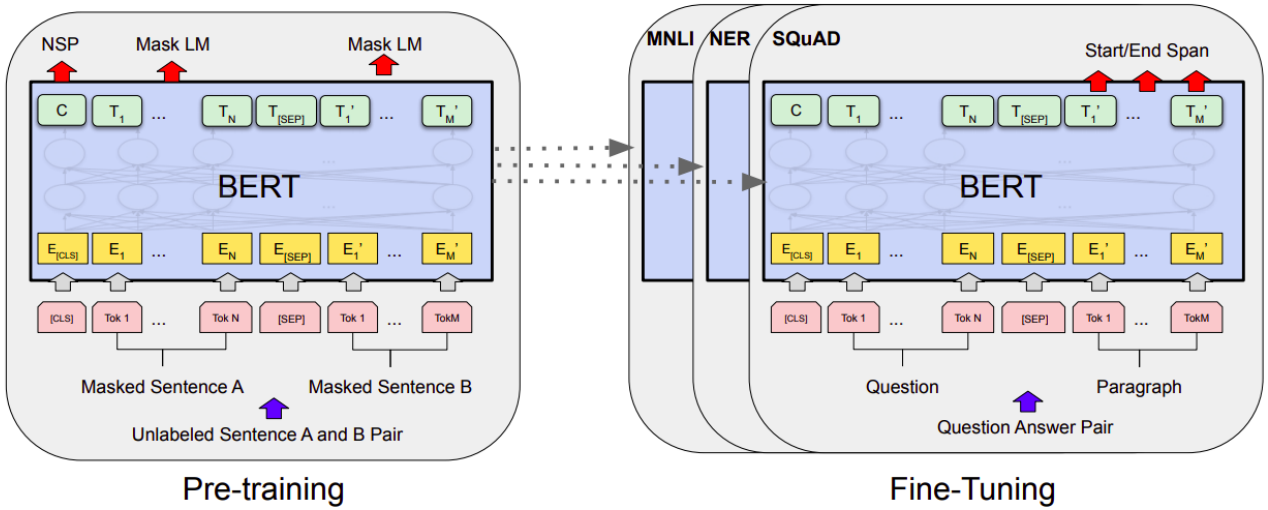


Figure: Overall pre-training and fine-tuning procedures for BERT. Fig taken from [2]

- Attention-based bidirectional language model
- Single encoder-style transformer block consisting of a multi-headed attention block followed by a small fully-connected network
- tokens are mapped to learned encodings and passed through L "encoder-style" transformer blocks to produce final representations

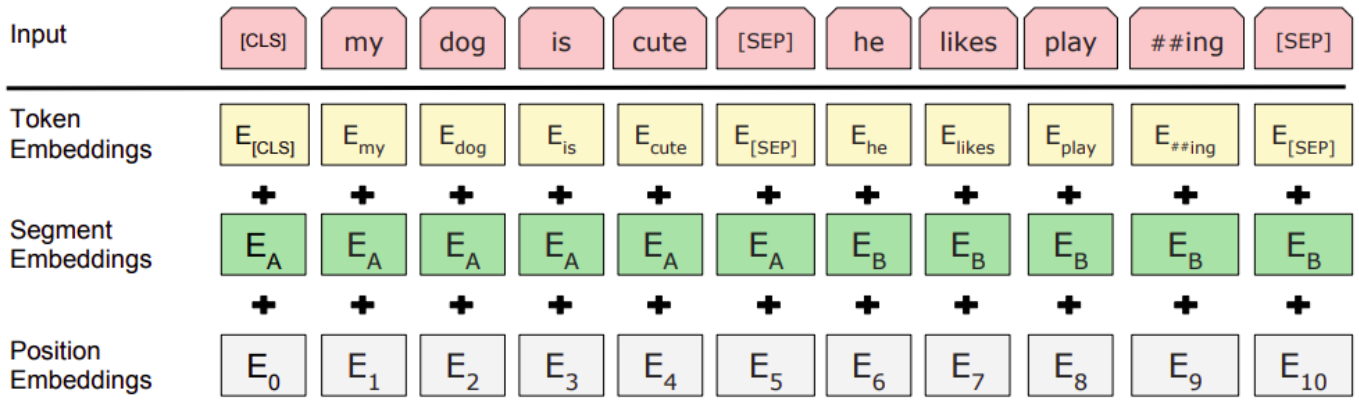


Figure: Overall pre-training and fine-tuning procedures for BERT. Fig taken from [2]

- **Text representation:** BERT operates over sequences of discrete tokens comprised of vocabulary words and a small set of special tokens: SEP, CLS, and MASK
- For a given token, the input representation is a **sum** of a **token-specific learned embedding, encodings embeddings for position** (i.e. token's index in the sequence), and **segment embeddings** (i.e. index of the token's sentence if multiple exist)

BERT Pretraining

- Trained end-end on two tasks
- **Masked language modelling:**
  - Randomly divides input tokens into disjoint sets masked X_M and observed X_O tokens (approximately 15% of tokens being masked).
  - Masked tokens are replaced with a special **[MASK]** token 80% of the time, a random word 10%, and unaltered 10%
  - Training objective: **reconstruct these masked tokens given the observed set** (cross-entropy)
  - The final hidden vectors corresponding to the mask tokens are fed into an output softmax over the vocabulary
- **Next sentence prediction:**
  - model takes tuple of sentence segments A and B following the format {CLS, w_{A1}, . . . , w_{AT} , SEP, w_{B1}, . . . , w_{BT} , SEP} and is trained to predict **whether or not B follows A** in the source text
  - CLS token representation is fed to an output layer which is trained to minimize a binary cross-entropy loss on this label

ViLBERT: Extending BERT to Jointly Represent Images and Text

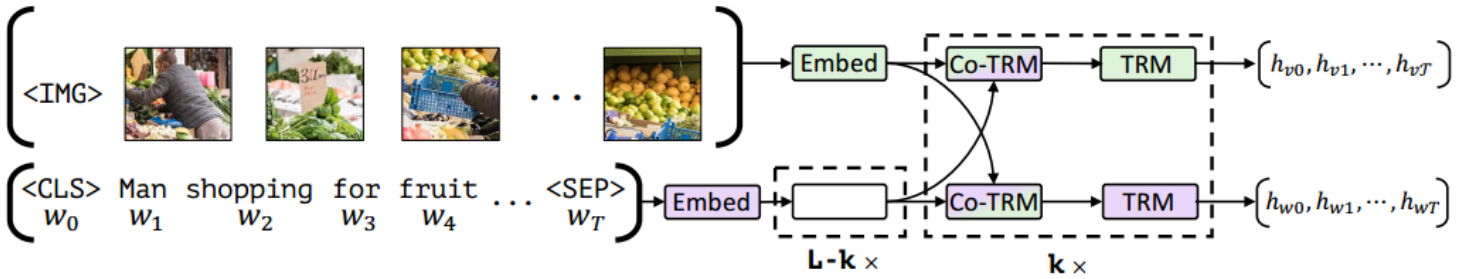
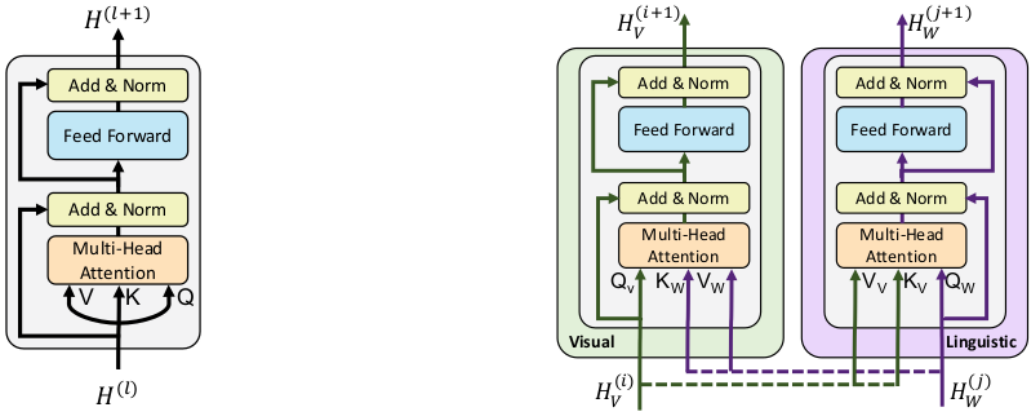


Figure: ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. Fig taken from [1]

- ViLBERT: Multi-modal co-attentional transformer module
- It consists of two parallel BERT-style models operating over image regions and text segments
- modelling each modality separately and then fusing them through a small set of attention-based interactions
- provides **interaction between modalities** at varying representation depths

- Each stream is a series of transformer blocks (TRM) and novel co-attentional transformer layers (Co-TRM) which deals in **information exchange between modalities**
- Given an image I represented as a set of region features $v_1, \ldots, v_T$ and a text input $w_0, \ldots, w_T$
- ViLBERT model outputs final representations $h_{v0}, \ldots, h_{vT}$ and $h_{w0}, \ldots, h_{wT}$



(a) Standard encoder transformer block

(b) Our co-attention transformer layer

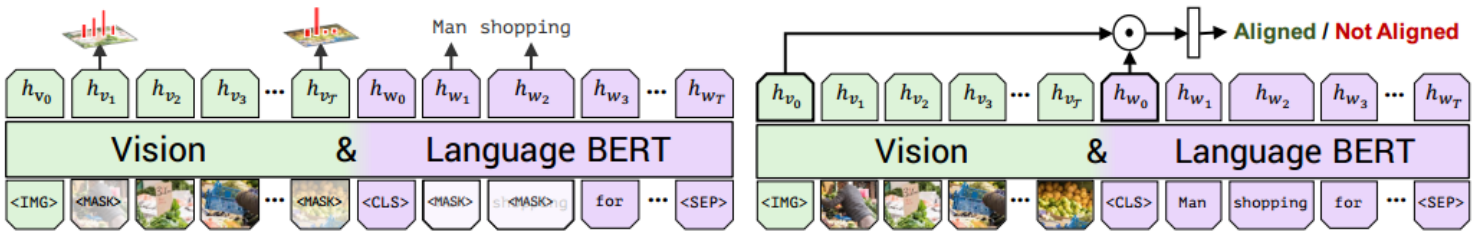
Figure: Novel co-attention mechanism on transformer architecture. Fig taken from [1]

Co-Attentional Transformer Layers

- Given intermediate visual and linguistic representations $H_V^{(i)}$ and $H_W^{(j)}$, the module computes query, key, and value matrices as in a standard transformer block
- the keys and values from each modality are passed as input to the other modality's multi-headed attention block
- produces attention-pooled features for each modality conditioned on the other
- **Effect:** image-conditioned language attention in the visual stream and language-conditioned image attention in the linguistic stream

Image Representation

- spatial location of images is encoded by construction of 5-d vector from region position (normalized top-left and bottom-right coordinates) and the fraction of image area covered
- the visual feature and spatial features of image are summed.
- append a special **IMG** token to the image region sequence representing the entire image (i.e. mean-pooled visual features with a spatial encoding corresponding to the entire image)



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

Figure: Overall Pretraining of ViLBERT. Fig taken from [1]

Training & Objectives

- **Masked multi-modal modeling**
  - masking approximately 15% of both words and image region inputs
  - Masked image regions: 90% of image features zeroed out of the time and are unaltered 10%
  - **Task:** reconstruct the masked regions the remaining inputs
  - model **predicts a distribution over semantic classes** for the corresponding image region rather than directly predicting the masked feature values
  - model is trained to **minimize the KL divergence** between these two distributions

- **Multi-modal alignment task**
  - image-text pair as input {IMG, $v_1, \ldots, v_T$, CLS, $w_1, \ldots, w_T$, SEP}
  - predict whether the image and text are aligned, i.e. whether the text describes the image
  - the outputs h_{IMG} and h_{CLS} represents the visual and linguistic inputs
  - **Overall representation:** $h_{IMG} \odot h_{CLS}$ (element-wise product)
  - learns a linear layer to make the binary prediction whether the image and text are aligned

Dataset

- Conceptual Captions: collection of 3.3 million aligned image-caption pairs automatically scraped from alt-text enabled web images.
- presents huge diversity of visual content
- To generate negatives image-caption pair, images and captions were randomly replaced

Training ViLBERT

- ViLBERT trained on **Conceptual Captions** dataset on 3.1 million image-caption pairs
- **linguistic stream** was initialized with a BERT language model
- **BERTBASE model**: 12 layers of transformer blocks with each block having a hidden state size of 762 and 12 attention heads
- **Faster R-CNN** (with ResNet-101) used to extract region features.
- **visual stream**: hidden state size of 1024 and 8 attention heads of Transformer and co-attentional transformer blocks
- **Hyperparameters**: learning rate 1e-4, batch size of 512, epochs:10, adam optimizer

Alt-text: A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

Conceptual Captions: a worker helps to clear the debris.

Alt-text: Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

Conceptual Captions: pop artist performs at the festival in a city.

Figure: Examples of images and image descriptions from the Conceptual Captions dataset

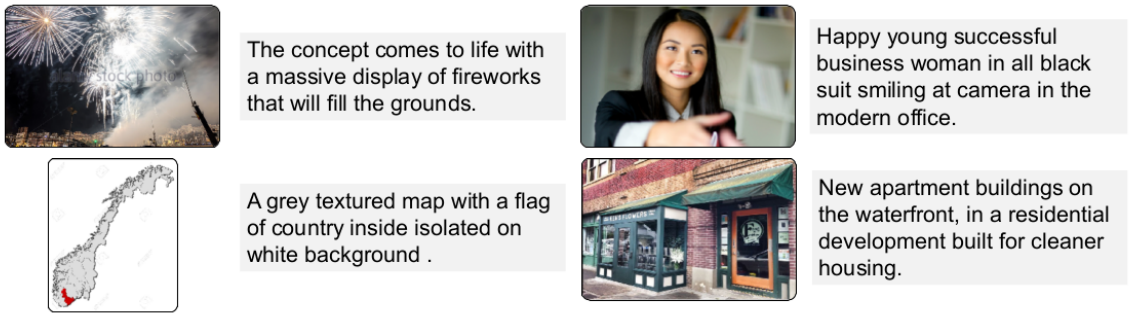


Figure: Qualitative examples of sampled image descriptions from ViLBERT model after pretraining tasks, but before fine-tuning. Fig taken from [1]

Vision-and-Language Transfer Tasks

- **fine-tuning strategy**: modify the pretrained base model by adding a classification layer on top to perform the new task and then train the entire model end-to-end

Visual Question Answering (VQA)

VQA

Figure: VQA example

- Train and evaluation on the VQA 2.0 dataset consisting of 1.1 million questions about COCO images each with 10 answers
- To fine-tune for VQA, learn a **two layer MLP on top** of the element-wise product of the h_{IMG} and h_{CLS}

Visual Commonsense Reasoning: visual question answering (Q→A) + answer justification (QA→R)

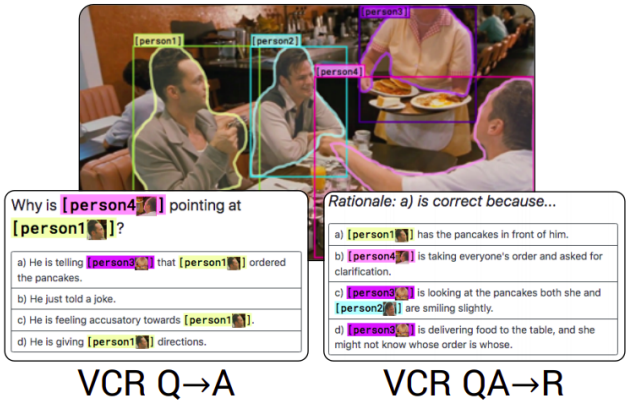


Figure: VCR example

- Visual Commonsense Reasoning (VCR) dataset consists of 290k multiple choice QA problems derived from 110k movie scenes
- To finetune on this task, **concatenate** the **question** and each possible **response** to form **four different text inputs** and pass each through ViLBERT along with the image
- prediction: learn a softmax layer on top of a linear layer

Grounding Referring Expressions

Figure: Referring expressions example

- **Task**: To localize an image region given a natural language reference
- Training and evaluation on the RefCOCO+ dataset
- For fine-tuning, the final representation h_{v_i} for each image region i is passed into a learned linear layer to predict a matching score

Caption-Based Image Retrieval

Figure: Image Retrieval example

- Training and evaluation on the Flickr30k dataset
- Train in a 4-way multiple-choice setting: randomly sampling three distractors for each image-caption pair: random caption, a random image, or a hard negative from among the 100 nearest neighbors of the target image
- softmax layer over the alignment score

'Zero-shot' Caption-Based Image Retrieval

- directly apply the pretrained multi-modal alignment prediction mechanism to caption-based image retrieval in Flickr30k **without finetuning**
- Goal: To demonstrate that the pretraining has developed the ability to ground text and that this can generalize to visual and linguistic variation without any task specific fine-tuning

Baseline:

- **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
  - The model is initialized with BERTBASE and trained identically to ViLBERT full model
  - Why this baseline? To establish the impact of the two-stream architecture of ViLBERT
- **ViLBERT$^+$**
  - ViLBERT architecture without pretraining on Conceptual Captions
  - Why this baseline? To isolate gains over task-specific baseline models as opposed to pretraining process on Conceptual Captions

	Method	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval		
		test-dev (test-std)	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
SOTA	DFAF [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-	-
SOTA	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-	-
SOTA	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-	-
SOTA	SCAN [35]	-	-	-	-	-	-	-	48.60	77.70	85.20	-	-	-
Ours	Single-Stream†	65.90	68.15	68.89	47.27	65.64	72.02	56.04	-	-	-	-	-	-
Ours	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-	-
Ours	ViLBERT†	68.93	69.26	71.01	49.48	68.61	75.97	58.44	45.50	76.78	85.02	0.00	0.00	0.00
Ours	ViLBERT	**70.55 (70.92)**	**72.42 (73.3)**	**74.47 (74.6)**	**54.04 (54.8)**	**72.34**	**78.52**	**62.61**	**58.20**	**84.90**	**91.52**	**31.86**	**61.12**	**72.80**

Figure: Transfer task results for the ViLBERT model compared with existing state-of-the-art and sensible architectural ablations. Fig taken from [1]

Takeaways:

- ViLBERT improves performance over a single-stream model
- pretraining tasks result in improved visiolinguistic representations (between 2% and 13% improvement)
- Finetuning from ViLBERT is a powerful strategy for vision-and-language tasks

Effect of Visual Stream Depth

	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval [26]		
Method	test-dev	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
ViLBERT (2-layer)	69.92	72.44	**74.80**	**54.40**	71.74	**78.61**	62.28	55.68	84.26	90.56	26.14	56.04	68.80
ViLBERT (4-layer)	70.22	**72.45**	74.00	53.82	72.07	78.53	**63.14**	55.38	84.10	90.62	26.28	54.34	66.08
ViLBERT (6-layer)	**70.55**	72.42	74.47	54.04	**72.34**	78.52	62.61	58.20	84.90	**91.52**	31.86	61.12	72.80
ViLBERT (8-layer)	70.47	72.33	74.15	53.79	71.66	78.29	62.43	**58.78**	**85.60**	91.42	**32.80**	**63.38**	**74.62**

Figure: Ablation study of the depth of our model with respect to the number of Co-TRM →TRM blocks. Fig taken from [1]

- VQA and Image Retrieval tasks benefit from greater depth - performance increases monotonically until a layer depth of 6
- Likewise, zero-shot image retrieval continues making significant gains as depth increases

Impact of Large Training set

	VQA [3]	VCR [25]			RefCOCO+ [32]			Image Retrieval [26]			ZS Image Retrieval [26]		
Method	test-dev	Q→A	QA→R	Q→AR	val	testA	testB	R1	R5	R10	R1	R5	R10
ViLBERT (0 %)	68.93	69.26	71.01	49.48	68.61	75.97	58.44	45.50	76.78	85.02	0.00	0.00	0.00
ViLBERT (25 %)	69.82	71.61	73.00	52.66	69.90	76.83	60.99	53.08	80.80	88.52	20.40	48.54	62.06
ViLBERT (50 %)	70.30	71.88	73.60	53.03	71.16	77.35	61.57	54.84	83.62	90.10	26.76	56.26	68.80
ViLBERT (100 %)	**70.55**	**72.42**	**74.47**	**54.04**	**72.34**	**78.52**	**62.61**	**58.20**	**84.90**	**91.52**	**31.86**	**61.12**	**72.80**

Figure: Transfer task results for ViLBERT as a function of the percentage of the Conceptual Captions dataset used during pre-training. Fig taken from [1]

- random subsets of 25% and 50% taken from the conceptual caption dataset, ViLBERT was pretrained and finetune ViLBERT using the same setup as before
- accuracy grows monotonically as the amount of data increases, which suggests that ViLBERT may benefit from even more pretraining data

Conclusion

- Intelligent way to extend BERT model into a joint model for image content and text and pretrain it on a large, automatically-collected dataset to learn visual grounding
- ViLBERT model introduces a novel **two-stream architecture** with co-attentional transformer blocks that exceeds state-of-the-art models for multiple vision-and-language tasks
- easy way to fine tuning by adding a classfier for each task
- strong baselines

Criticism

- No significance tests were performed
- No plots of training loss, test loss and other metrics across epochs as just optimizing the accuracy often leads to overfitting
- Not much explanation as why performance on VCR, and Ground refering tasks do not increase with the stream depth of the model
- It would be nice if the divergence measures between two modalitites were reported

Reference


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


Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805v2, 2019*.

Thank you very much for your attention.

Do you have any question?