

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

Jiasen Lu, Dhruv Batra, Devi Parikh, Stefan Lee

Paper Published in NeurIPS 2020

Presented by:

Arka Mukherjee


Balaji Balasubramanian

November 9th, 2021

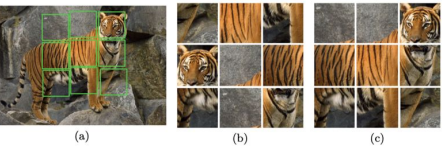
Proxy Tasks!

- Supervised learning is expensive.
- Proxy-tasks are implicit tasks generated from the data itself.

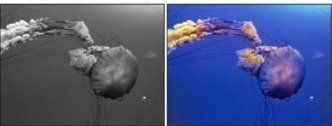
Example:



Predicting relative location



Solving a jigsaw puzzle



Colorizing an image

Visual Grounding

- A lot of progress has been made in training models specific to images such as Resnet, Vision Transformers and training models specific to text data such as BERT.
- Text based models lack the real world understanding of the texts and Image based model lack the understanding of language that humans use to communicate with each other.
- Using transformers to learn representations on visual and language data to perform Visual Grounding is a promising direction as the model learn to build a good joint representation for two important modalities.

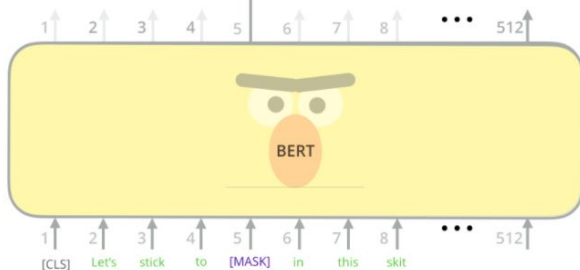
Revisiting BERT

Use the output of the masked word's position to predict the masked word

Possible classes:
All English words

0.1%	Aardvark
...	...
10%	Improvisation
...	...
0%	Zyzyva

FFNN + Softmax



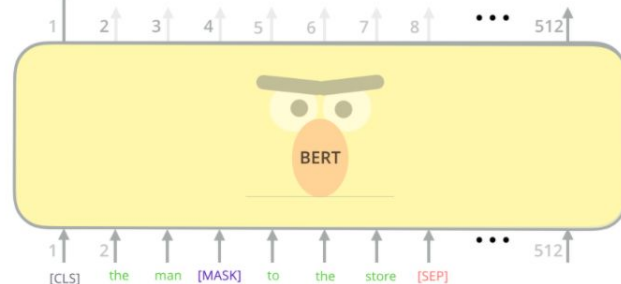
Randomly mask
15% of tokens

Input

BERT's clever language modeling task masks 15% of words in the input and asks the model to predict the missing word.

Predict likelihood
that sentence B
belongs after
sentence A

FFNN + Softmax



Tokenized
Input

Input

[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flightless birds [SEP]

Sentence A Sentence B

<https://jalammar.github.io/illustrated-transformer/>

Discretize Visual Space

Discretize the visual space using clustering to create visual tokens and pass it to pre-trained BERT model.

This approach has few drawbacks:

- Loss of information during the discretization process.
- Treating text and visual modalities identically and ignoring the inherent differences in both modalities.
- Damaging the language understanding of BERT.

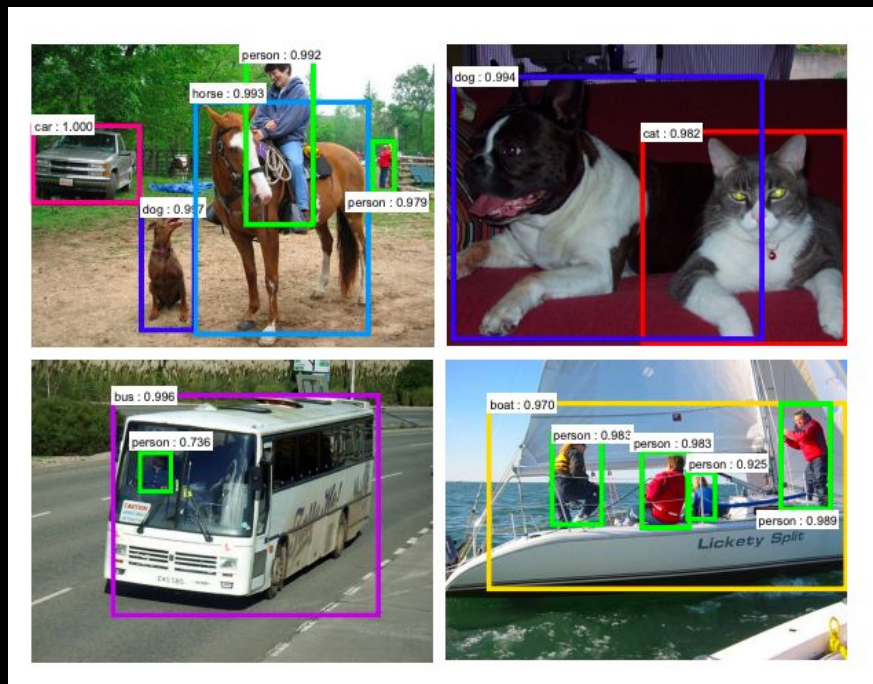
Instead, the authors create a two stream architecture that treats visual and language modalities separately and fuse them later.

- the authors say: "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", could you elaborate?

Creating visual tokens of objects in the images.

Image Region Features

- Image region features are generated from the bounding box predictions of Faster R-CNN model with Resnet 101 backbone.
- 5d output is created consisting of the bounding box coordinates and fraction of image area covered.
- The object detection labels are used for the self-supervised learning task later.



Self Attention

Query vectors: $Q = XW_Q$

Key vectors: $K = XW_K$

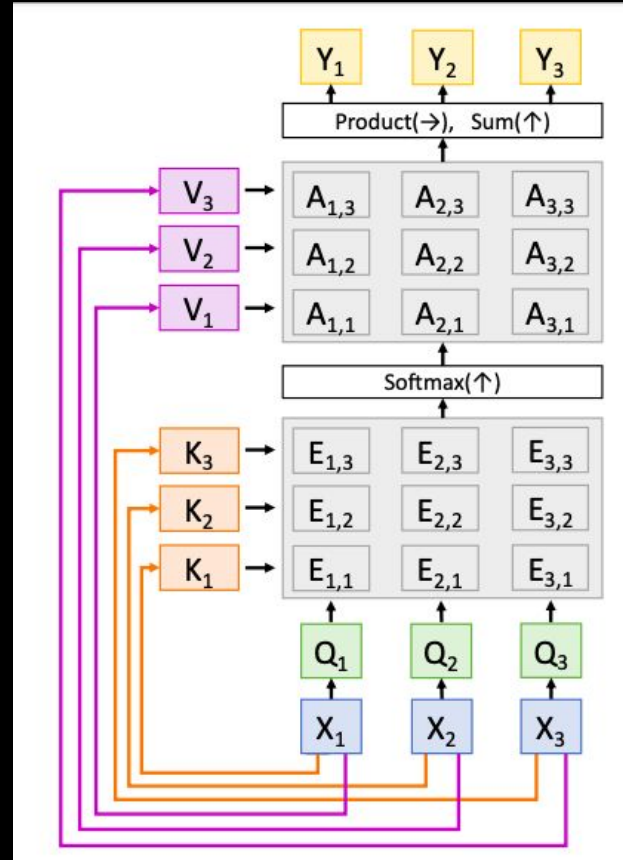
Value Vectors: $V = XW_V$

Similarities: $E = QK^T$

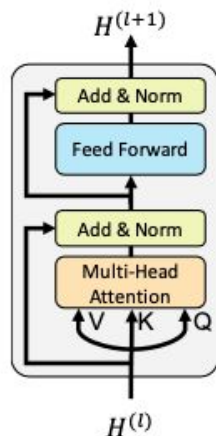
Attention weights: $A = \text{softmax}(E, \text{dim}=1)$

Output vectors: $Y = AV$

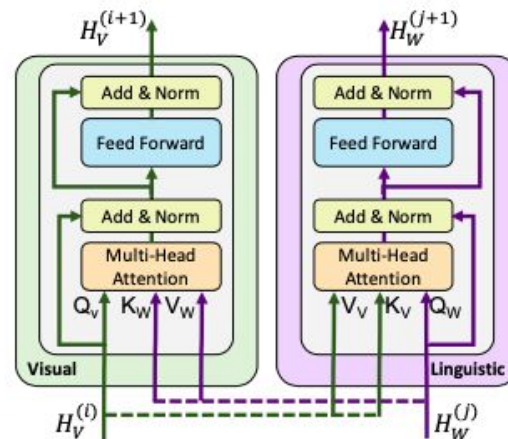
https://www.youtube.com/watch?v=YAgjfMR9R_M



Co-attention transformer block



(a) Standard encoder transformer block



(b) Our co-attention transformer layer

- Co-attention mechanism is performed by exchanging key-value pairs in visual and linguistic streams.
- It results in image-conditioned language attention in the visual stream and language-conditioned image attention in the linguistic stream

Dataset - Conceptual Captions



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.

Contains 3.3M Image-Text pairs

Dataset - Conceptual Captions

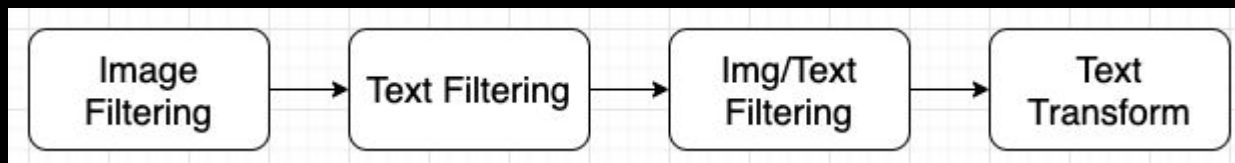


Image Filtering: JPEG Images, both dimensions greater than 400 pixels, ratio of dimension less than 2:1

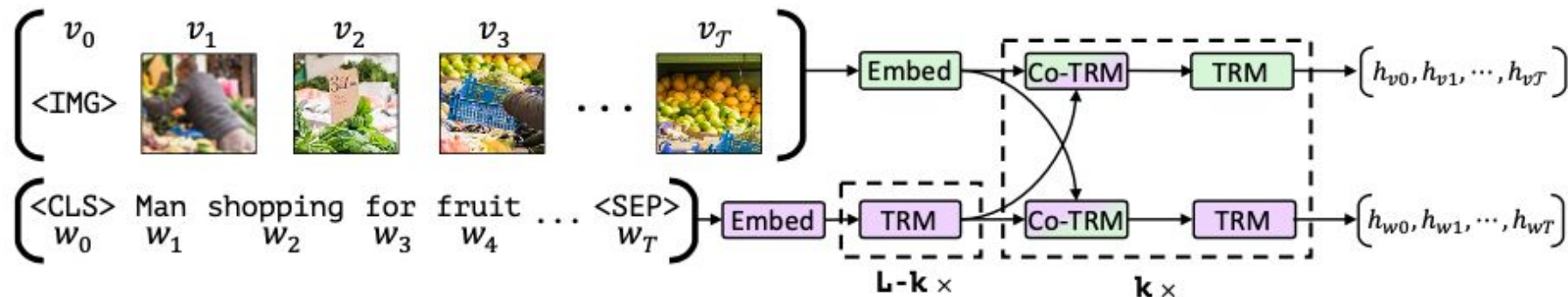
Text Filtering : Using Google Cloud Natural Language APIs to filter the text captions. 97% text captions removed. Profanity, token repetition, capitalization of first word, etc detected.

Image and Text based filtering- Mapping the text tokens to images using Google Cloud Vision APIs.

Text Transform- Google Cloud Natural Language API (named entity recognition) and Google Knowledge Graph (KG) Search API (Harrison Ford to actor) used.

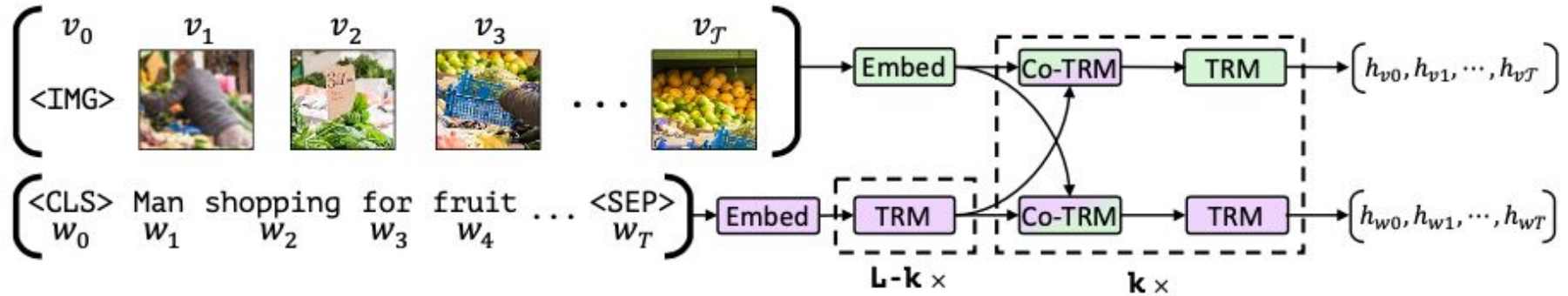
Original Alt-text	Harrison Ford and Calista Flockhart attend the premiere of 'Hollywood Homicide' at the 29th American Film Festival September 5, 2003 in Deauville, France.
Conceptual Captions	actors attend the premiere at festival.
what-happened	"Harrison Ford and Calista Flockhart" mapped to "actors"; name, location, and date dropped.

ViLBERT



- Two parallel BERT-style model operate over visual(green) and language(purple) stream.
- The image is represented with image region features $v_0 \dots v_T$ and text is represented with tokens $w_0 \dots w_T$.
- Language features have larger amount of preprocessing than Visual features.

ViLBERT



- Each stream contains of vanilla transformer blocks (TRM) and co-attentional transformer layers (Co-TRM)
- The blocks inside the dotted boxes are being repeated.

Questions



Nehal Pandey 4:45 PM

I was wondering as the models are only tested on language and vision datasets, so these generic representations be significant in case of unimodal tasks?

- Text based models lack the real world understanding of the texts. The trophy did not fit in the suitcase because it was big. What does 'it' refer to here.
- Image based model lack the understanding of language that humans use to communicate with each other and is rich with information. An image can speak a thousand words!



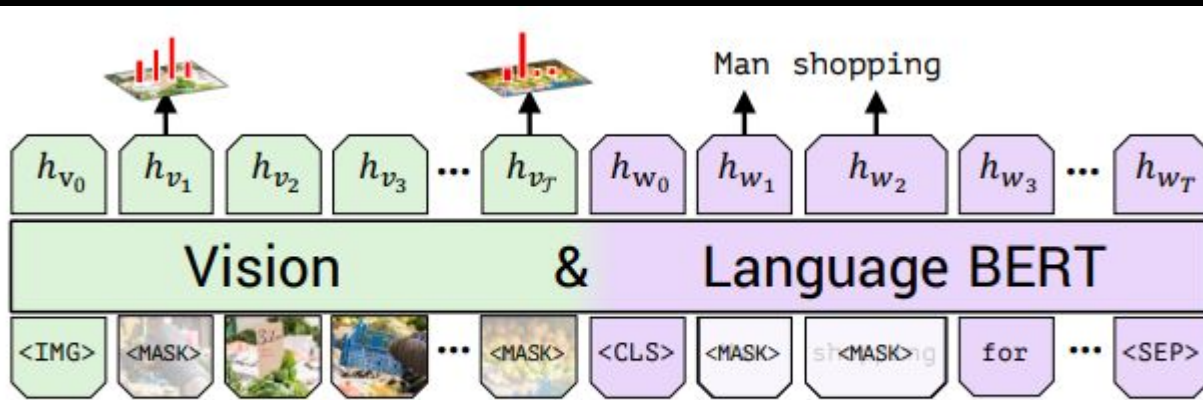
Abhinav Moudgil 1:23 PM

Why do you think two-stream architecture in ViLBERT works better than single-stream?

- Two-stream architecture can accommodate different processing needs of each modality.
- Text data was given more preprocessing than image data because the image features were already high level.
- Hard to finetune Bert to image and text data.

Pretraining tasks: Masked Multi-modal Modelling

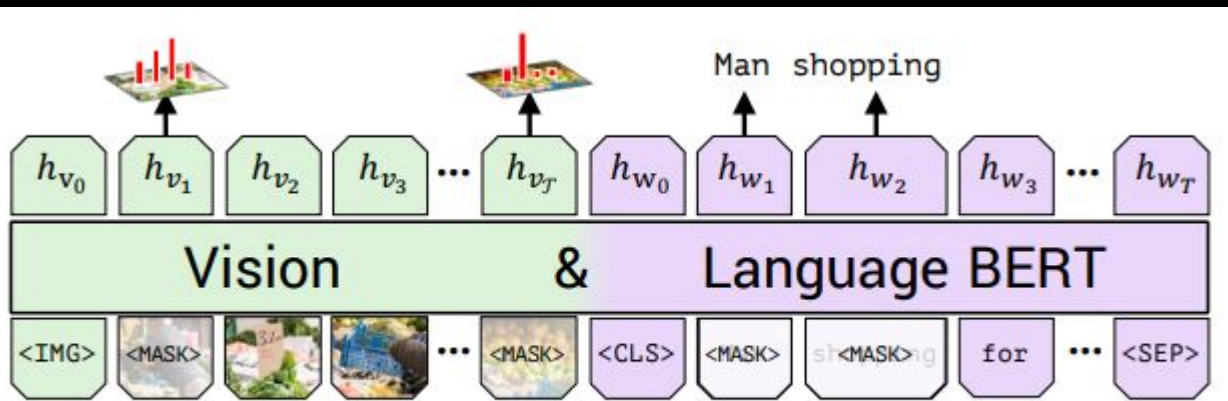
- Mask 15% of image and text regions. The model needs to reconstruct those based on the remaining inputs.
- **Text:** Similar to BERT, for the 15% masked tokens, 80% are zeroed out, 10% are unaltered and 10% are replaced by a random text.



(a) Masked multi-modal learning

Pretraining tasks: Masked Multi-modal Modelling

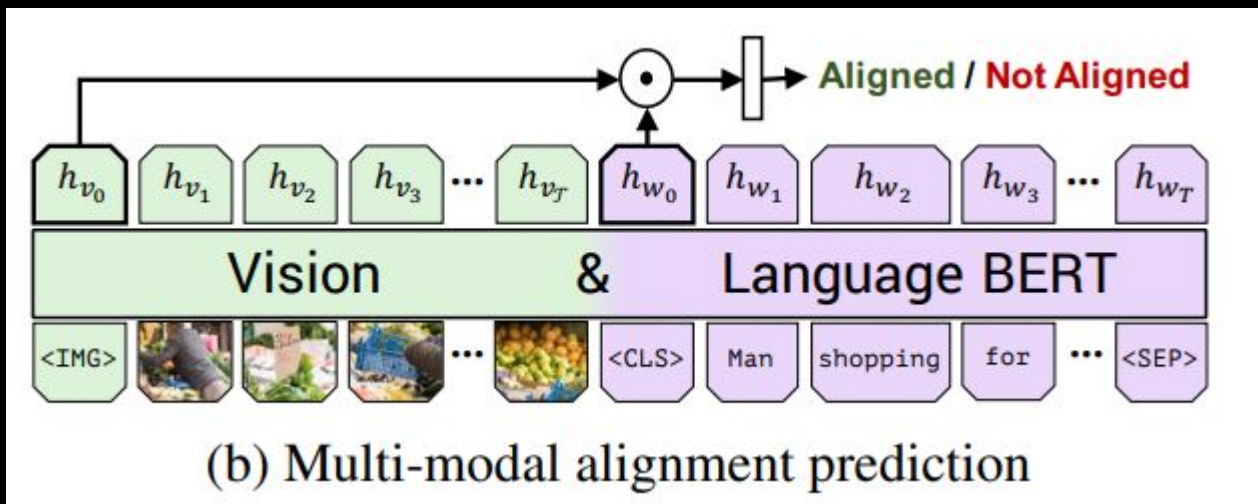
- Images: The masked tokens are zeroed out 90% of the times and unaltered 10% of the times because masks do not appear during the finetuning stage.
- Predict the image class i.e. reduce the KL divergence loss between the Faster R-CNN prediction distribution and the prediction of the ViLBERT.



(a) Masked multi-modal learning

Multi Modal Alignment Prediction

- This task involves a binary prediction where the model must predict whether the caption describes the image content or not.



Experiments

Vision:

- Faster R-CNN with ResNet 101 Backbone (pretrained on visual genome dataset).
- ADAM optimizer for convergence; LR of $1e-4$.
- Batch Size = 512.
- Number of Epochs = 10.
- Caption Alignment Loss & Masked Loss are weighted equally.
- 8 Attention Heads.
- Hidden state size = 1024.

Text:

- BERT_{base} chosen due to training time concerns. (BookCorpus)
- BERT_{large} likely to be better.
- Hidden state size = 762.
- 12 Attention Heads.

VQA - Visual Question & Answers

- VQA 2.0 dataset used. Over 200k images, 3+ questions per image,
- Learn a 2 Layer MLP on top of the vision and text representations mapping to 3129 answers.
- The answers could be yes/no, counting, single word, open ended.
- LR $4e-5$, batch size 256, 20 epochs.



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



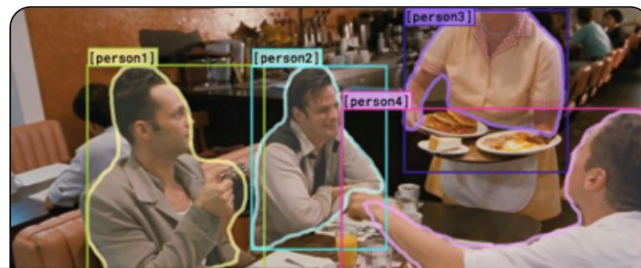
Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

VCR - Visual Commonsense Reasoning (VCR)

- Visual Commonsense Reasoning (290k MCQ Questions, 110k movie scenes)- It contains MCQ VQA and Answer Justification.
- The image and four texts (separately) are passed as input to the model. A linear layer will predict a score for each of the 4 image-text pair and softmax would convert it into probability. LR- $2e-5$, 20 epochs, batch size 64.



Why is [person4] pointing at [person1]?

- a) He is telling [person3] that [person1] ordered the pancakes.
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

VCR Q→A

Rationale: a) is correct because...

- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

VCR QA→R

Grounding Referring Expressions

- Localize an image region given a natural language reference.
- Dataset - RefCoco+
- Bounding box proposals of Mask RCNN are fed to visual stream, the text is passed to the language stream. Matching score is predicted for each region. LR 4e-5, batch size 256, 20 epochs.

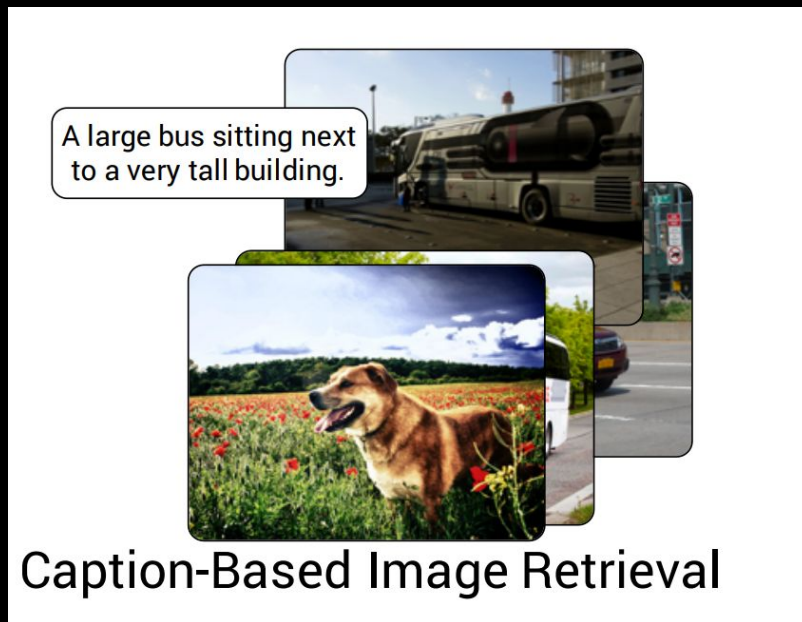


Guy in yellow dribbling ball

Referring Expressions

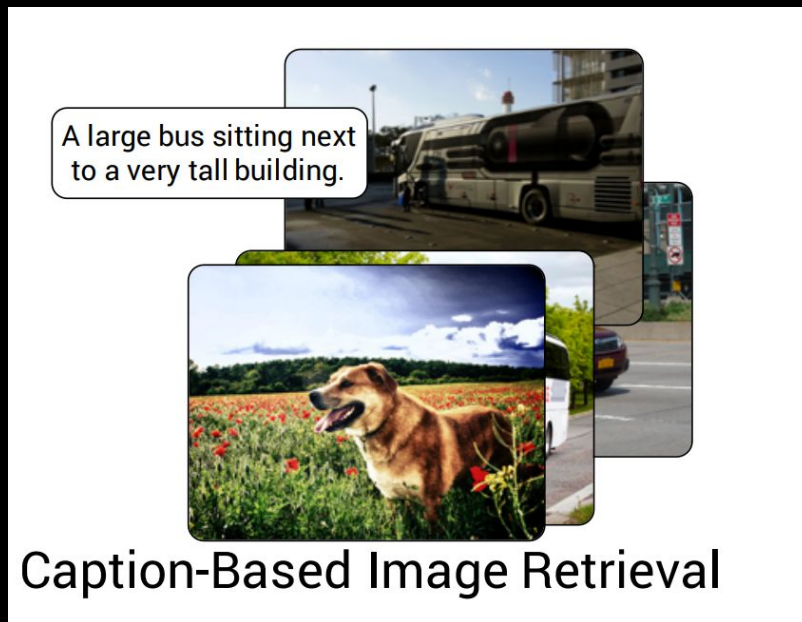
Caption Based Image Retrieval

- Identifying an image from a pool (4 way MCQ) given a caption describing its content.
- Trained on Flickr30k dataset (31000 images, 5 captions for each image)
- Compute alignment score and apply softmax activation to it.
- LR- $2e-5$, 20 epochs, batch size 64.



Zero-Shot Caption Based Image Retrieval

- Directly using the model pretrained on multi-modal alignment prediction without finetuning on Flickr30k dataset.
- The alignment prediction score is passed to softmax activation.



Questions



Venkatesh 3:34 PM

This has been asked before but can you comment on their decision to use an object detector for extracted regions?

- The object detection labels are used to pretrain the model.
- Most questions in problems like VQA are directly related to the objects in context.
 - Objects like sky, grass are missed.
 - Possible Solution: Instance Segmentation.

Who is wearing glasses?

man



woman



Where is the child sitting?

fridge



arms



Is the umbrella upside down?

yes



no



How many children are in the bed?

2



1

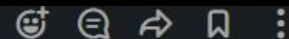


turn over

Questions



Ali Ghelmani 8:38 AM



The authors state that the image regions are extracted using the Faster R-CNN object detector. Object detectors are usually trained on a limited number of object classes. Considering that their training dataset is the result of web crawling, I wonder how the model handles objects, and their associated captions, which the object detector does not recognize, and how valid would the suggested image regions be in such a case. What are your thoughts on this?

- We think it's a problem with Transfer Learning in general (i.e, the ability to generalize to new data).
- Which is why there are additional pre-training tasks on datasets like COCO / VQA 2.0 / VCR.
- Faster R-CNN is trained on COCO + ImageNet.
- Open point: SSL tests using the JFT dataset.

Test baselines for analysis

1. Singular Stream BERT architecture: A vanilla BERT model accepts inputs of both modalities through the same set of transformer blocks, and it has missing co-TRM. This is computationally expensive(*) as we can't cache across a single streams as both visio + linguistic streams don't interact with one another.

* hence, not used for image retrieval

Test baselines for analysis

2. ViLBERT: A regular ViLBERT that has not undergone additional pre-training on Conceptual Captions. This helps us analyze performance gains that occur due to pre-training.



Task Specific Baselines / Results

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)	-	-	-	-	-	-	-	-	-	-	-
	R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-
	MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-
	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	-	-	-	-	-
	Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-
	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
	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

As expected, **pre-trained ViLBERT > vanilla ViLBERT!**

What does pre-training learn?



The concept comes to life with a massive display of fireworks that will fill the grounds.



A grey textured map with a flag of country inside isolated on white background .



Happy young successful business woman in all black suit smiling at camera in the modern office.



New apartment buildings on the waterfront, in a residential development built for cleaner housing.

What does pre-training learn?

M

Matthew Riemer 11:03 PM

The sampled image descriptions from ViLBERT in Figure 5 are very interesting. It seems like the biggest issue in these cases is that the descriptions are long and include information that may be speculative or incorrect. Do you think that this is just an artifact of the pre-training dataset used as the authors suggest? Or maybe this could have to do with the sampling procedure as well?

- Issue in verbosity could be due to lack of task-specific fine-tuning.
- The Conceptual Captions dataset usually isn't famous for brevity.
- Also potentially linked to HTML based data collection / web-scraping for the dataset.

Questions



Julia Hindel 8:14 AM



The authors state that the linguistic stream is initialised with a pre-trained BERT LM. How about the visual stream? It's not clear to me if it is also initialised with pre-trained weights.

Why does it improve the efficiency if linguistic representation is frozen for the caption-based retrieval?

- The visual stream is built on a Faster R-CNN with a ResNet-101 backbone.
- It improves the computational cost, as we reduce some re-computations for the linguistics stream (Caching).
- The text stream has significantly more processing before interacting with visual features.
- The context aggregation requirement is less.

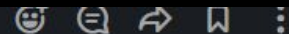
Ablation Study: Effect of Visual Stream Depth

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

- There could be overfitting going on with a higher capacity for VCR and RefCOCO+.
- Plus, some tasks like Image Retrieval need more context aggregation.
- Context Aggregation \propto ViLBERT Depth



Vishal Ghorpade 12:22 PM



- Looking at Table 2, it seems like there needs to be exploration to find layers-size that are beneficial for a specific task as the performance for different task on different layers varies i.e it is not consistent that if Vilbert-6layer is good for VQA but not for VCR. Is this correct?

Ablation Study: Percentage of Data used

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

- Data Size \propto Performance!
- **Open Point:** Still scope for increase w.r.t data collection.

Questions

Z

Zhixuan Lin 4:49 PM

I imagine if the model is trained on only extract object patches then it will focus on object descriptions and have trouble learning the global scene semantics. But in Figure 5, it seems that the model has a pretty decent understanding of the overall semantics of the image. Can you provide some intuition why the ViLBERT can learn this?

- This could be potentially due to the two-stream architecture of the ViLBERT.
- The model could learn context-information from the associated texts fed to the Linguistic stream, even if the bounding boxes miss the vision context.



Happy young successful business woman in all black suit smiling at camera in the modern office.

Questions

P

Philippe Brouillard 11:58 AM

Interestingly, the results of the single-stream method (uses directly BERT) are better than some SOTA methods. It makes me wonder if the SOTA are really adequate? Also, compared to ViLBERT, the single-stream accuracy is often around 2% lower. Is it a convincing difference in the field (I mean there is no standard deviation to truly compare the methods)?

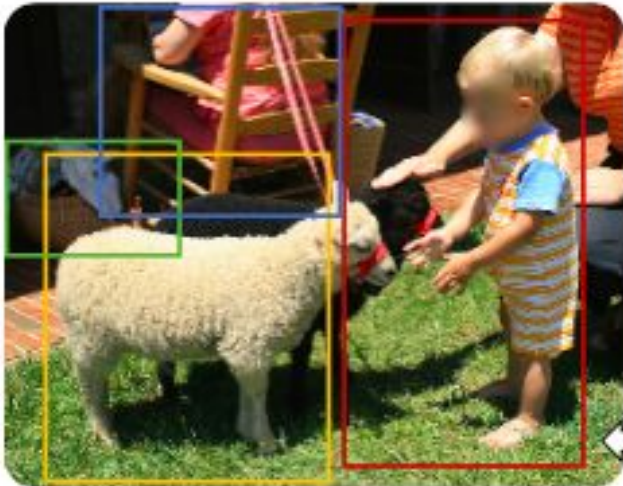
- Previous SOTA methods had lack of pre-training, fine-tuning, lack of BERT in the pipeline, etc.

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 DFAP [36]	70.22 (70.34)	-	-	-	-	-	-	-	-	-	-	-	-
R2C [25]	-	63.8 (65.1)	67.2 (67.3)	43.1 (44.0)	-	-	-	-	-	-	-	-	-
MAttNet [33]	-	-	-	-	65.33	71.62	56.02	-	-	-	-	-	-
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	-	-	-	-	-	-
Single-Stream	68.85	71.09	73.93	52.73	69.21	75.32	61.02	-	-	-	-	-	-
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
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

Future of ViLBERT!

12-in-1: Multi-Task Vision and Language Representation Learning

- Train ViLBERT on 12 datasets from 4 categories of tasks.
- Reduce the number of parameters from 3 billion to 270 million while improve the performance by 2.05 on an average.



Visual Question Answering
What color is the child's outfit? Orange

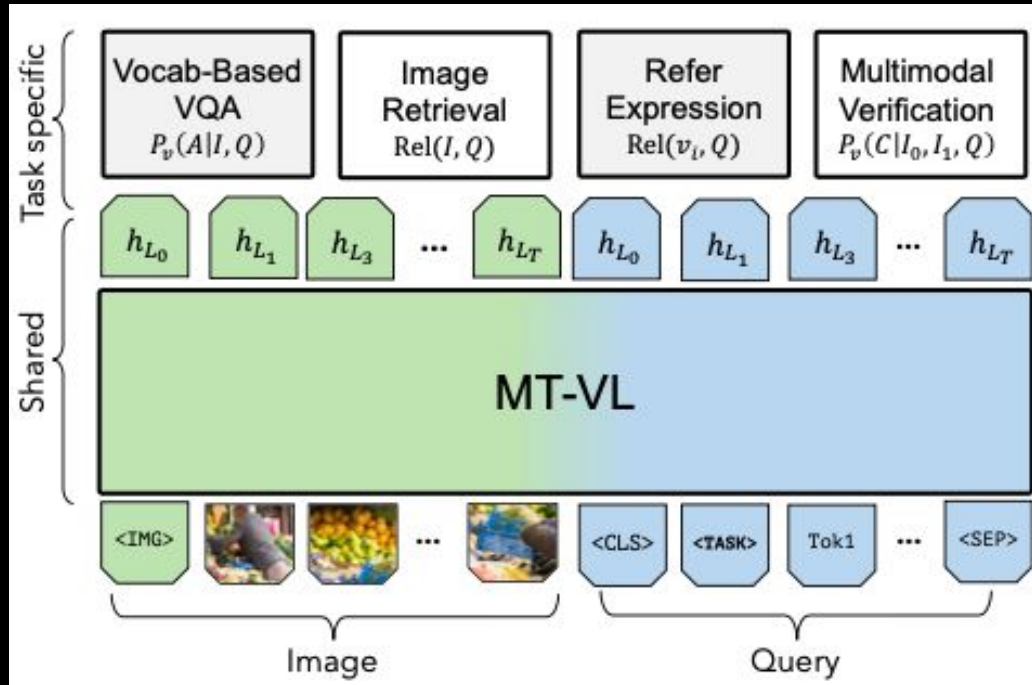
Referring Expressions
child sheep basket people sitting on chair

Multi-modal Verification
The child is petting a dog. **false**

Caption-based Image Retrieval
A child in orange clothes plays with sheep.

12-in-1- Model Architecture

- The ViLBERT model is shared between the four tasks (with a new pre-training scheme).
- Each of the four tasks have task specific layers on top of the shared model.



4 tasks

Vocab Based VQA

For an input image and text question, select the answer from a fixed vocabulary (classification)

Image Retrieval

For a set of images and a text caption, select the image most relevant to the caption.

Referring Expressions

Given a text expression and image, detect the image region referred to by the expression.

Multi-modal Verification

Given one or more images and a natural language statement, judge the correctness or predict their semantic relationship.

12-in-1: Multi-Task Vision and Language Representation Learning

Pretraining

- The model is pretrained on Conceptual Caption dataset.

Round Robin Batch Level Sampling

- For every multi task iteration, each task sends a batch of dataset and updates the parameters in sequence.

Dynamic Stop-and-Go

- There are two modes :- stop and go
- Monitor the validation loss of the task once per epoch.
- If the validation performance improvement is less than 0.1%, move the task to stop mode.
- Smaller updates are made in the stop mode.
- If the validation performance is less than 0.5%, move the task to go mode.

Curriculum Learning

- Curriculum Learning moves from easiest (fastest converging) to hardest (slowest converging) task.
- Anti Curriculum Learning moves from hardest (VQA) to easiest task (RE).
- Both performed worse than no Curriculum Learning. This is probably due to catastrophic forgetting).

12-in-1: Multi-Task Vision and Language Representation Learning

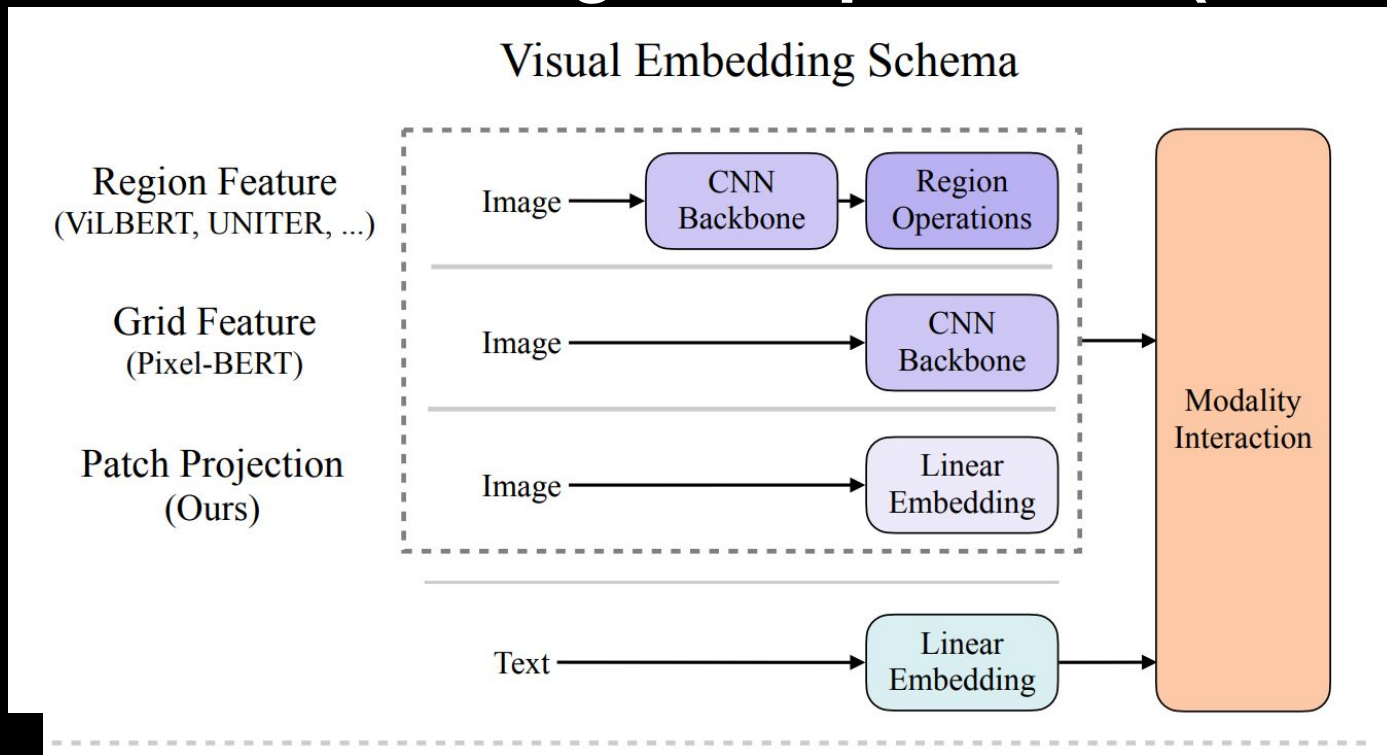
Algorithm 1: DSG for Multi-Task Learning

```
 $n_t \leftarrow$  number of iterations per epoch for task  $t$ 
 $\Delta \leftarrow$  size of gap between iterations in stop mode
 $DSG_t \leftarrow go$ 
for  $i \leftarrow 1$  to  $MaxIter$  :
    for  $t \in Tasks$  :
        if  $DSG_t = go$  or ( $DSG_t = stop$  and  $i \bmod \Delta = 0$ ) :
            Compute task loss  $L_t(\theta)$  and gradient  $\nabla_t(\theta)$ 
            Update  $\theta \leftarrow \theta - \epsilon \nabla_t(\theta)$ , where  $\theta = \theta_s \cup \theta_t$ 
        if  $i \bmod n_t = 0$  :
            Compute validation score  $s_t$  on task  $t$ 
            if  $DSG_t = go$  and  $Converged(s_t)$  :
                |  $DSG_t \leftarrow stop$ 
            else if  $DSG_t = stop$  and  $Diverged(s_t)$  :
                |  $DSG_t \leftarrow go$ 
        end
    end
end
```

12-in-1: Multi-Task Vision and Language Representation Learning - Ablation Study

	Task Token	Dynamic Stop-and-Go	G1	G2	G3	G4	All Tasks Average
AT (our)							
1 token per dataset	✓	✓	56.35	63.61	75.52	77.61	69.08
2 token per head	✓	✓	55.95	61.48	75.35	77.37	68.52
3 w/o task token		✓	55.67	62.55	75.38	76.73	68.53
4 w/o DSG	✓		55.50	62.92	75.24	76.31	68.52
5 w/ curriculum			54.68	61.21	75.19	76.70	67.24
6 w/ anti-curriculum			55.82	59.58	73.69	75.94	67.98
7 vanilla multitask			54.09	61.45	75.28	76.71	67.92

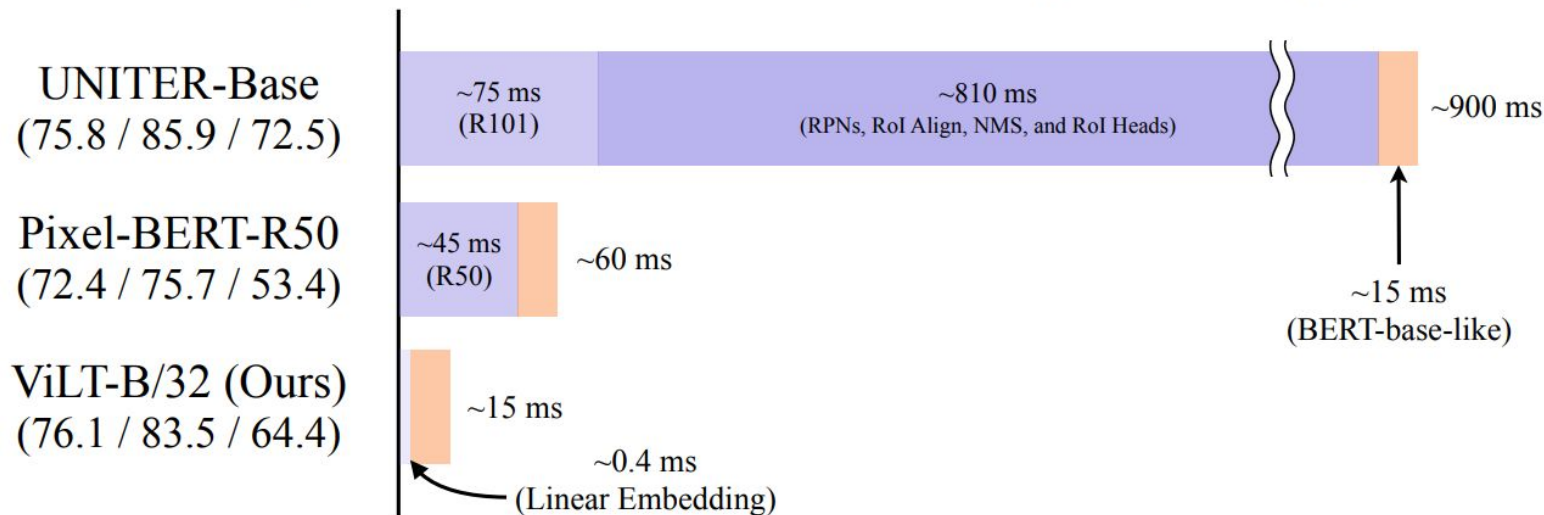
ViLT - Vision-and-Language Transformer Without Convolution or Region Supervision (Kim et. al)



ViLT - Vision-and-Language Transformer Running Time

Running Time

(Performances : NLVR2 test-P Acc. / F30K TR R@1 / F30K IR R@1)



ViLT - Vision-and-Language Transformer Running Time

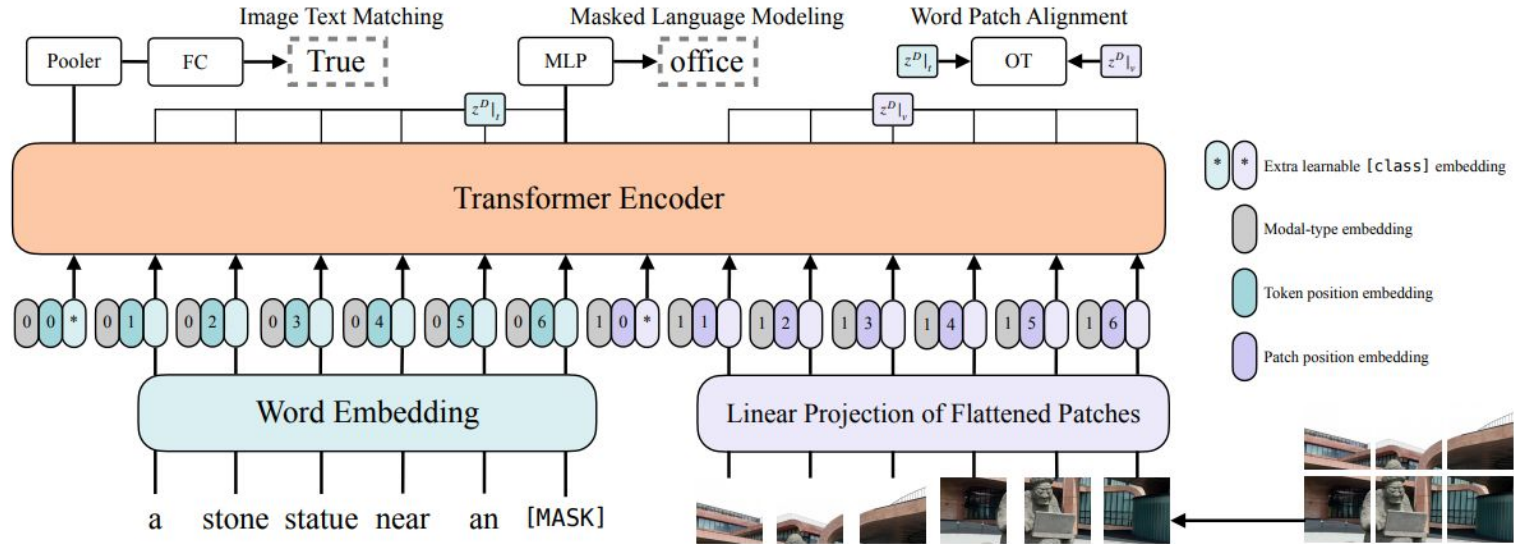


Figure 3. Model overview. Illustration inspired by Dosovitskiy et al. (2020).

Patches ~ Only around 2.4M parameters

ViLT - Vision-and-Language Transformer Running Time



a display of **flowers** growing out and over the retaining **wall** in front of **cottages** on a **cloudy** day.



flowers



wall



cottages



cloudy



a room with a **rug**, a **chair**, a **painting**, and a **plant**.



rug



chair



painting



plant

ViLT - Vision-and-Language Transformer Running Time (Kim et. al)

- **Augmentation:**
- All strategies same as RandAugment, which is also used in Noisy Student, **except cutout and colour inversion.**
- **Open Point:** Gaussian / RBF Noise might potentially help.

<https://arxiv.org/pdf/2102.03334.pdf>

ViLT - Vision-and-Language Transformer Running Time (Kim et. al)

Visual Embed	Model	#Params (M)	#FLOPs (G)	Time (ms)
Region	ViLBERT ³⁶⁺³⁶	274.3	958.1	~900
	VisualBERT ³⁶⁺¹²⁸	170.3	425.0	~925
	LXMERT ³⁶⁺²⁰	239.8	952.0	~900
	UNITER-Base ³⁶⁺⁶⁰	154.7	949.9	~900
	OSCAR-Base ⁵⁰⁺³⁵	154.7	956.4	~900
	VinVL-Base ⁵⁰⁺³⁵	157.3	1023.3	~650
	Unicoder-VL ^{100+?}	170.3	419.7	~925
	ImageBERT ¹⁰⁰⁺⁴⁴	170.3	420.6	~925
Grid	Pixel-BERT-X152 ^{146+?}	144.3	185.8	~160
	Pixel-BERT-R50 ^{260+?}	94.9	136.8	~60
Linear	ViLT-B/32 ²⁰⁰⁺⁴⁰	87.4	55.9	~15

<https://arxiv.org/pdf/2102.03334.pdf>

ViLT - Vision-and-Language Transformer

Performance

Visual Embed	Model	Time (ms)	Zero-Shot Text Retrieval						Zero-Shot Image Retrieval					
			Flickr30k (1K)			MSCOCO (5K)			Flickr30k (1K)			MSCOCO (5K)		
			R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Region	ViLBERT	~900	-	-	-	-	-	-	31.9	61.1	72.8	-	-	-
	Unicoder-VL	~925	64.3	85.8	92.3	-	-	-	48.4	76.0	85.2	-	-	-
	UNITER-Base	~900	80.7	95.7	98.0	-	-	-	66.2	88.4	92.9	-	-	-
	ImageBERT [†]	~925	70.7	90.2	94.0	44.0	71.2	80.4	54.3	79.6	87.5	32.3	59.0	70.2
Linear	ViLT-B/32	~15	69.7	91.0	96.0	53.4	80.7	88.8	51.3	79.9	87.9	37.3	67.4	79.0
	ViLT-B/32 [⊕]	~15	73.2	93.6	96.5	56.5	82.6	89.6	55.0	82.5	89.8	40.4	70.0	81.1

<https://arxiv.org/pdf/2102.03334.pdf>