Visual Question Answering
Models

Fall 2021

Overview

Introduction

Bottom-up and Top-down Attention Model

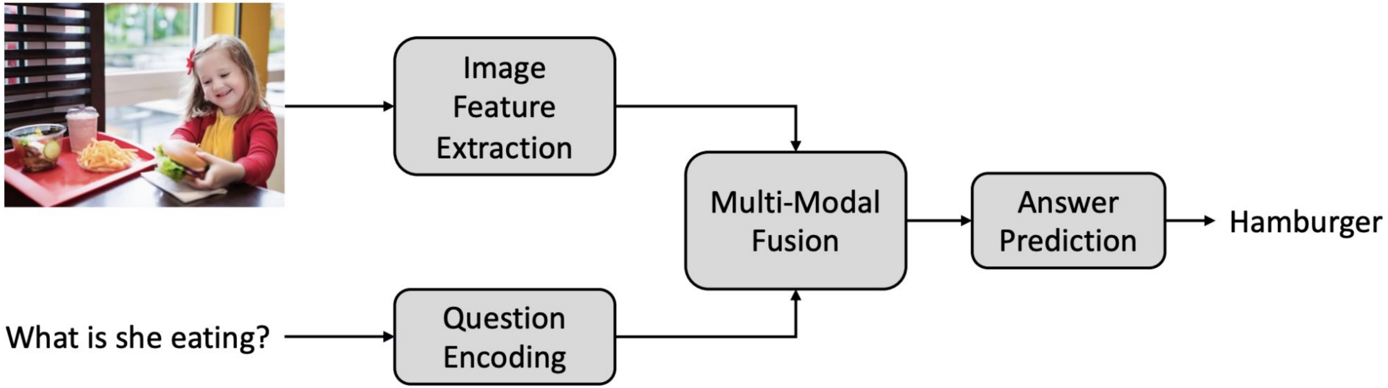
Vision Language Pre-training

Oscar Model

Grid Features vs Region Based Features

Without Convolution or Region Supervision

How does a typical VQA system work?



Slide: datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf

Overview

Introduction

Bottom-up and Top-down Attention Model

Vision Language Pre-training

Oscar Model

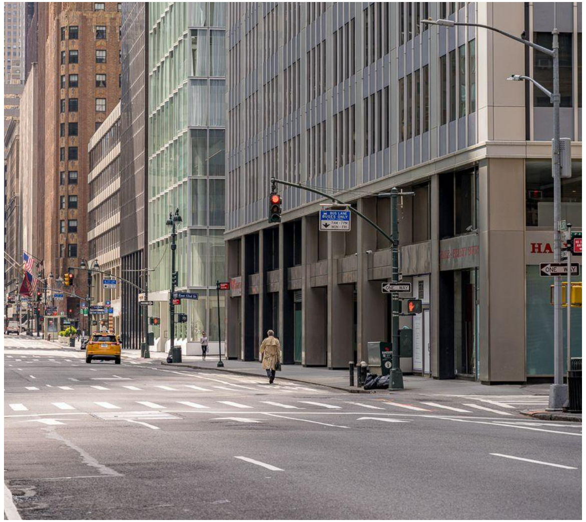
Grid Features vs Region Based Features

Without Convolution or Region Supervision

Visual Attention

Fine-grained visual processing is often essential for visual and language tasks.

Learn to focus on image regions related to the task.



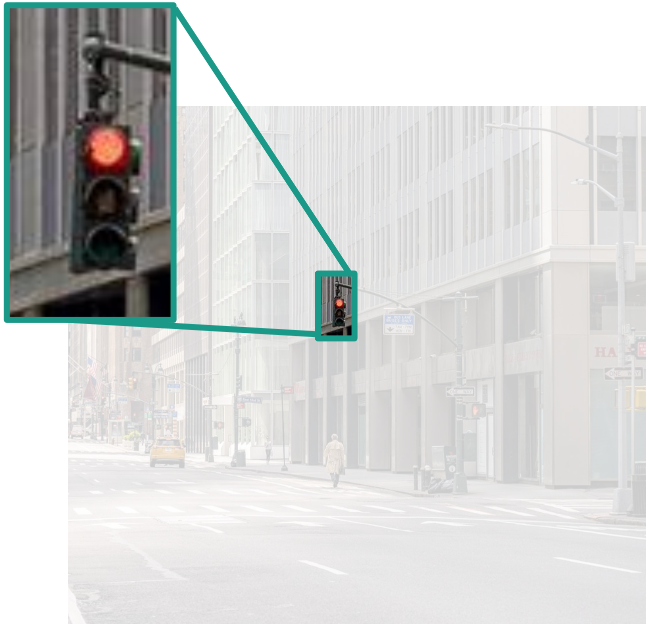
Q: What color is the traffic light?

Image: metouhey.com
Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Visual Attention

Fine-grained visual processing is often essential for visual and language tasks.

Learn to focus on image regions related to the task.



Q: What color is the traffic light?

A: Red

Image: metouhey.com
Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Visual Attention

Fine-grained visual processing is often essential for visual and language tasks.

Learn to focus on image regions related to the task.

Q: Is the child holding a bottle or a can?

Image: visualqa.org
Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Visual Attention

Fine-grained visual processing is often essential for visual and language tasks.

Learn to focus on image regions related to the task.

Q: Is the child holding a bottle or a can?

A: Bottle

Image: visualqa.org
Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Visual Attention

Learn to focus on image regions related to the task.

attended feature ← $\hat{\boldsymbol{v}} = f(\boldsymbol{h}, V)$

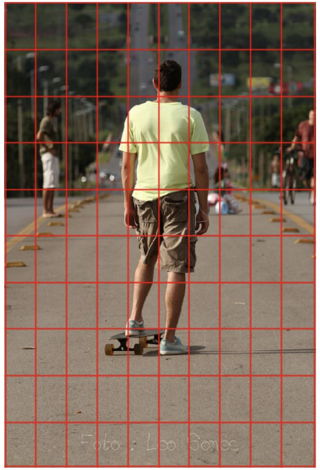
1. Set of attention candidate

2. Text content representation

3. Learned attention function

Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

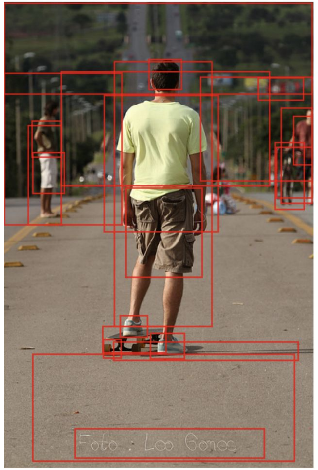
Attention Candidates



Spatial output of a CNN

$$V = \{v_1, \ldots, v_{100}\}$$

10 x 10 grids



Object-based attention

$$V = \{v_1, \ldots, v_k\}$$

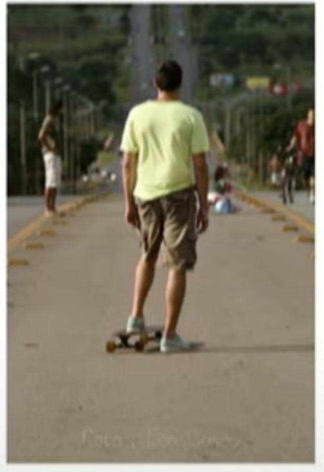
k regions

Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Visual Attention

Enabling attention to be calculated at the level of objects and other salient image regions.

It is the natural basis for attention to be considered.



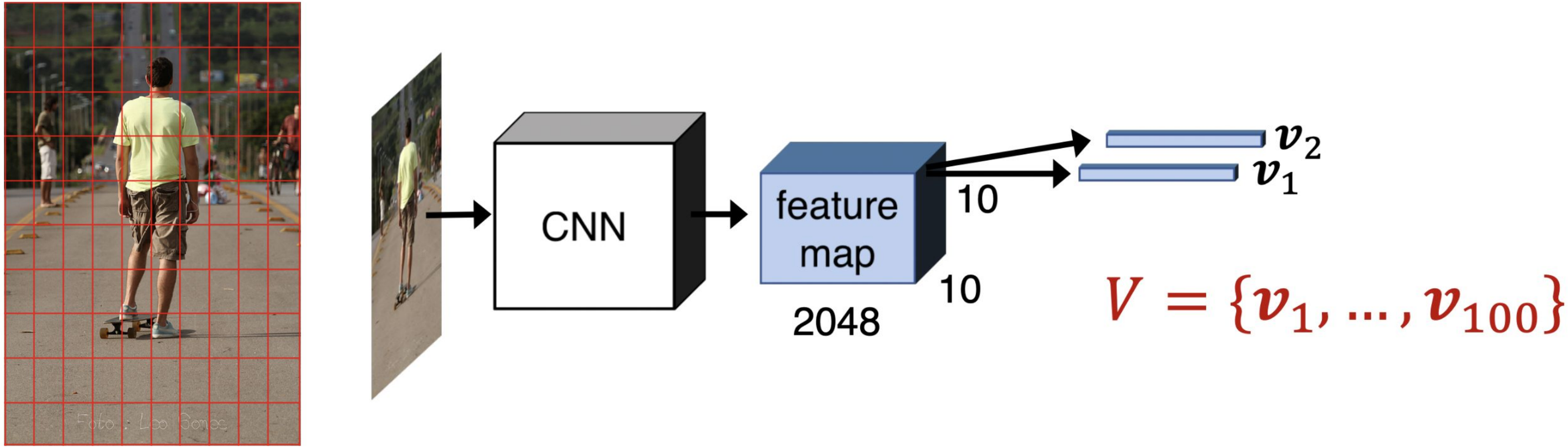
C: A **young man** on a **skateboard** looking down **street** with **people** watching.

Q: Is the **boy** in the **yellow shirt** wearing **head protective gear**?

A: No

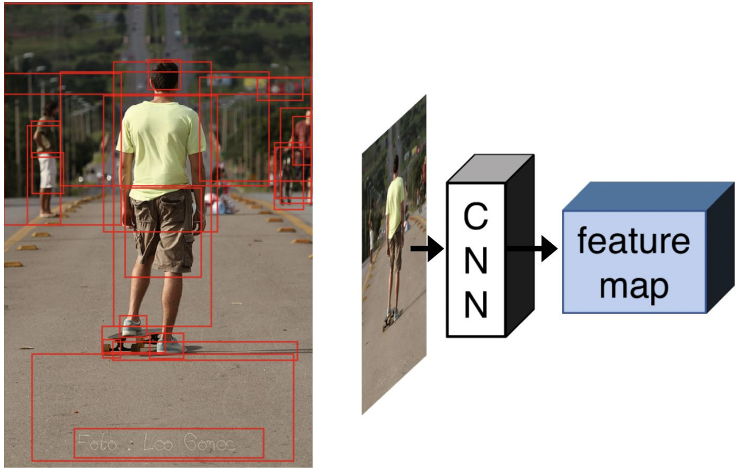
Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Spatial output of a CNN



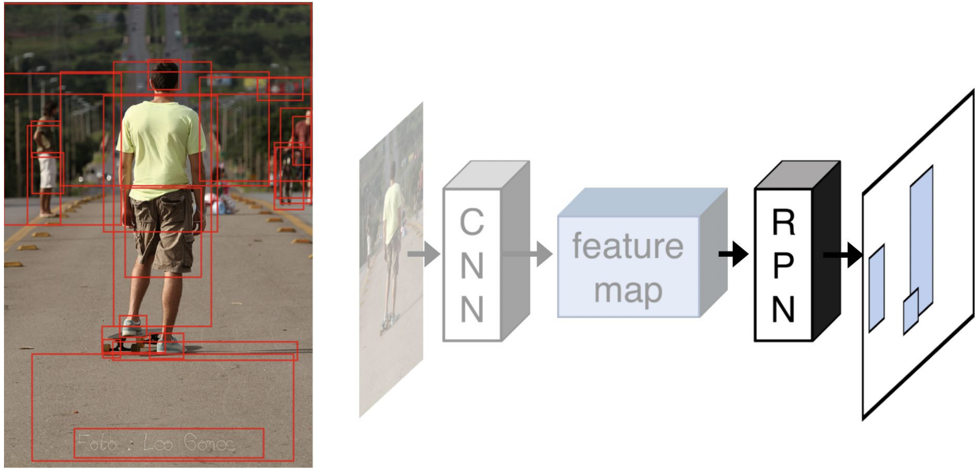
Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Bottom-Up Attention - Fast R-CNN



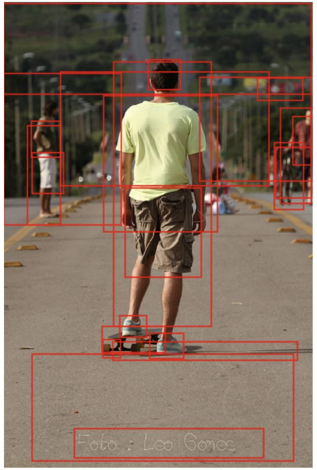
Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

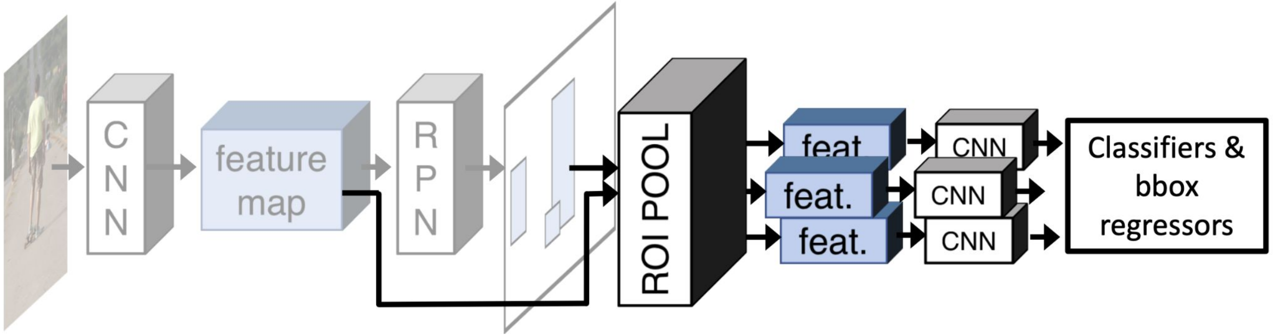
Bottom-Up Attention - Fast R-CNN



Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Bottom-Up Attention - Fast R-CNN



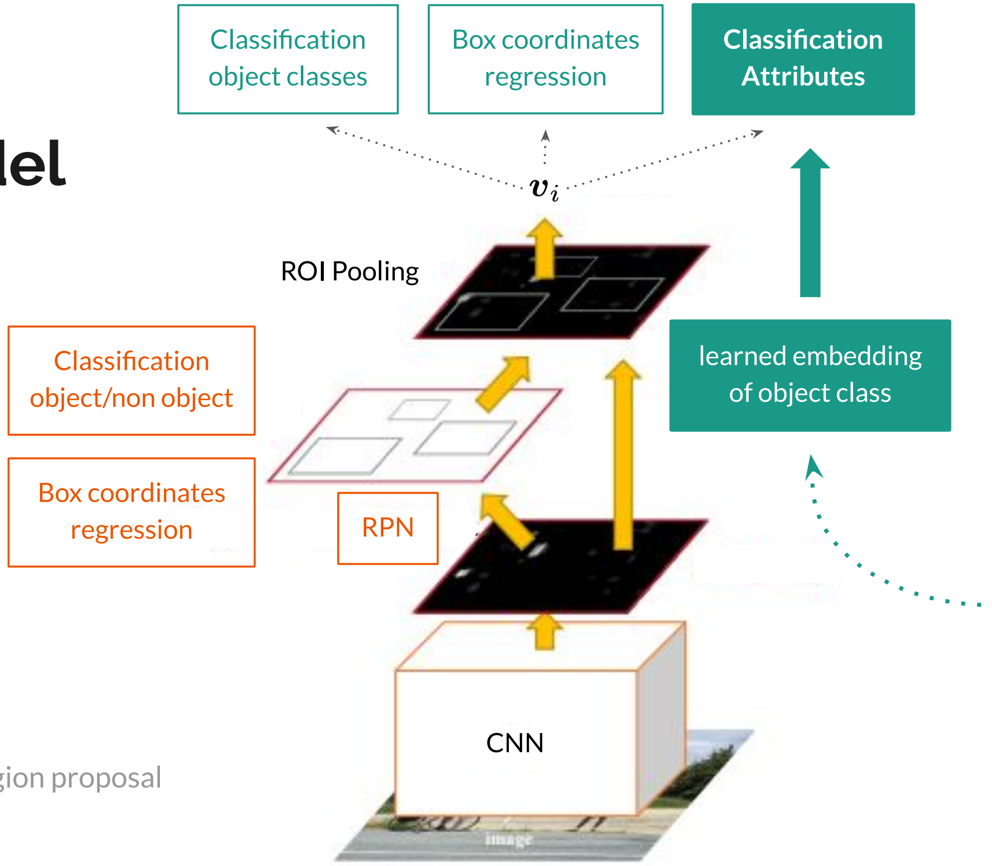


Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Bottom-up Attention Model

For each selected region i, $\boldsymbol{v}_i$ is defined as the mean-pooled convolutional feature from this region.

The original Faster R-CNN multi-task loss function contains four components. They add an additional multi-class loss component to train the attribute predictor.

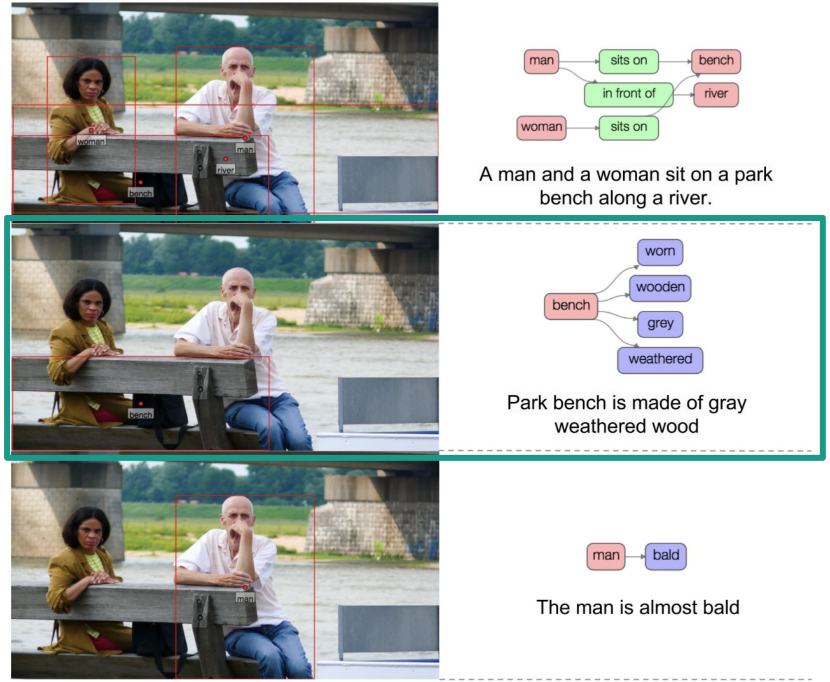


Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015

Pre-training
Visual Genome Dataset

Visual Genome is a dataset, a knowledge base, an ongoing effort to connect structured image concepts to language.

To aid the learning of good feature representations, we add an additional training output for predicting attribute classes (in addition to object classes).

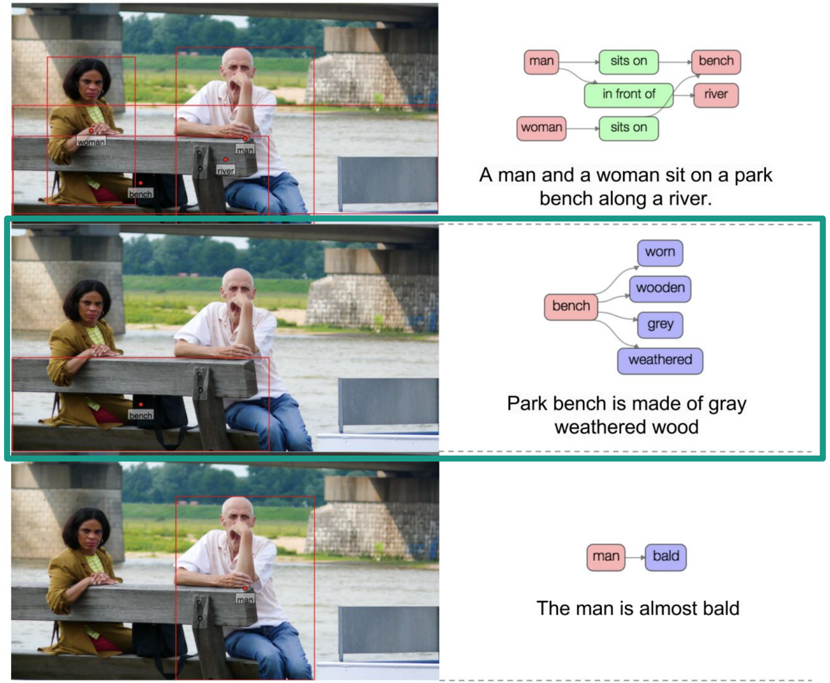


Visual genome: Connecting language and vision using crowdsourced dense image annotations. In arXiv, 2016.

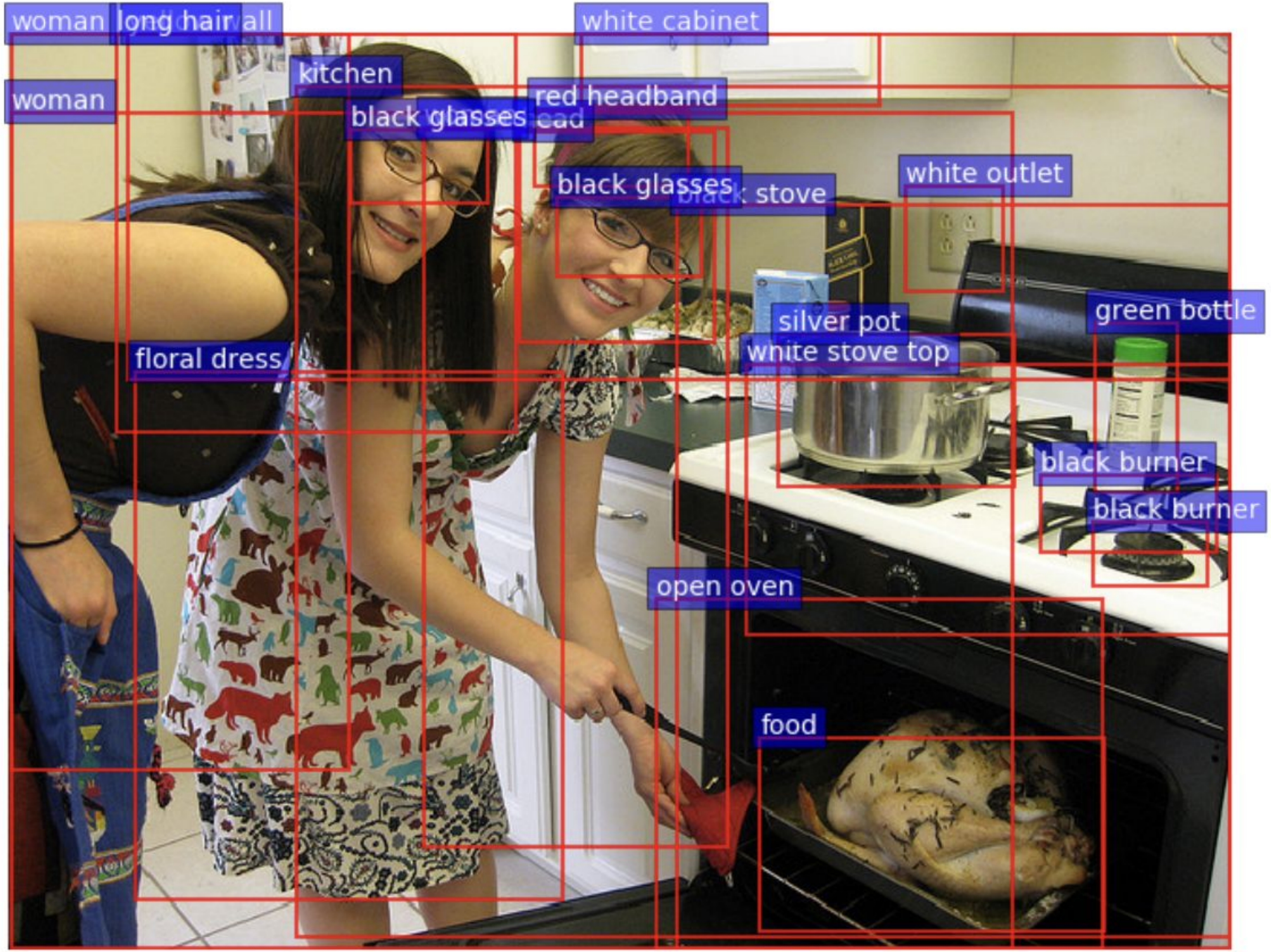
Pre-training
Visual Genome Dataset

1600 Object classes.

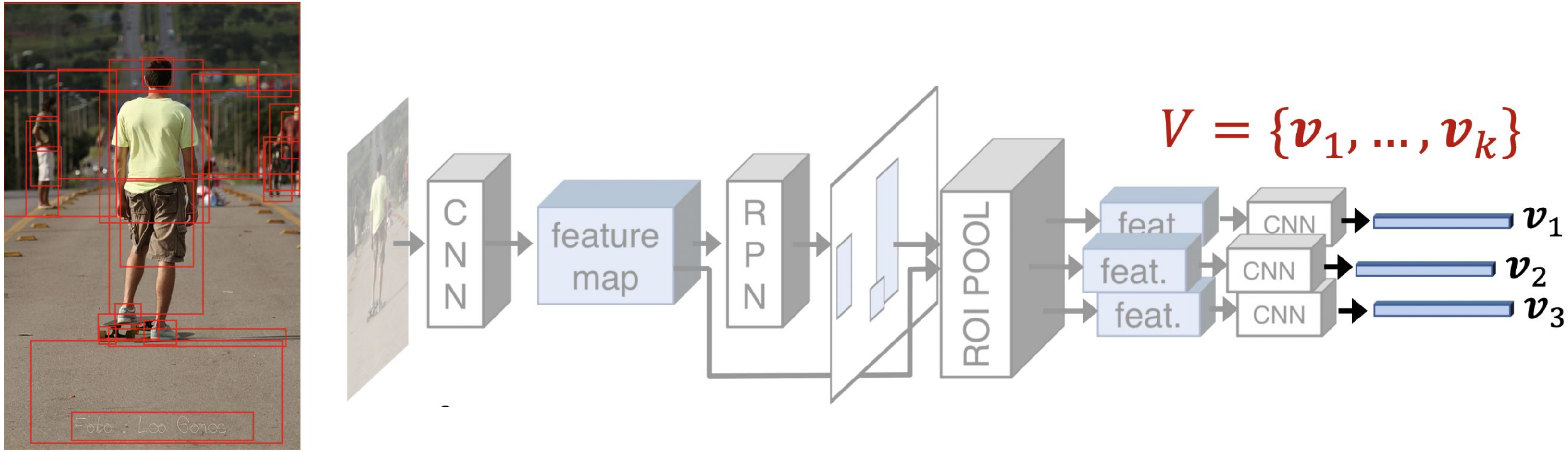
400 Attribute classes.



Visual genome: Connecting language and vision using crowdsourced dense image annotations. In arXiv preprint arxiv:1602.07332, 2016.


woman
long hair
white cabinet
kitchen
woman
red headband
black glasses
black glasses
white outlet
green bottle
silver pot
white stove top
floral dress
black burner
black burner
open oven
food

Bottom-Up Attention - Fast R-CNN

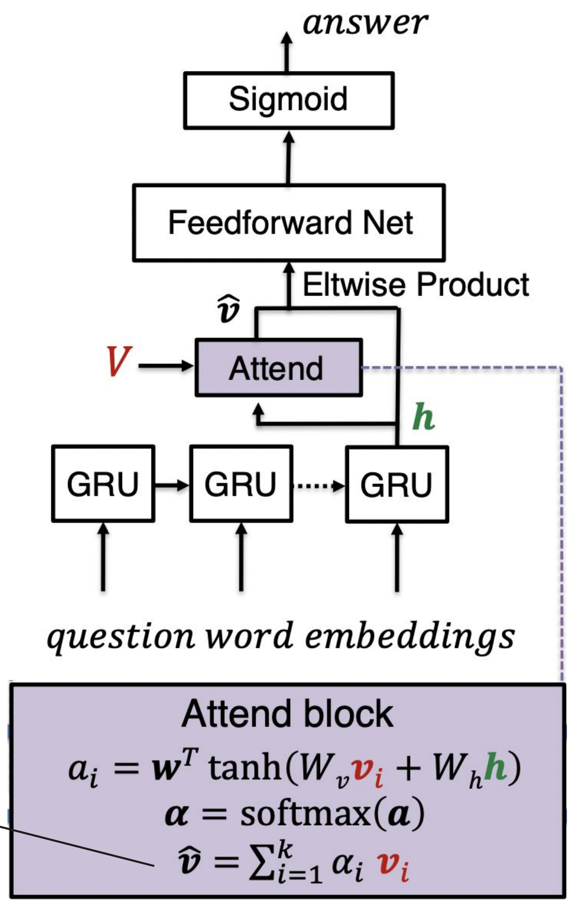


Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

VQA Model

Given a set of spatial image features V , their proposed VQA model also uses a ‘soft’ top-down attention mechanism to weight each feature, using the question representation as context.

$$\hat{\boldsymbol{v}} = f(\boldsymbol{h}, V)$$

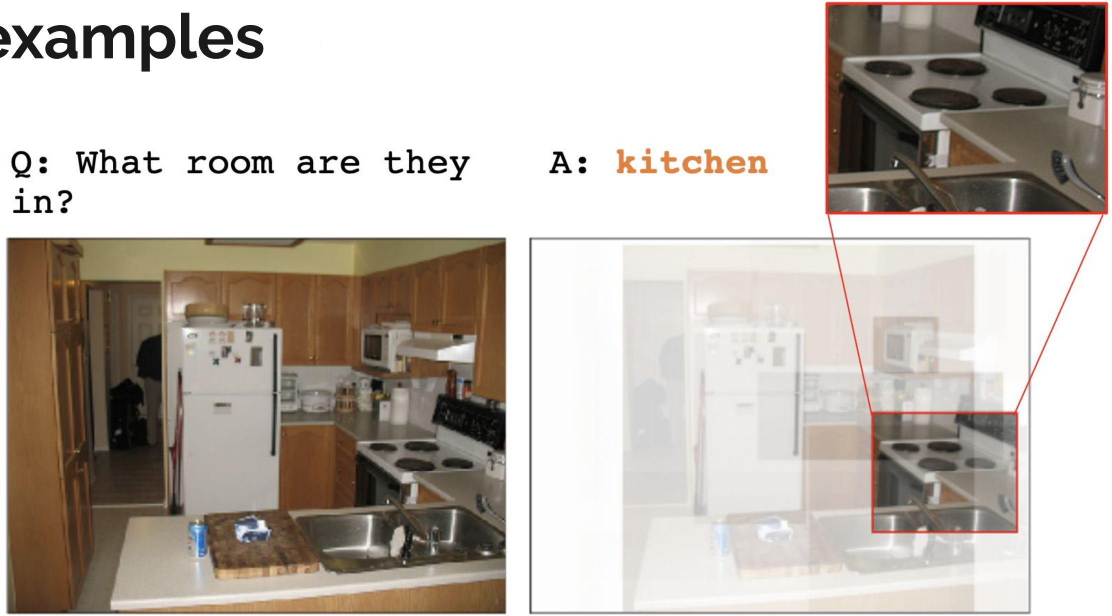


Bottom-up and top-down attention for image captioning and vqa.
In CVPR. IEEE, 2018.

ResNet (10×10): A man sitting on a ~~toilet~~ in a bathroom.

Up-Down (Ours): A man sitting on a **couch** in a bathroom.

VQA examples



Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

VQA examples - Counting

Q: How many oranges are on pedestals?

A: ~~two~~

Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

VQA examples - Reading

Q: What is the name of the realty company?

A: ~~none~~

Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Results

VQA v2 val set (single-model):

	Yes/No	Number	Other	Overall
ResNet (1×1)	76.0	36.5	46.8	56.3
ResNet (14×14)	76.6	36.2	49.5	57.9
ResNet (7×7)	77.6	37.7	51.5	59.4
Up-Down (Ours)	80.3	42.8	55.8	63.2

+4%

Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Benefits

Natural approach

Unifies vision & language tasks with object detection models

Transfer learning by pre-training on object detection datasets

Complementary to other models (just swap attention candidates)

Can be fine-tuned

Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.

Overview

Vision Language Pre-training

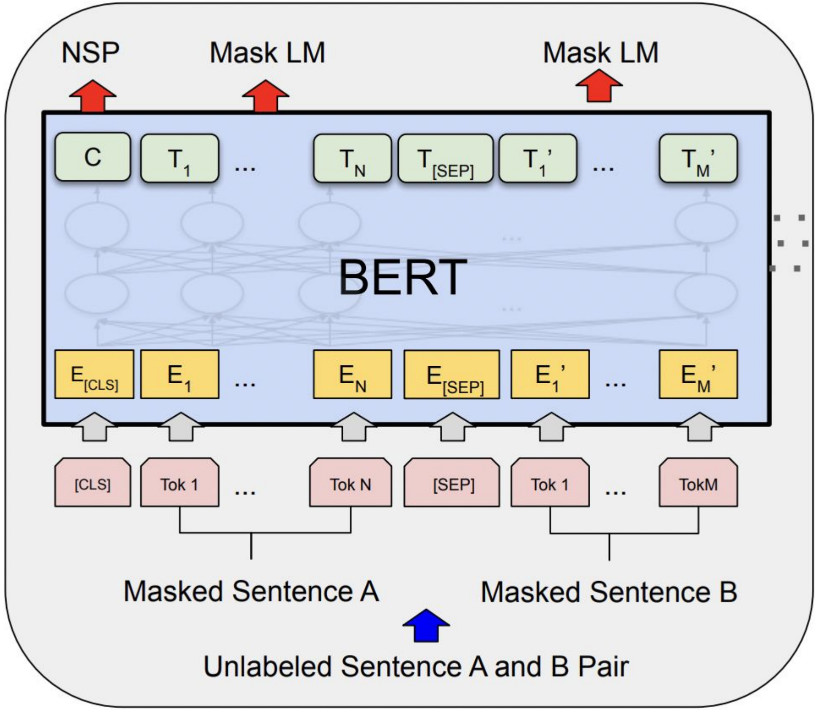
One of the important factors for performance improvement is pre-training on massive amounts of datasets.

IMAGENET

BERT
(Delvin et al, 2018)

Google's Conceptual Captions

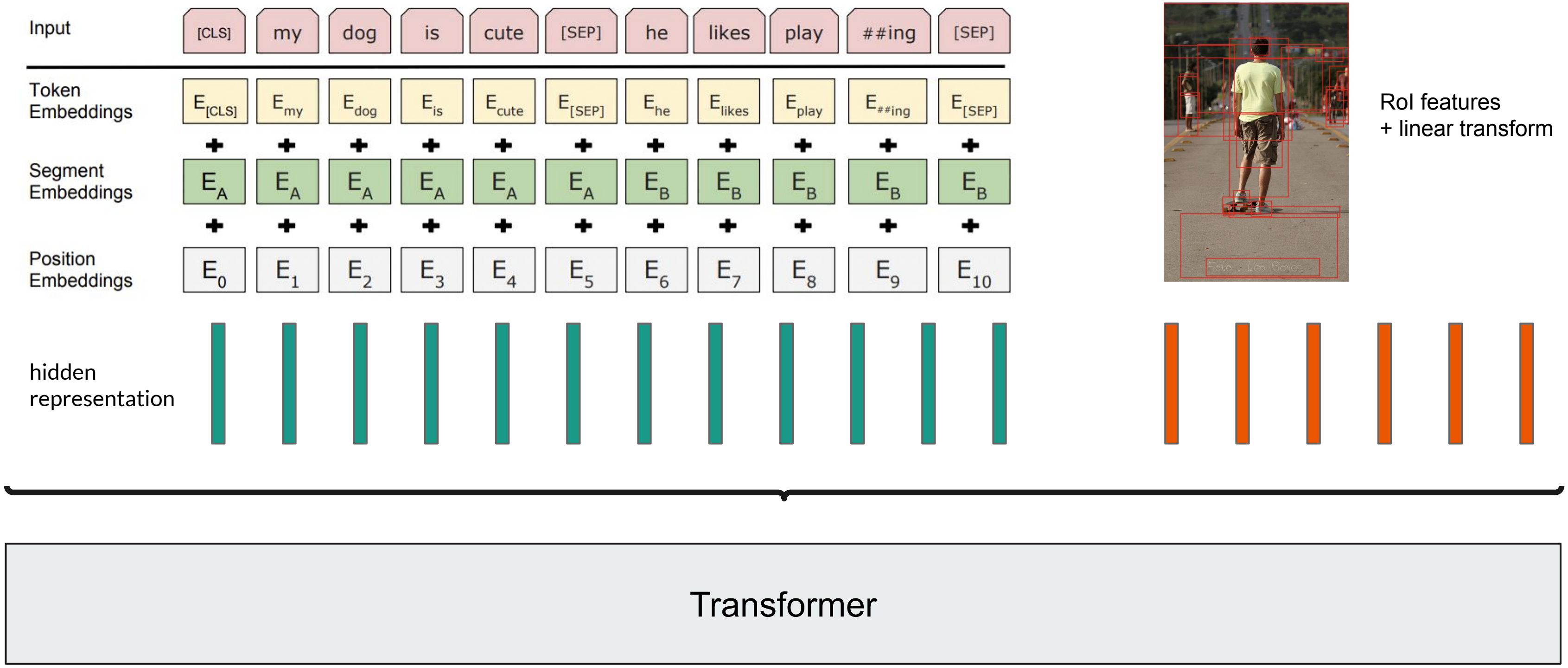
www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/



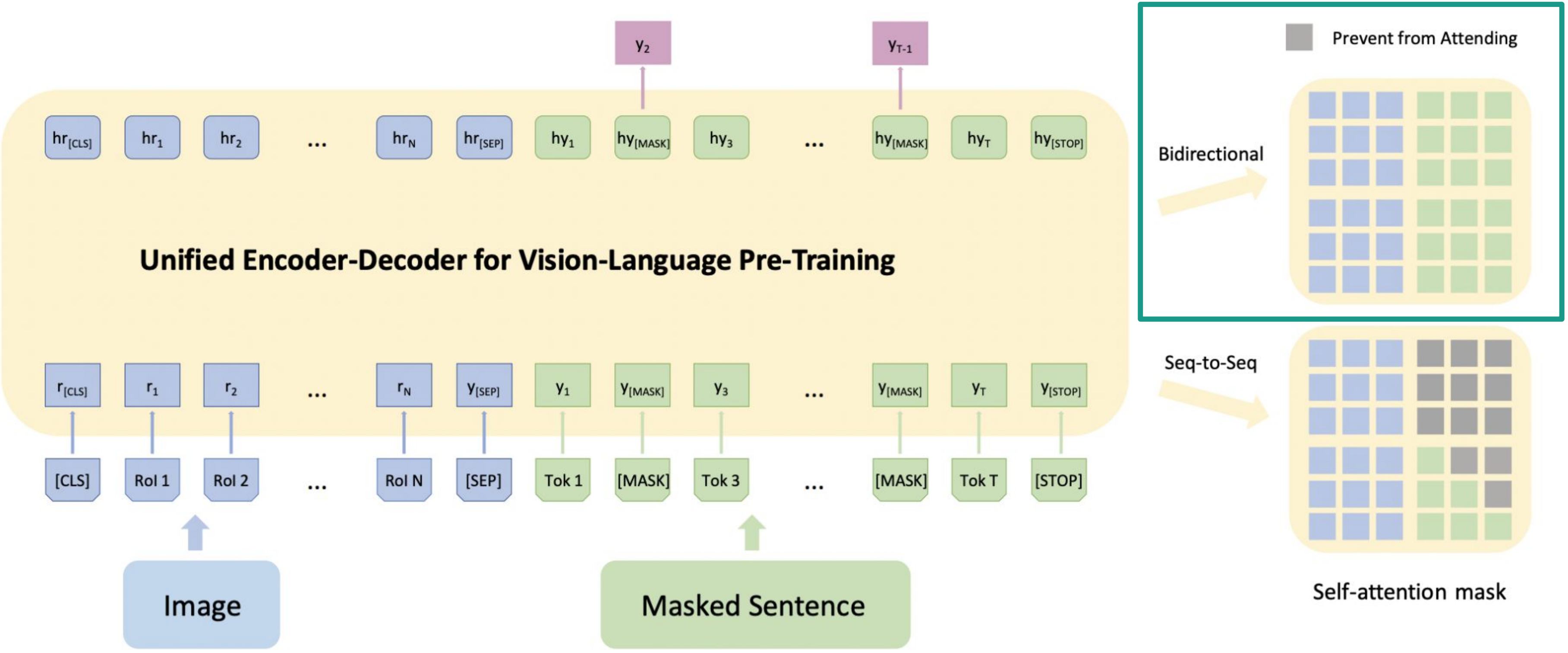
Masked Language Model (MLM)

input: regular sentence with [MASK] token
output: hidden representation of [MASK]
objective: predict vocabulary ID

www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/
Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018



www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/
Bert: Pre-training of deep bidirectional transformers for language understanding. In arXiv preprint arXiv:1810.04805, 2018



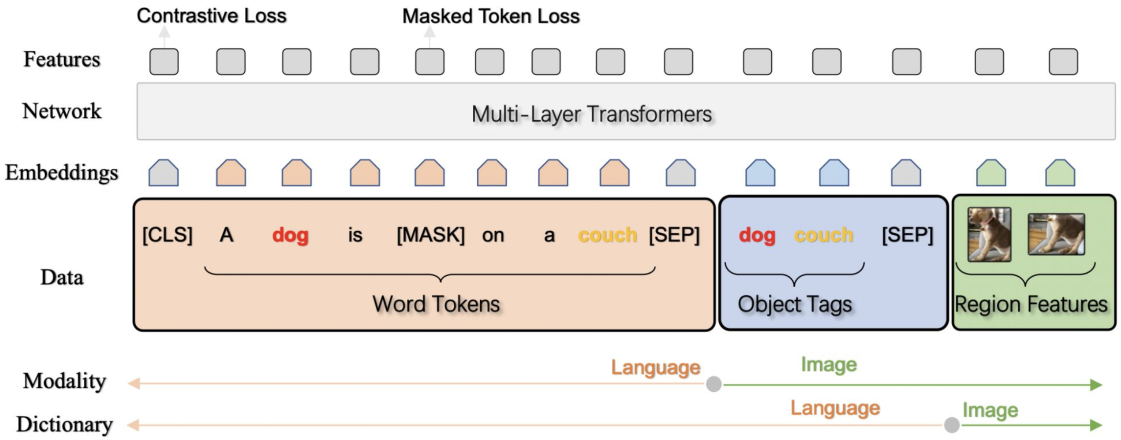
www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks/
Unified Vision-Language Pre-Training for Image Captioning and VQA. In AAAI, 2020.

Overview

OSCAR Model

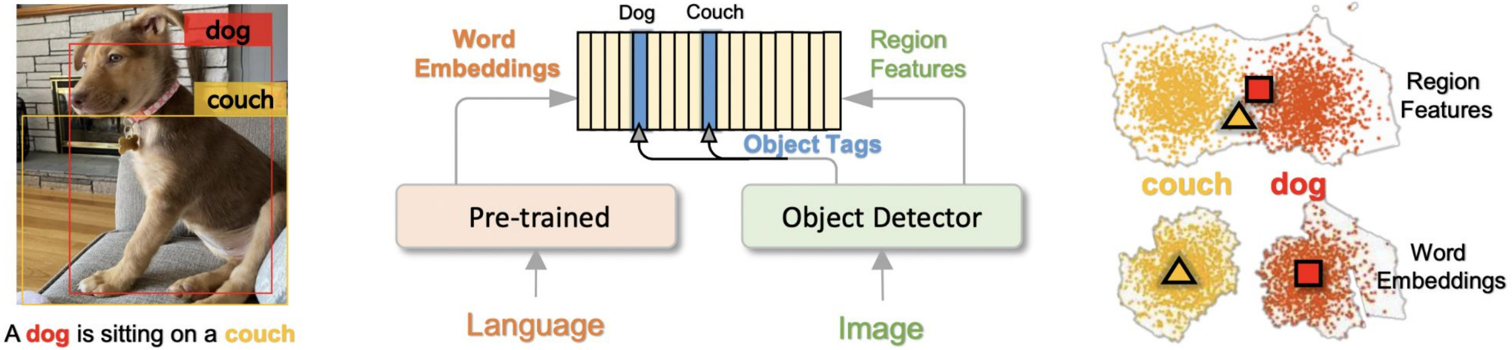
Objects tags is a language representation of visual concepts.

This is the key reason why OSCAR outperforms previous methods.



www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks
Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In ECCV, 2020

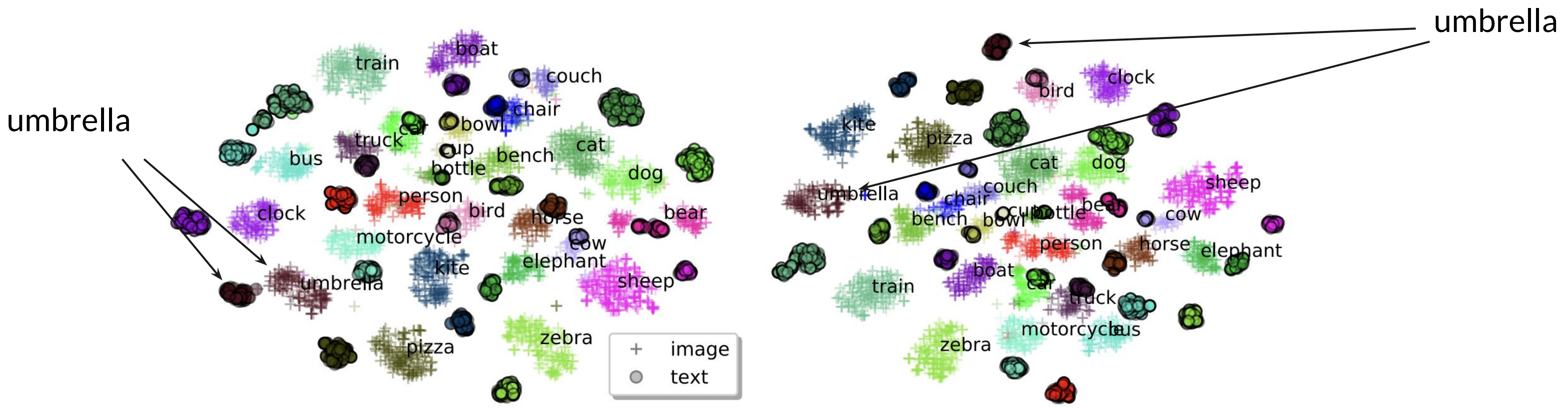
OSCAR Model



(a) Image-text pair (b) Objects as anchor points (c) Semantics spaces

www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks
Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In ECCV, 2020

OSCAR Model - Feature space visualization



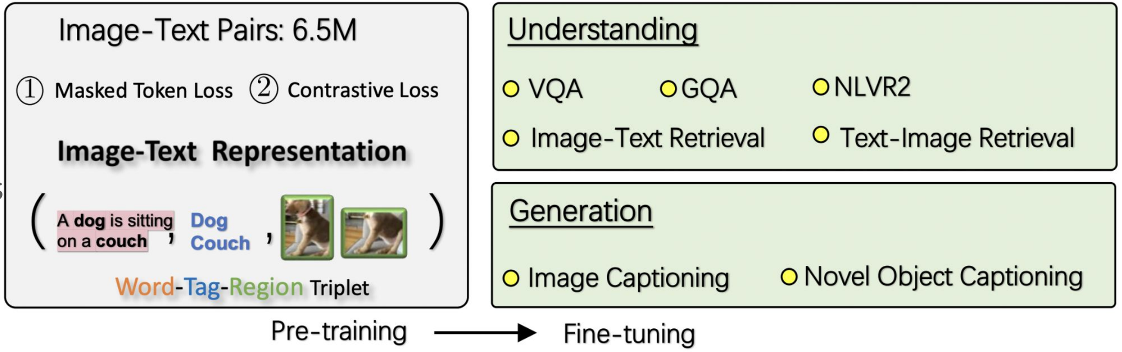
(a) OSCAR (b) Baseline (No tags)

www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks
Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In ECCV, 2020

OSCAR Model

A masked token loss over words & tags

A contrastive loss between tags and others.



www.microsoft.com/en-us/research/publication/oscar-object-semantics-aligned-pre-training-for-vision-language-tasks
Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In ECCV, 2020

Dataset

Table 5: Statistics of the pre-training corpus.

Source	COCO (train)	CC (all)	SBU (all)	Flicker30k (train)	VQA (train)	GQA (bal-train)	VG-QA (train)	Total
Image/Text	112k/560k	3.0M/3.0M	840k/840k	29k/145k	83k/444k	79k/1026k	48k/484k	4.1M/6.5M

Oscar model is pre-trained on a large-scale V+L dataset composed of 6.5 million pairs

Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In ECCV, 2020

Contrastive Loss

a contrastive loss for the modality view, which measures the model's capability of distinguishing an original triple and its "polluted" version (that is, where an original object tag is replaced with a randomly sampled one).

$$\mathcal{L}_{\mathrm{C}} = -\mathbb{E}_{(\boldsymbol{h}', \boldsymbol{w}) \sim \mathcal{D}} \log p(y | f(\boldsymbol{h}', \boldsymbol{w}))$$

Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In ECCV, 2020

Masked Token Loss

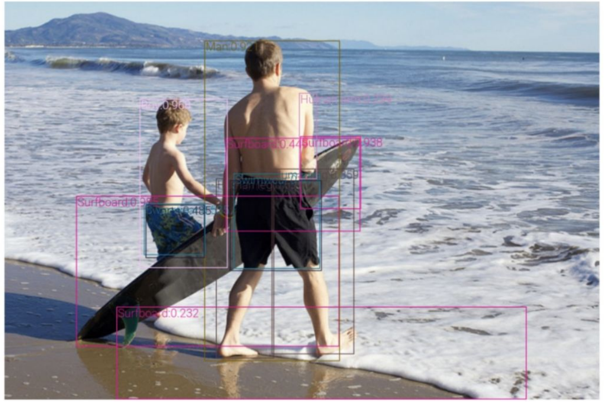
a masked token loss for the dictionary view, which measures the model's capability of recovering the masked element (word or object tag) based on its context

$$\mathcal{L}_{\text{MTL}} = -\mathbb{E}_{(\boldsymbol{v},\boldsymbol{h})\sim\mathcal{D}} \log p(h_i | \boldsymbol{h}_{\setminus i}, \boldsymbol{v})$$

www.microsoft.com/en-us/research/blog/objects-are-the-secret-key-to-revealing-the-world-between-vision-and-language
Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In ECCV, 2020

VinVL: Revisiting Visual Representations in Vision-Language Models

VinVL captures much richer image semantics



X152-FPN model trained on OpenImages

X152-C4 model trained on four public object detection datasets

datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
VinVL: Revisiting Visual Representations in Vision-Language Models. In CVPR, 2021

Visual feature	VQA		GQA		Image Captioning				NoCaps		Image Retrieval			Text Retrieval			NLVR2	
	test-dev	test-std	test-dev	test-std	B@4	M	C	S	C	S	R@1	R@5	R@10	R@1	R@5	R@10	dev	test-P
Anderson *et al.* [2]	73.16	73.44	61.58	61.62	40.5	29.7	137.6	22.8	86.58	12.38	54.0	80.8	88.5	70.0	91.1	95.5	78.07	78.36
Ours	**75.95**	**76.12**	**65.05**	**64.65**	**40.9**	**30.9**	**140.6**	**25.1**	**92.46**	**13.07**	**58.1**	**83.2**	**90.1**	**74.6**	**92.6**	**96.3**	**82.05**	**83.08**
Δ	**2.79** ↑	**2.68** ↑	**3.47** ↑	**3.03** ↑	**0.4** ↑	**1.2** ↑	**3.0** ↑	**2.3** ↑	**5.9** ↑	**0.7** ↑	**4.1** ↑	**2.4** ↑	**1.6** ↑	**4.6** ↑	**1.5** ↑	**0.8** ↑	**3.98** ↑	**4.71** ↑

Table 1: Uniform improvements on seven VL tasks by replacing visual features from Anderson *et al.* [2] with ours. The NoCaps baseline is from VIVO [9], and our results are obtained by directly replacing the visual features. The baselines for rest tasks are from OSCAR [21], and our results are obtained by replacing the visual features and performing OSCAR+ pre-training. All models are BERT-Base size. As analyzed in Section 5.2, the new visual features contributes 95% of the improvement.

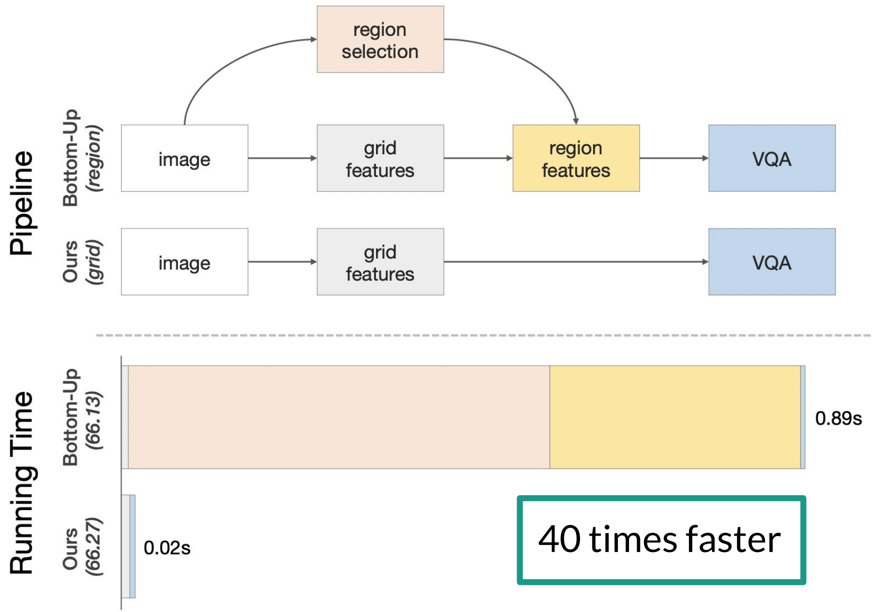
Visual feature matter!

datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
VinVL: Revisiting Visual Representations in Vision-Language Models. In CVPR, 2021

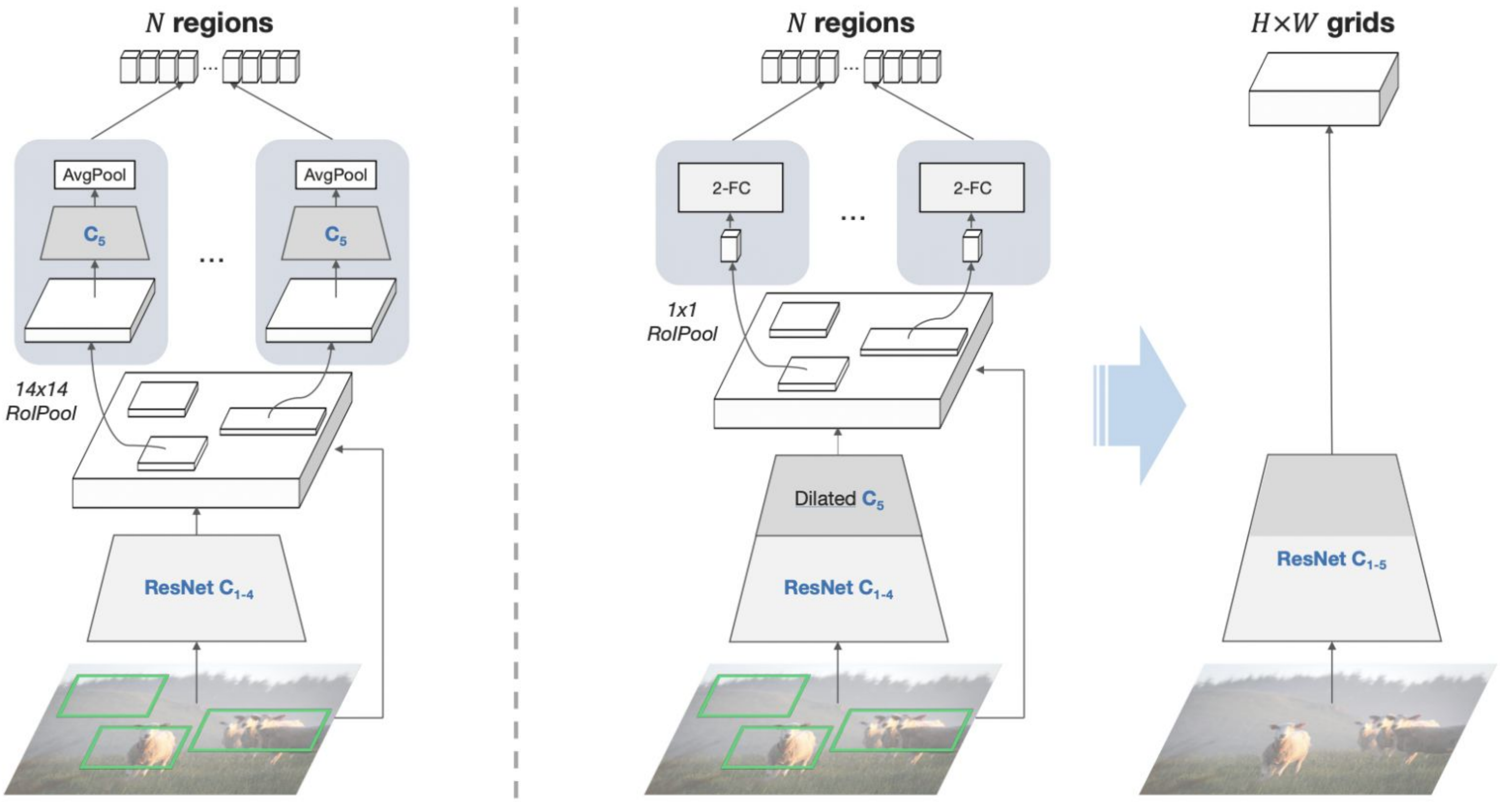
Overview

In Defense of Grid Features for Visual Question Answering

Using grid features is fast, and it can achieve comparable performance with regional features.



datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
In Defense of Grid Features for Visual Question Answering. In CVPR, 2020.



In Defense of Grid Features for Visual Question Answering. In CVPR, 2020.

In Defense of Grid Features for Visual Question Answering

Why previous methods based on grid features cannot outperform **Bottom-Up and Top-Down Attention** features?

datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
In Defense of Grid Features for Visual Question Answering. In CVPR, 2020.

In Defense of Grid Features for Visual Question Answering

Why previous methods based on grid features cannot outperform **Bottom-Up and Top-Down Attention** features?

1. **Pre-training task**
2. **Input image size**

datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
In Defense of Grid Features for Visual Question Answering. In CVPR, 2020.

In Defense of Grid Features for Visual Question Answering

	accuracy	time (ms)
Pythia [16]	68.31	-
R	68.21	929
G	67.76	39

(a)

	accuracy	time (ms)
MCAN [50]	70.93	-
R	72.01	963
G	72.59	72

(b)

	accuracy	time (ms)
Pythia [16]	54.22	-
R	54.28	874
G	54.17	38

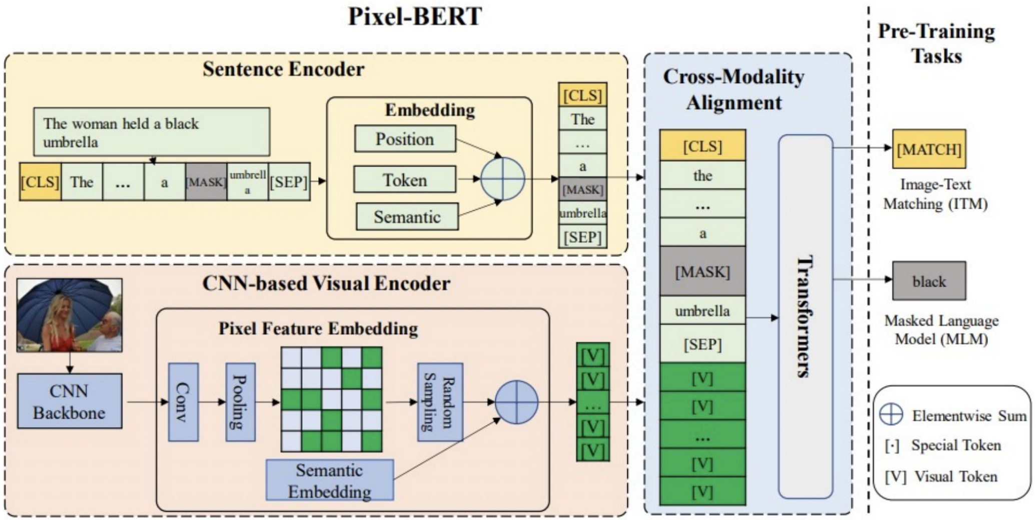
(c)

	B4	M	C	S	time (ms)
BUTD [2]	36.2	27.0	113.5	20.3	-
R	36.2	27.7	113.9	20.8	1101
G	36.4	27.4	113.8	20.7	240

(d)

datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
In Defense of Grid Features for Visual Question Answering. In CVPR, 2020.

Pixel-BERT: An E2E Pre-training Framework



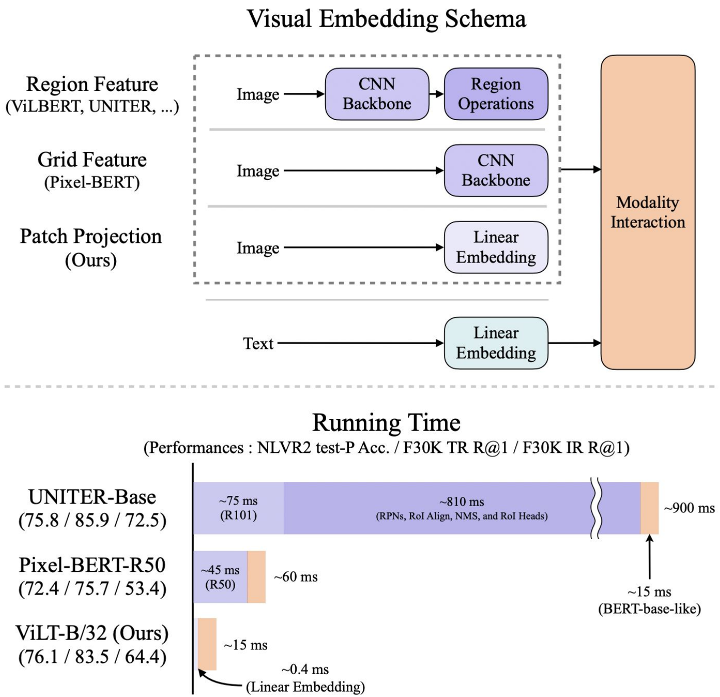
datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers. In arXiv, 2020

From grid features to region features, and to grid features again?

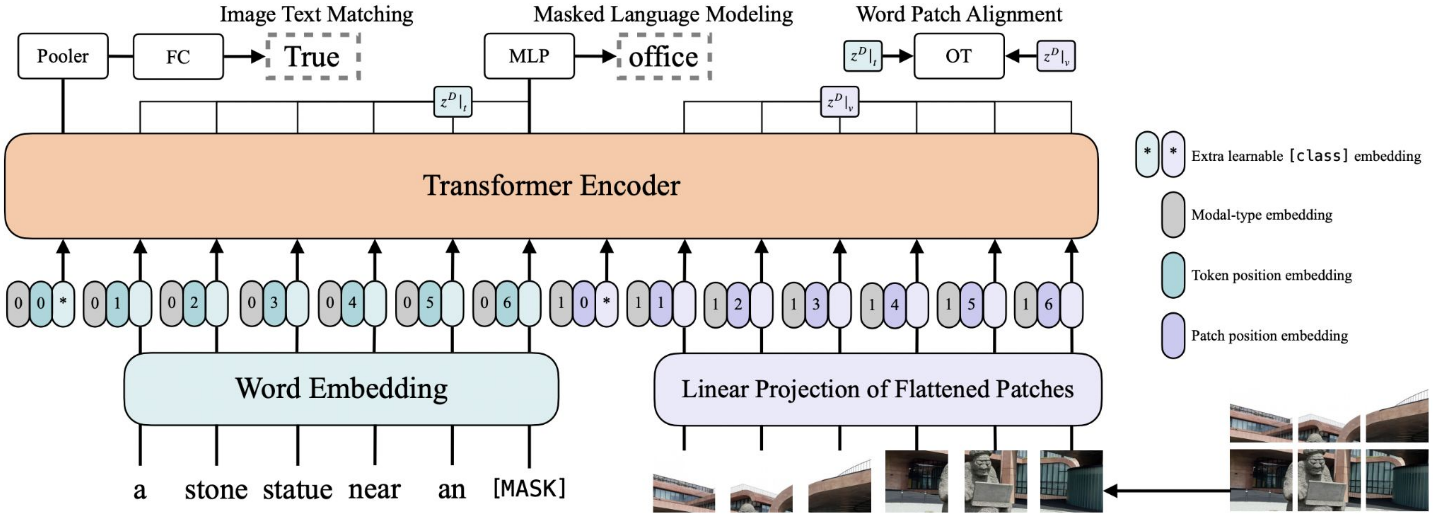
Overview

ViLT: Without Convolution or Region Supervision

ViLT is very fast since both object detection models and CNNs are not used.

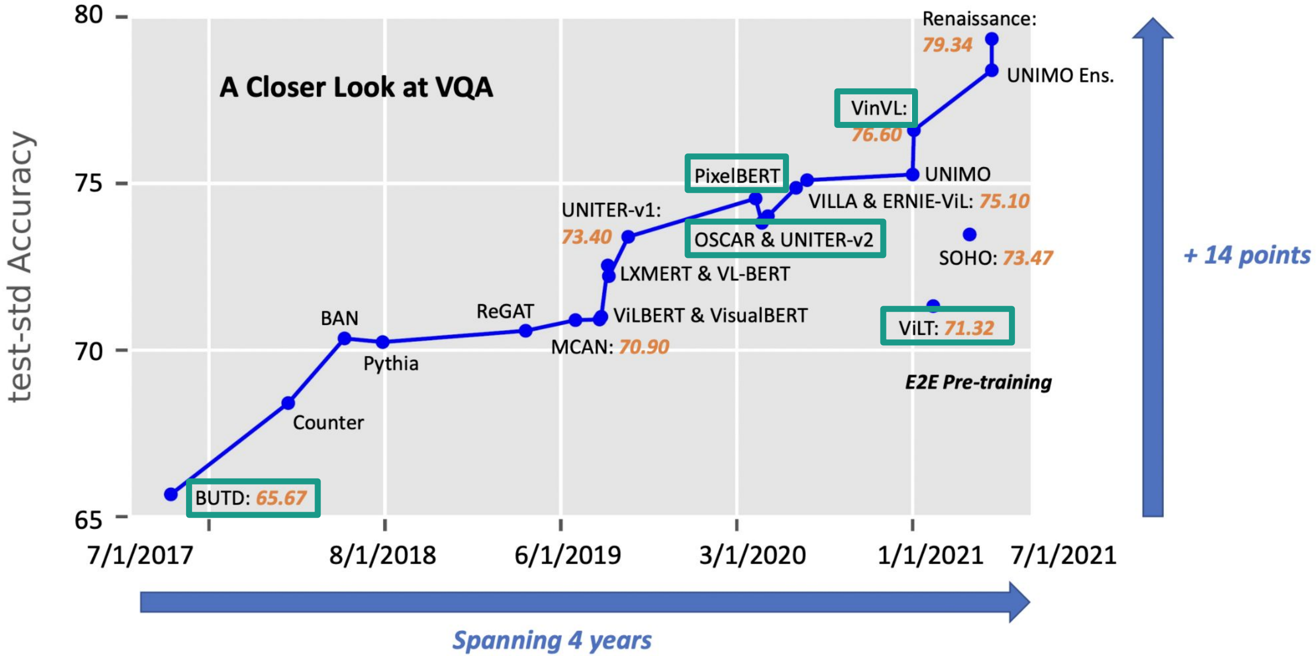


datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In ICML, 2021



A single unified transformer is learned

datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In ICML, 2021



However, performance-wise, it is still not ideal!

datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf

Thank you! Any Questions?

References

Papers:

- Bottom-up and top-down attention for image captioning and vqa. In CVPR. IEEE, 2018.
- Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015
- ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. In ICML, 202
- In Defense of Grid Features for Visual Question Answering. In CVPR, 2020.
- Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In ECCV, 2020
- Bert: Pre-training of deep bidirectional transformers for language understanding. In arXiv, 2018.
- In Defense of Grid Features for Visual Question Answering. In CVPR, 2020.
- VinVL: Revisiting Visual Representations in Vision-Language Models. In CVPR, 2021
- Visual genome: Connecting language and vision using crowdsourced dense image annotations. In arXiv, 2016.

References

Others:

- datarelease.blob.core.windows.net/tutorial/VQA2VLN2021/VLP_part1.pdf
- microsoft.com/en-us/research/blog/objects-are-the-secret-key-to-revealing-the-world-between-vision-and-language
- youtube.com/watch?v=A5Lzjpjiyzc
- youtube.com/watch?v=TBOnKekODCI
- youtube.com/watch?v=QNesnXfyYq8