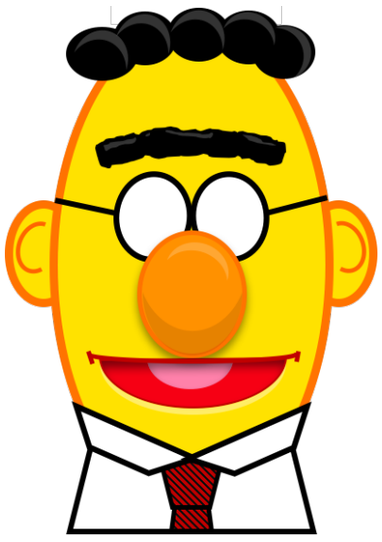
# **ViLBERT**: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks[1]

(Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee)
Presented by - **Sidharth Singla**, 20774908



**Semi-supervised Sequence Learning**

**context2Vec**

**Pre-trained seq2seq**

**ELMo**

**ULMFiT**

Multi-lingual

**MultiFiT**

Transformer

**BERT**

Bidirectional LM

**GPT**

Larger model
More data

**GPT-2**

Defense

**Grover**

Cross-lingual

**XLM**

**UDify**

Multi-task

**MT-DNN**

Knowledge distillation

**MT-DNN$_{KD}$**

+ Generation

**MASS**

**UniLM**

Span prediction
Remove NSP

**SpanBERT**

Longer time
Remove NSP
More data

**RoBERTa**

Permutation LM
Transformer-XL
More data

**XLNet**

+Knowledge Graph

**ERNIE (Tsinghua)**

Neural entity linker

**KnowBert**

Cross-modal

Whole Word Masking

**VideoBERT**
**CBT**
**ViLBERT**
**VisualBERT**
**B2T2**
**Unicoder-VL**
**LXMERT**
**VL-BERT**
**UNITER**

**ERNIE (Baidu)**

**BERT-wwm**

By Xiaozhi Wang & Zhengyan Zhang @THUNLP

# INTRODUCTION

- Vision-and-Language BERT.
- A model for learning task-agnostic joint representations of image content and natural language.
- Two-stream model. Visual and textual processing in separate streams that interact through co-attentional transformer layers.
- 'Vision and Language' tasks.
- Pretrain-then-transfer learning approach to vision-and-language tasks.
- Github - https://github.com/jiasenlu/vilbert_beta

# MODEL

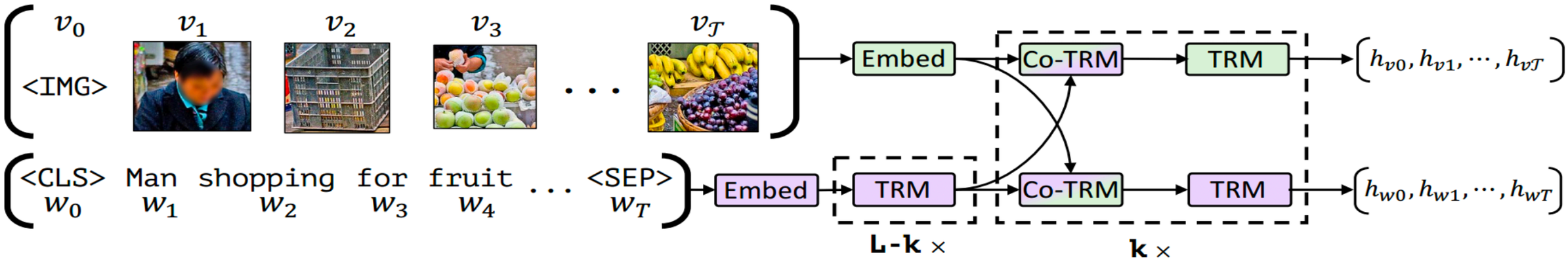
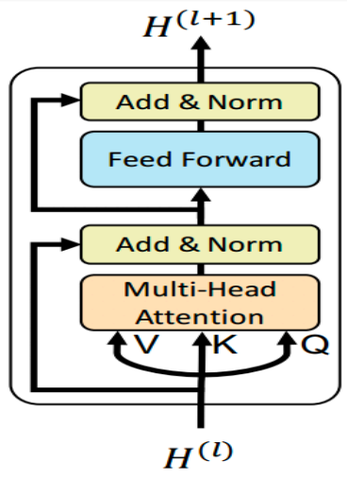
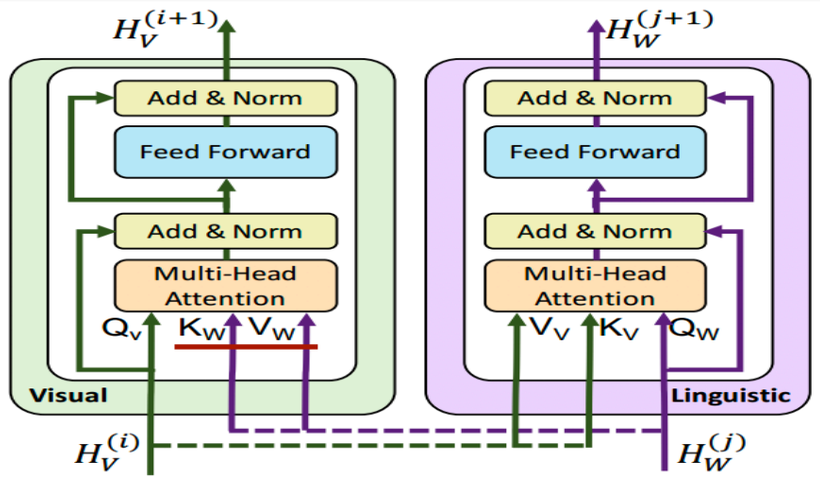



Figure 1: Our ViLBERT model consists of two parallel streams for visual (green) and linguistic (purple) processing that interact through novel co-attentional transformer layers. This structure allows for variable depths for each modality and enables sparse interaction through co-attention. Dashed boxes with multiplier subscripts denote repeated blocks of layers.




(a) Standard encoder transformer block

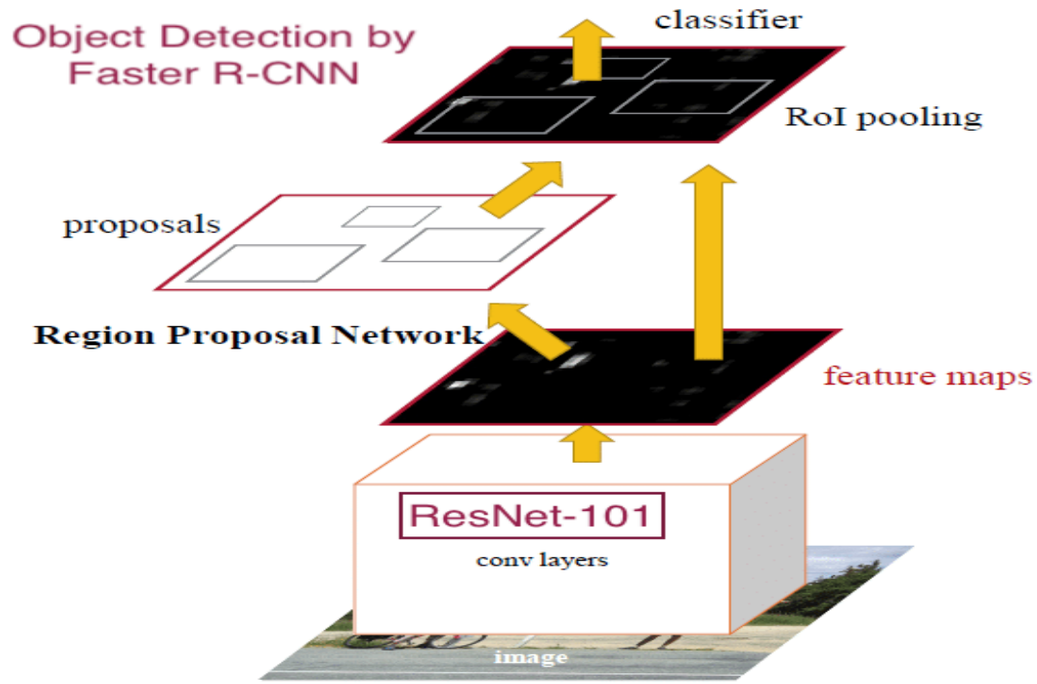



(b) Our co-attention transformer layer

Figure 2: We introduce a novel co-attention mechanism based on the transformer architecture. By exchanging key-value pairs in multi-headed attention, this structure enables vision-attended language features to be incorporated into visual representations (and vice versa).

# IMAGE REPRESENTATIONS

- Image region features are generated by extracting bounding boxes and their visual features. For region i, $v_i$ is the mean-pooled convolutional feature from that region.
- Spatial locations(Location Embeddings) encoded as a 5-d vector: Region position (normalized top-left and bottom-right coordinates) and the fraction of image area covered.
- embeddings = image_embeddings + token_type_embeddings + loc_embeddings
- Image region sequence begins with an IMG token. Represents the entire image.




Pooling--Average (mean) pooling

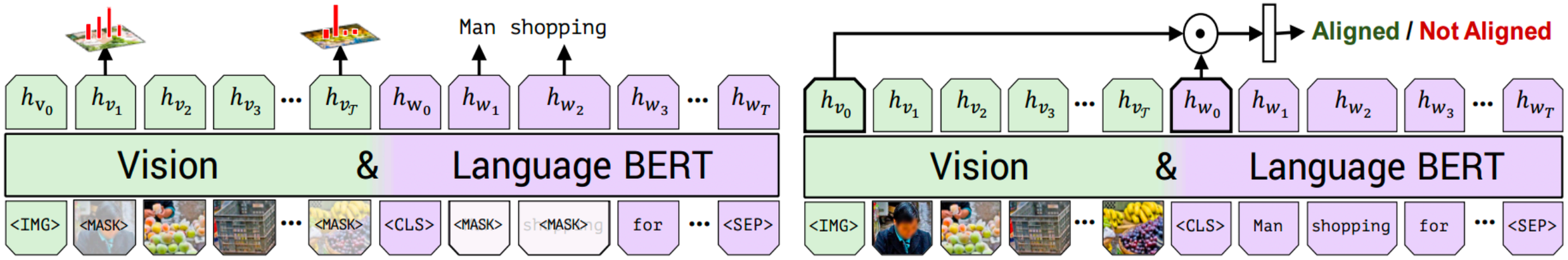
| | | | |
|---|---|---|---|
| 12 | 7 | 0 | 86 |
| 19 | 8 | 0 | 12 |
| 27 | 5 | 23 | 4 |
| 97 | 12 | 35 | 60 |

→

| | |
|---|---|
| 11.5 | 24.5 |
| 35.25 | 30.5 |

$$a = \frac{\sum_{N}^{N} a_i}{N}$$

# PRE-TRAINING




(a) Masked multi-modal learning

(b) Multi-modal alignment prediction

Figure 3: We train ViLBERT on the Conceptual Captions [24] dataset 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.

- **Dataset**
  - Conceptual Captions: Collection of 3.3 million image-caption pairs automatically scraped from alt-text enabled web images[2].

- **Masked Multi-Modal Learning Task**
  - Approximately 15% of both words and image region are masked and reconstructed given the remaining inputs.
  - Image features zeroed out 90% and unaltered 10%. Masked text inputs are handled as in BERT[3].
  - Model predicts a distribution over semantic classes rather than directly regressing the masked feature values for the corresponding image region.
    - Supervision by output distribution for the region from the pretrained detection model used. Minimize KL divergence.
- **Multi-modal alignment task**
  - Prediction whether the text describes the image(image aligned with the text).
  - Element-wise product between $h_{IMG}$ and $h_{CLS}$ and a linear layer is learnt to make the binary prediction.

# IMPLEMENTATION

BERTBASE - 12 layers of transformer blocks. Each block having hidden state size of 762 and 12 attention heads.

- Linguistic stream initialized with a BERTBASE language model pre-trained on the BookCorpus and English Wikipedia.
- BASE model chosen due to concerns over training time. BERTLARGE model can further boost performance.
- Faster R-CNN[4](with ResNet-101 backbone) pretrained on Visual Genome dataset is used to extract region features. 10 to 36 high-scoring image region boxes are selected.
- Transformer and co-attentional transformer blocks in the visual stream have hidden state size of 1024 and 8 attention heads.
- Trained on 8 TitanX GPUs with a total batch size of 512 for 10 epochs.
- Adam optimizer with initial learning rates of 1e-4 is used with a linear decay learning rate schedule.
- Both training task losses are weighed equally.

# TRANSFER TASKS

- Pretrained ViLBERT model transferred to a set of four established vision-and-language tasks and one diagnostic task.

- Fine-tuning strategy to modify the pretrained base model and perform the new task by training the entire model end-to-end.

- **Visual Question Answering (VQA)**

  - Training and Evaluation on VQA 2.0 dataset.

  - Fine-tuning: Two layer MLP is learnt on top of the element-wise product of the image and text representations $h_{IMG}$ and $h_{CLS}$.

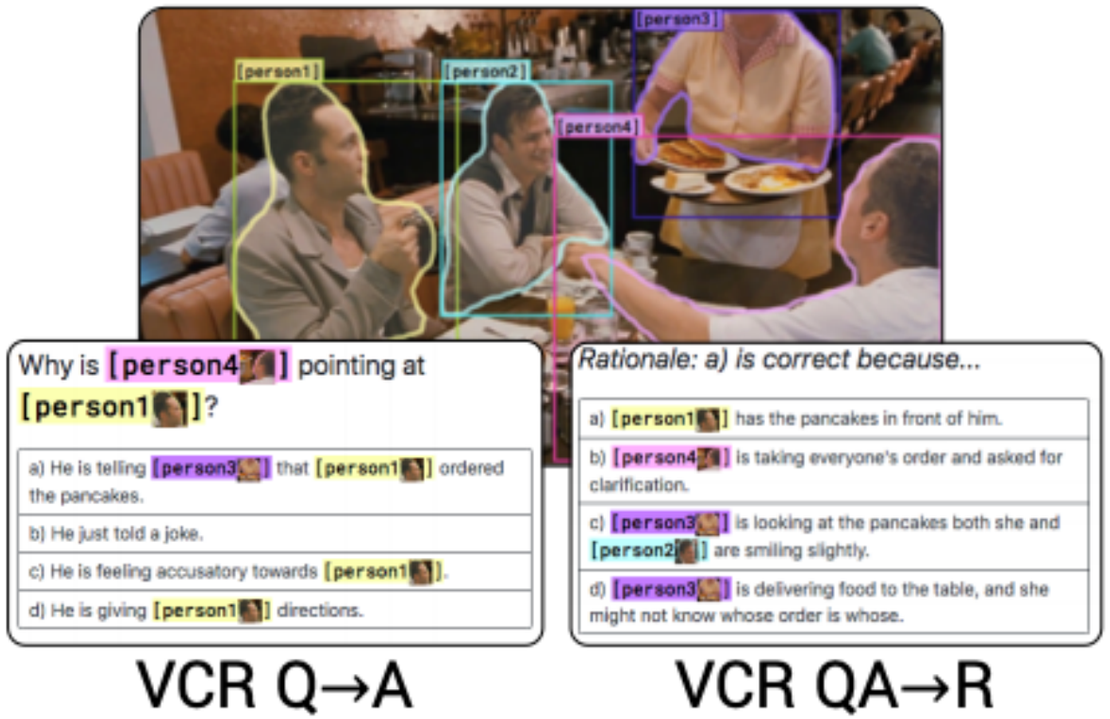
  - Multi-label classification task. Loss - Binary cross-entropy loss. Batch size 256. Maximum 20 epochs. Initial learning rate 4e-5.


VQA

- **Visual Commonsense Reasoning (VCR)**

  - Given an image, Visual Question Answering (Q->A) and Answer justification (QA->R).

  - Trained on Visual Commonsense Reasoning (VCR) dataset having object tags integrated into the language providing direct grounding supervision and explicitly excludes referring expressions.




  - Fine-tuning: Question and each possible response is concatenated and four different text inputs are passed along with the image. A linear layer is learnt on top of the post-element-wise product representation.

  - Softmax prediction. Loss - Cross-entropy loss. 20 epochs. Batch size 64. Initial learning rate 2e-5.

- **Grounding Referring Expressions**

  - Localize an image region given a natural language reference.

  - Training and Evaluation is done on RefCOCO+ dataset.

  - Bounding box proposals provided by *MAttNet*[5], which use a Mask R-CNN are directly used.

  - Fine-tuning: Final representation $h_{vi}$ is passed into a learned linear layer to predict a matching score. IoU is computed with the ground truth box thresholding at 0.5.

  - Loss - Binary cross-entropy loss. Maximum 20 epochs. Batch size 256. Initial learning rate 4e-5.

- **Caption-Based Image Retrieval**
  - Identifying an image from a pool given a caption describing its content.
  - Training and Evaluation is done on the Flickr30k dataset. Trained in a 4-way multiple-choice setting by randomly sampling three distractors for each image-caption pair - substituting a random caption, a random image, or a hard negative from among the 100 nearest neighbors of the target image.
  - Alignment score(same as in alignment prediction pretraining) is computed for each. Softmax applied. Loss - Cross-entropy loss. 20 epochs. Batch size 64. Initial learning rate 2e-5.
- **'Zero-shot' Caption-Based Image Retrieval**
  - Pre-trained multi-modal alignment prediction model on Conceptual Captions dataset is used directly. No fine-tuning.
  - Demonstrates that the pretraining has developed the ability to ground text. Tested on the caption-based image retrieval task test-set.

# BASELINES

- **Single-Stream Model**
  - Single BERT architecture processing both modality inputs through same set of transformer blocks - **sharing parameters** and processing stacks for both visual and linguistic inputs.
  - No changes to the BERT architecture, resulting in significantly deeper visual processing and earlier interaction between the modalities than ViLBERT. Trained identically.
- **ViLBERT†**
  - ViLBERT architecture that has **not undergone** any pre-training tasks.
  - BERT initialization for the linguistic stream and represents image regions.
  - Baseline is compared to isolate gains over task-specific baseline models that might be due to the architecture, language initialization, or visual features as opposed to the pre-training process.

# RESULTS

Table 1: Transfer task results for our ViLBERT model compared with existing state-of-the-art and sensible architectural ablations. † indicates models without pretraining on Conceptual Captions. For VCR and VQA which have private test sets, we report test results (in parentheses) only for our full model. Our full ViLBERT model outperforms task-specific state-of-the-art models across all tasks.

| | | VQA [3] | VCR [25] | | | RefCOCO+ [32] | | | Image Retrieval [26] | | | ZS Image Retrieval | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Method | 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** |

Table 2: Ablation study of the depth of our model with respect to the number of Co-TRM→TRM blocks (shown in a dashed box in Fig. 1). We find that different tasks perform better at different network depths – implying they may need more or less context aggregation.

| | 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** |

Table 3: Transfer task results for ViLBERT as a function of the percentage of the Conceptual Captions dataset used during pre-training. We see monotonic gains as the pretraining dataset size grows.

| | 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** |

# OTHER MULTIMODAL MODELS (BERT BASED/RELATED )

# VL-BERT[6]

| Model | Q → A | | QA → R | | Q → AR | |
|---|---|---|---|---|---|---|
| | val | test | val | test | val | test |
| R2C (Zellers et al., 2019) | 63.8 | 65.1 | 67.2 | 67.3 | 43.1 | 44.0 |
| ViLBERT (Lu et al., 2019)† | 72.4 | 73.3 | 74.5 | 74.6 | 54.0 | 54.8 |
| VisualBERT (Li et al., 2019b)† | 70.8 | 71.6 | 73.2 | 73.2 | 52.2 | 52.4 |
| B2T2 (Alberti et al., 2019)† | 71.9 | 72.6 | 76.0 | 75.7 | 54.9 | 55.0 |
| VL-BERT$_{BASE}$ w/o pre-training | 73.1 | - | 73.8 | - | 54.2 | - |
| VL-BERT$_{BASE}$ | 73.8 | - | 74.4 | - | 55.2 | - |
| VL-BERT$_{LARGE}$ | 75.5 | 75.8 | 77.9 | 78.4 | 58.9 | 59.7 |

Table 1: Comparison to the state-of-the-art methods with single model on the VCR dataset. † indicates concurrent works.

| Model | test-dev | test-std |
|---|---|---|
| BUTD (Anderson et al., 2018) | 65.32 | 65.67 |
| ViLBERT (Lu et al., 2019)† | 70.55 | 70.92 |
| VisualBERT (Li et al., 2019b)† | 70.80 | 71.00 |
| LXMERT (Tan & Bansal, 2019)† | 72.42 | 72.54 |
| VL-BERT$_{BASE}$ w/o pre-training | 69.58 | - |
| VL-BERT$_{BASE}$ | 71.16 | - |
| VL-BERT$_{LARGE}$ | 71.79 | 72.22 |

Table 2: Comparison to the state-of-the-art methods with single model on the VQA dataset. † indicates concurrent works.

#### 4.2.3 Referring Expression Comprehension

| Model | Ground-truth Regions | | | Detected Regions | | |
|---|---|---|---|---|---|---|
| | val | testA | testB | val | testA | testB |
| MAttNet (Yu et al., 2018) | 71.01 | 75.13 | 66.17 | 65.33 | 71.62 | 56.02 |
| ViLBERT (Lu et al., 2019)† | - | - | - | 72.34 | 78.52 | 62.61 |
| VL-BERT$_{BASE}$ w/o pre-training | 74.41 | 77.28 | 67.52 | 66.03 | 71.87 | 56.13 |
| VL-BERT$_{BASE}$ | 79.88 | 82.40 | 75.01 | 71.60 | 77.72 | 60.99 |
| VL-BERT$_{LARGE}$ | 80.31 | 83.62 | 75.45 | 72.59 | 78.57 | 62.30 |

Table 3: Comparison to the state-of-the-art methods with single model on the RefCOCO+ dataset. † indicates concurrent work.

# UNITER[7]

| Tasks | | SOTA | ViLBERT | VLBERT | Unicoder -VL | VisualBERT | LXMERT | UNITER BASE | UNITER LARGE |
|---|---|---|---|---|---|---|---|---|---|
| VQA | test-dev | 70.63 | 70.55 | 70.50 | - | 70.80 | 72.42 | 72.27 | **73.24** |
| | test-std | 70.90 | 70.92 | 70.83 | - | 71.00 | 72.54 | 72.46 | **73.40** |
| VCR | Q→A | 72.60 | 73.30 | 74.00 | - | 71.60 | - | 75.00 | **77.30** |
| | QA→R | 75.70 | 74.60 | 74.80 | - | 73.20 | - | 77.20 | **80.80** |
| | Q→AR | 55.00 | 54.80 | 55.50 | - | 52.40 | - | 58.20 | **62.80** |
| NLVR$^2$ | dev | 54.80 | - | - | - | 67.40 | 74.90 | 77.14 | **78.40** |
| | test-P | 53.50 | - | - | - | 67.00 | 74.50 | 77.87 | **79.50** |
| SNLI-VE | val | 71.56 | - | - | - | - | - | 78.56 | **79.28** |
| | test | 71.16 | - | - | - | - | - | 78.02 | **78.98** |
| ZS IR (Flickr) | R@1 | - | 31.86 | - | 42.40 | - | - | 62.34 | **65.82** |
| | R@5 | - | 61.12 | - | 71.80 | - | - | 85.62 | **88.88** |
| | R@10 | - | 72.80 | - | 81.50 | - | - | 91.48 | **93.52** |
| IR (Flickr) | R@1 | 48.60 | 58.20 | - | 68.30 | - | - | 71.50 | **73.66** |
| | R@5 | 77.70 | 84.90 | - | 90.30 | - | - | 91.16 | **93.06** |
| | R@10 | 85.20 | 91.52 | - | 94.60 | - | - | 95.20 | **95.98** |
| IR (COCO) | R@1 | 38.60 | - | - | 44.50 | - | - | 48.42 | **51.72** |
| | R@5 | 69.30 | - | - | 74.40 | - | - | 76.68 | **78.41** |
| | R@10 | 80.40 | - | - | 84.00 | - | - | 85.90 | **86.93** |
| ZS TR (Flickr) | R@1 | - | - | - | 61.60 | - | - | 75.10 | **77.50** |
| | R@5 | - | - | - | 84.80 | - | - | 93.70 | **96.30** |
| | R@10 | - | - | - | 90.10 | - | - | 95.50 | **98.50** |
| TR (Flickr) | R@1 | 67.90 | - | - | 82.30 | - | - | 84.70 | **88.20** |
| | R@5 | 90.30 | - | - | 95.10 | - | - | 97.10 | **98.40** |
| | R@10 | 95.80 | - | - | 97.80 | - | - | 99.00 | **99.00** |
| TR (COCO) | R@1 | 50.40 | - | - | 59.60 | - | - | 63.28 | **66.60** |
| | R@5 | 82.20 | - | - | 85.10 | - | - | 87.04 | **89.42** |
| | R@10 | 90.00 | - | - | 91.80 | - | - | 93.08 | **94.26** |
| Ref-COCO | val | 87.51 | | - | - | - | - | 91.64 | **91.84** |
| | testA | 89.02 | - | - | - | - | - | 92.26 | **92.65** |
| | testB | 87.05 | - | - | - | - | - | 90.46 | **91.19** |
| | val$^d$ | 77.48 | - | - | - | - | - | 81.24 | **81.41** |
| | testA$^d$ | 83.37 | - | - | - | - | - | 86.48 | **87.04** |
| | testB$^d$ | 70.32 | - | - | - | - | - | 73.94 | **74.17** |
| Ref-COCO+ | val | 75.38 | - | 78.44 | - | - | - | 82.84 | **84.04** |
| | testA | 80.04 | - | 81.30 | - | - | - | 85.70 | **85.87** |
| | testB | 69.30 | - | 71.18 | - | - | - | 78.11 | **78.89** |
| | val$^d$ | 68.19 | 72.34 | 71.84 | - | - | - | 74.72 | **74.94** |
| | testA$^d$ | 75.97 | 78.52 | 77.59 | - | - | - | 80.65 | **81.37** |
| | testB$^d$ | 57.52 | 62.61 | 60.57 | - | - | - | 65.15 | **65.35** |
| Ref-COCOg | val | 81.76 | - | - | - | - | - | 86.52 | **87.85** |
| | test | 81.75 | - | - | - | - | - | 86.52 | **87.73** |
| | val$^d$ | 68.22 | - | - | - | - | - | 74.31 | **74.86** |
| | test$^d$ | 69.46 | - | - | - | - | - | 74.51 | **75.77** |

Table 4: Results on downstream V+L tasks from UNITER model, compared with task-specific state-of-the-art (SOTA) and concurrent pre-trained models. ZS: Zero-Shot, IR: Image Retrieval and TR: Text Retrieval.

# REFERENCES


1) Lu J, Batra D, Parikh D, Lee S. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. InAdvances in Neural Information Processing Systems 2019 (pp. 13-23).

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

3) Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

4) Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In NuerIPS, pages 91–99, 2015.

5) Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In CVPR, 2018.

6) Su W, Zhu X, Cao Y, Li B, Lu L, Wei F, Dai J. Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530. 2019 Aug 22.

7) Chen YC, Li L, Yu L, Kholy AE, Ahmed F, Gan Z, Cheng Y, Liu J. Uniter: Learning universal image-text representations. arXiv preprint arXiv:1909.11740. 2019 Sep 25.