Images and Languages

CS 6384 Computer Vision

Professor Yu Xiang

The University of Texas at Dallas

Image Classification

- ImageNet dataset
  - Training: 1.2 million images
  - Testing and validation: 150,000 images
  - 1000 categories

n02119789: kit fox, Vulpes macrotis
n02100735: English setter
n02096294: Australian terrier
n02066245: grey whale, gray whale, devilfish, Eschrichtius gibbosus, Eschrichtius robustus
n02509815: lesser panda, red panda, panda, bear cat, cat bear, Ailurus fulgens
n02124075: Egyptian cat
n02417914: ibex, Capra ibex
n02123394: Persian cat
n02125311: cougar, puma, catamount, mountain lion, painter, panther, Felis concolor
n02423022: gazelle

https://image-net.org/challenges/LSVRC/2012/index.php

Understand Images with Natural Languages

- Image captioning
- Object grounding
- Visual question answering
- Representation learning with images and languages

Image Captioning

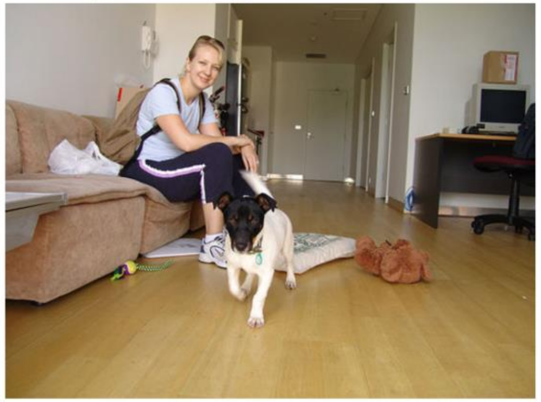
- Automatically generate texture descriptions of images

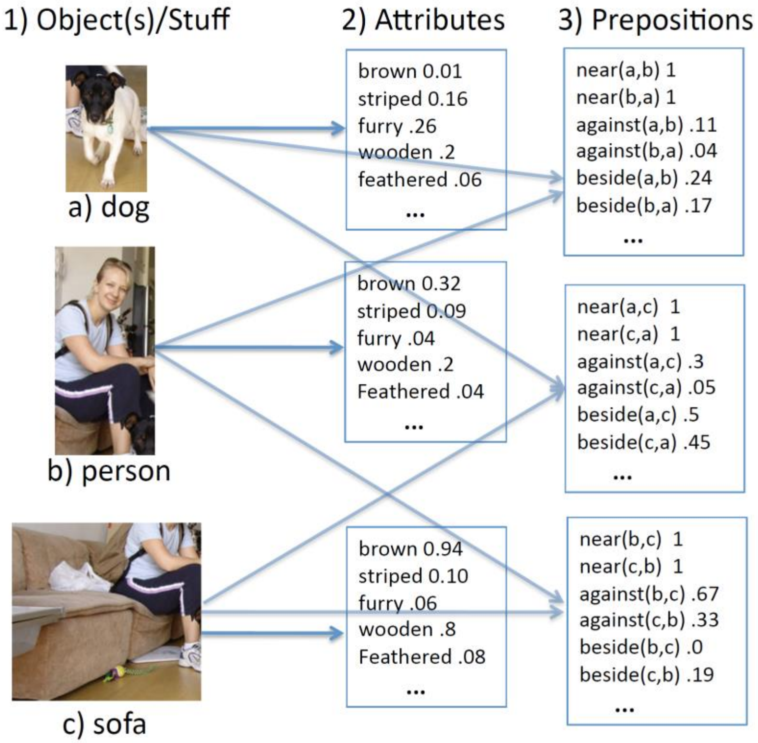
```
the person is riding a surfboard in the ocean
```

https://www.tensorflow.org/tutorials/text/image_captioning

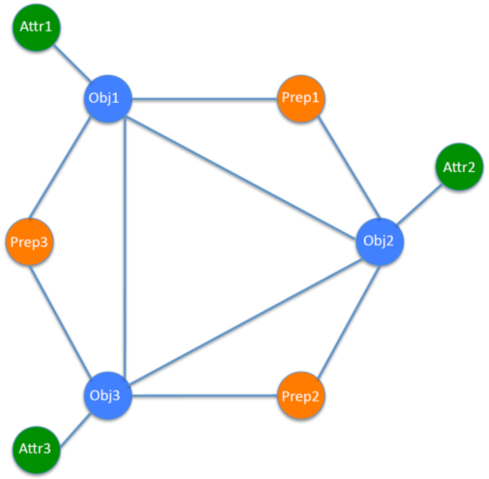
A Traditional Method for Image Captioning

Input Image





4) Constructed CRF



5) Predicted Labeling

<<null,person_b>,against,<brown,sofa_c>>
<<null,dog_a>,near,<null,person_b>>
<<null,dog_a>,beside,<brown,sofa_c>>

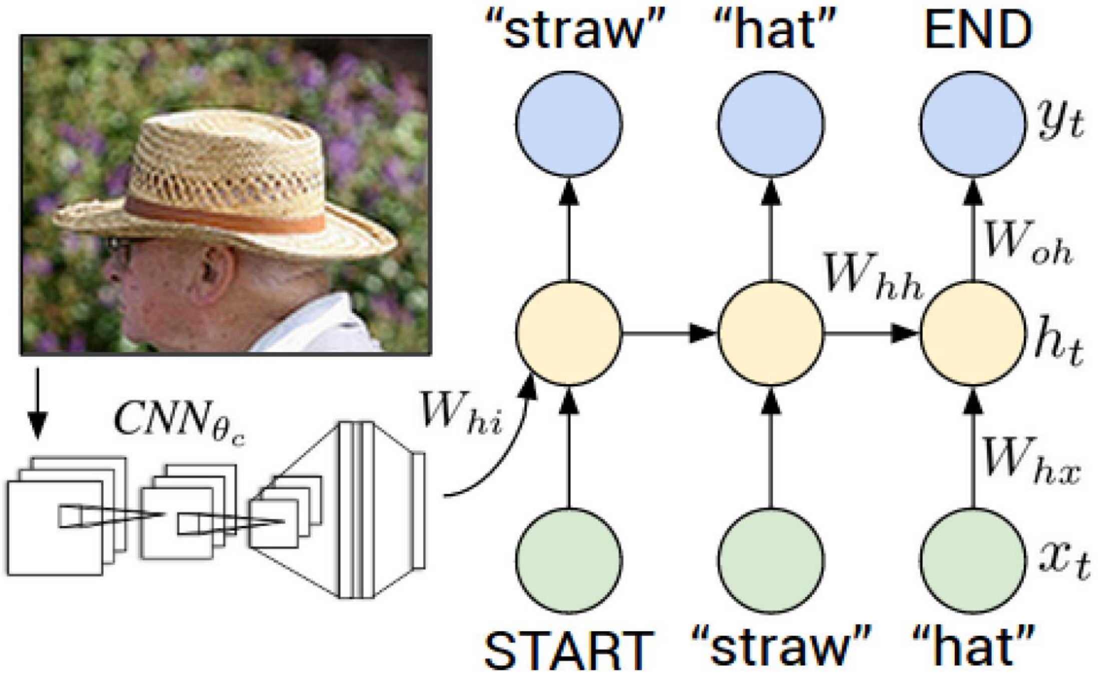
6) Generated Sentences

This is a photograph of one person and one brown sofa and one dog. The person is against the brown sofa. And the dog is near the person, and beside the brown sofa.

Using templates

Baby Talk: Understanding and Generating Image Descriptions. Kulkarni et al., CVPR, 2011

Image Captioning with RNNs



- Image embedding

$$b_v = W_{hi}[CNN_{\theta_c}(I)]$$

- Hidden state at time t

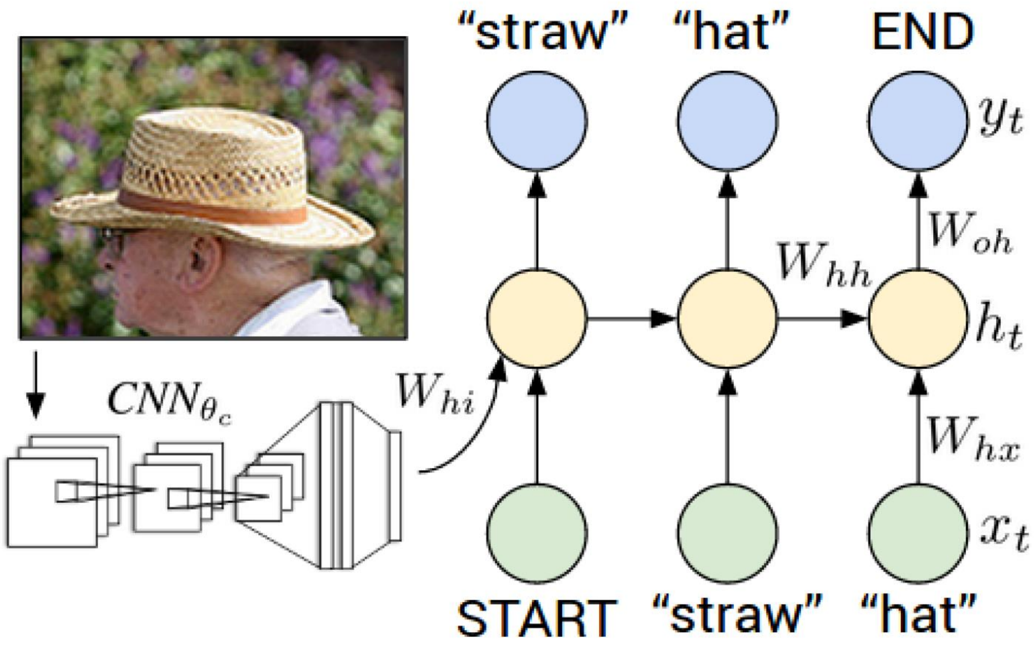
$$h_t = f(W_{hx}x_t + W_{hh}h_{t-1} + b_h + \mathbb{1}(t=1) \odot b_v)$$

Parameters

- Word embedding $x_t = W_w \mathbb{I}_t$
- Output $y_t = softmax(W_{oh}h_t + b_o)$

Deep Visual-Semantic Alignments for Generating Image Descriptions. Karpathy & Fei-fei, CVPR, 2015

Image Captioning with RNNs

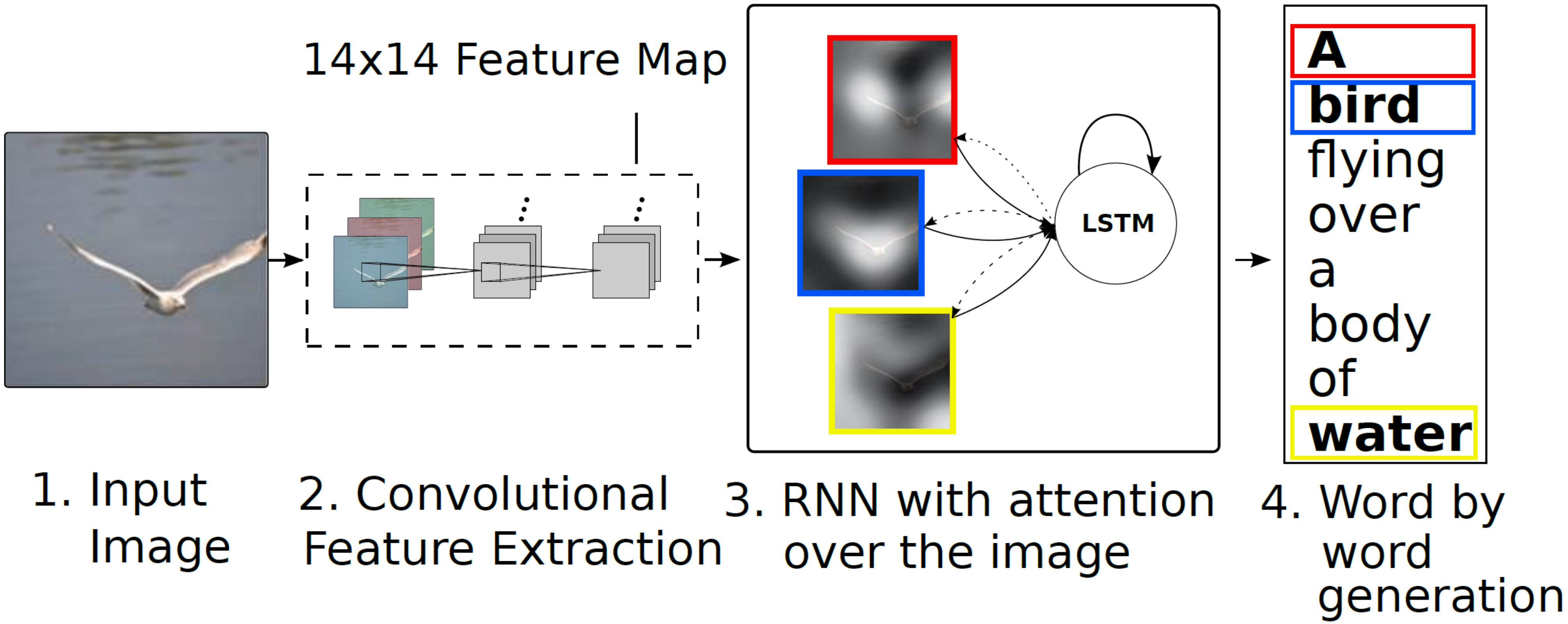


man in black shirt is playing guitar.

construction worker in orange safety vest is working on road.

Deep Visual-Semantic Alignments for Generating Image Descriptions. Karpathy & Fei-fei, CVPR, 2015

Image Captioning with Attentions



Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Xu et al., PMLR, 2015.

Image Captioning with Attentions

14x14 Feature Map

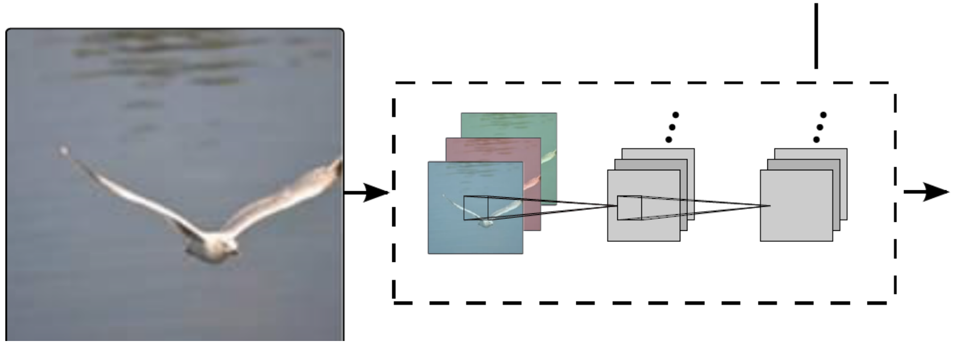


Image features for different locations

$$a = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \ \mathbf{a}_i \in \mathbb{R}^D$$

LSTM for caption generation

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \mathbf{g}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} T_{D+m+n,n} \begin{pmatrix} \mathbf{E}\mathbf{y}_{t-1} \\ \mathbf{h}_{t-1} \\ \hat{\mathbf{z}}_t \end{pmatrix}$$

Word embedding

Context vector

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t).$$

Attention

$$e_{ti} = f_{\text{att}}(\mathbf{a}_i, \mathbf{h}_{t-1})$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}$$

$$\hat{\mathbf{z}}_t = \phi\left(\{\mathbf{a}_i\}, \{\alpha_i\}\right)$$

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Xu et al., PMLR, 2015.

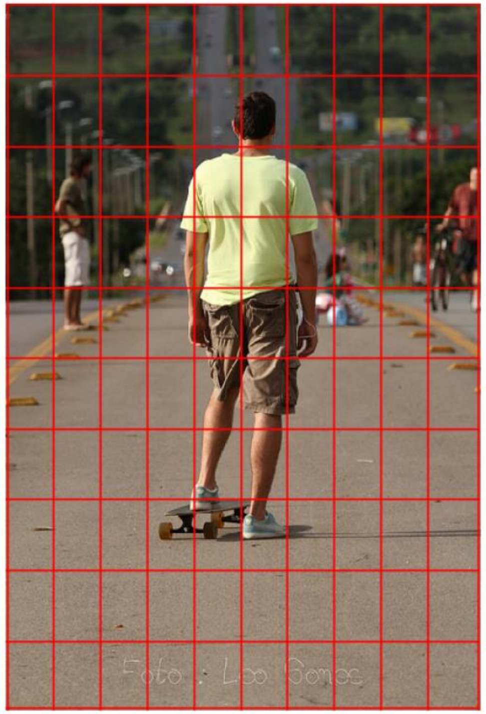
Image Captioning with Attentions

Dataset	Model	BLEU				
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)$^{\dagger\Sigma}$	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)$^{\circ}$	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC$^{\dagger\circ\Sigma}$	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)^{a}	—	—	—	—	20.41
	MS Research (Fang et al., 2014)$^{\dagger a}$	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)$^{\circ}$	64.2	45.1	30.4	20.3	—
	Google NIC$^{\dagger\circ\Sigma}$	66.6	46.1	32.9	24.6	—
	Log Bilinear$^{\circ}$	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

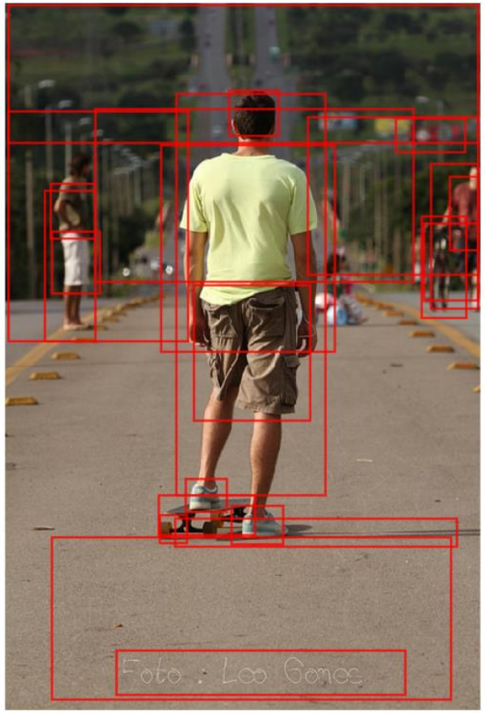
BLEU (BiLingual Evaluation Understudy) **METEOR (Metric for Evaluation of Translation with Explicit ORdering)**

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. Xu et al., PMLR, 2015.

Image Captioning with Object Detection



Grid-based attention



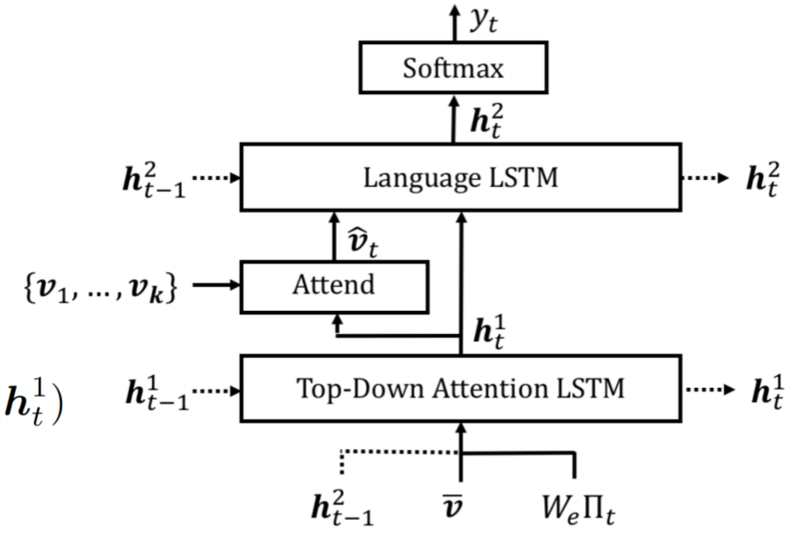
Object detection-based attention

Object detection features $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\}$

RoI pooling from Faster R-CNN

LSTM-based model

$$\bar{\boldsymbol{v}} = \frac{1}{k}\sum_i \boldsymbol{v}_i$$



Attention

$$a_{i,t} = \boldsymbol{w}_a^T \tanh\left(W_{va}\boldsymbol{v}_i + W_{ha}\boldsymbol{h}_t^1\right)$$
$$\boldsymbol{\alpha}_t = \text{softmax}\left(\boldsymbol{a}_t\right)$$

$$\hat{\boldsymbol{v}}_t = \sum_{i=1}^{K} \alpha_{i,t}\boldsymbol{v}_i$$

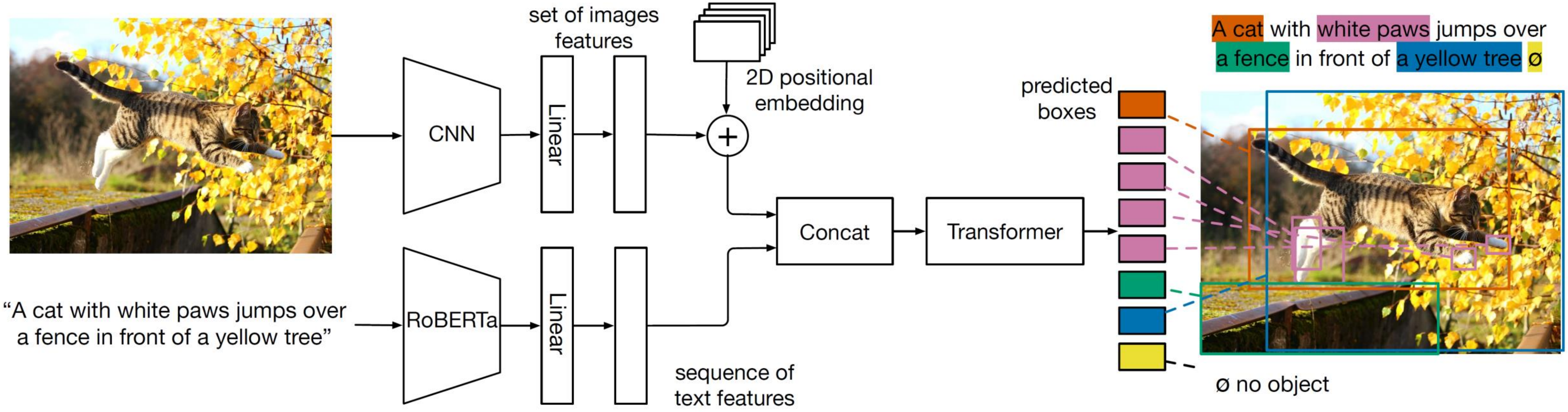
Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Anderson et al., CVPR, 2018

Object Grounding

A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. Plummer et al., ICCV, 2015.

Object Grounding



MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. Kamath et al., 2021

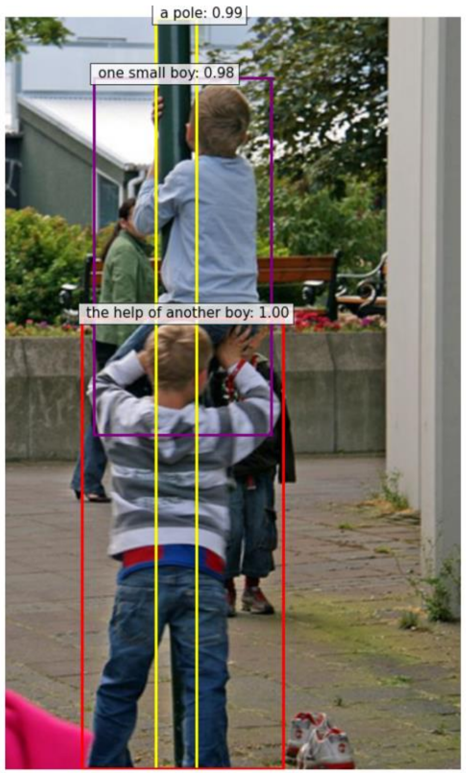
Object Grounding

- Soft token prediction
  - For each detected bounding, predict a probability distribution over the tokens in the input phase

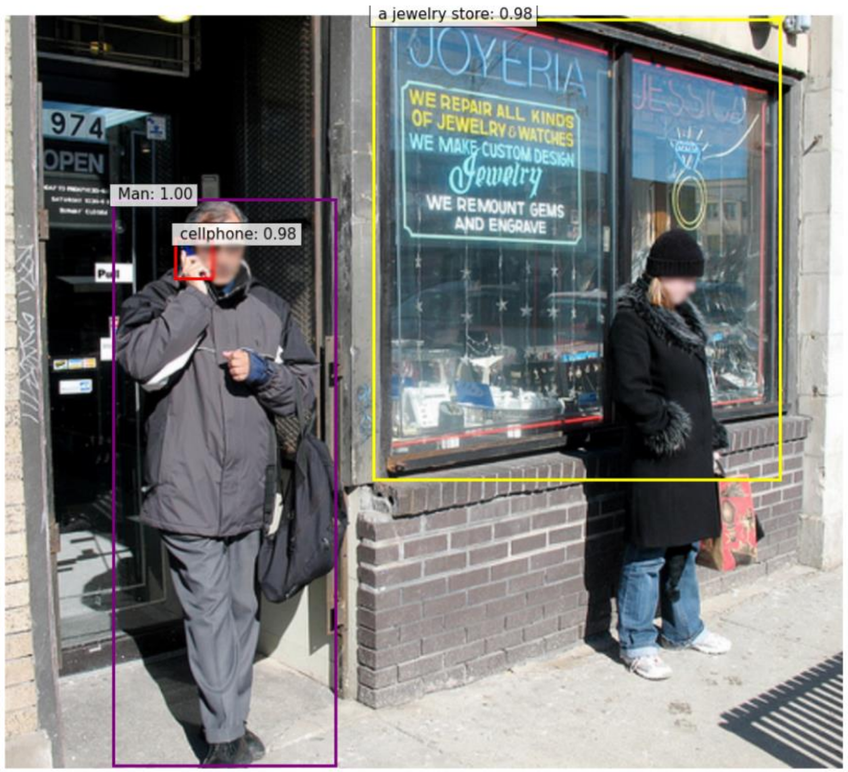
maximum number of tokens: 256

MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. Kamath et al., 2021

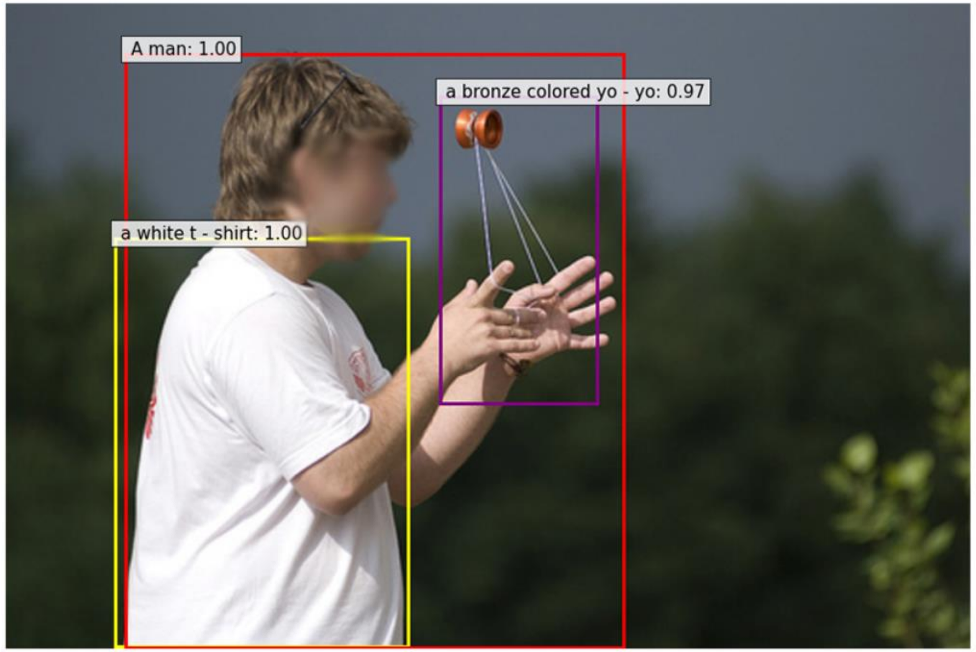
Object Grounding



(a) "one small boy climbing a pole with the help of another boy on the ground"



(b) "A man talking on his cellphone next to a jewelry store"



(c) "A man in a white t-shirt does a trick with a bronze colored yo-yo"

MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. Kamath et al., 2021

Visual Question Answering

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

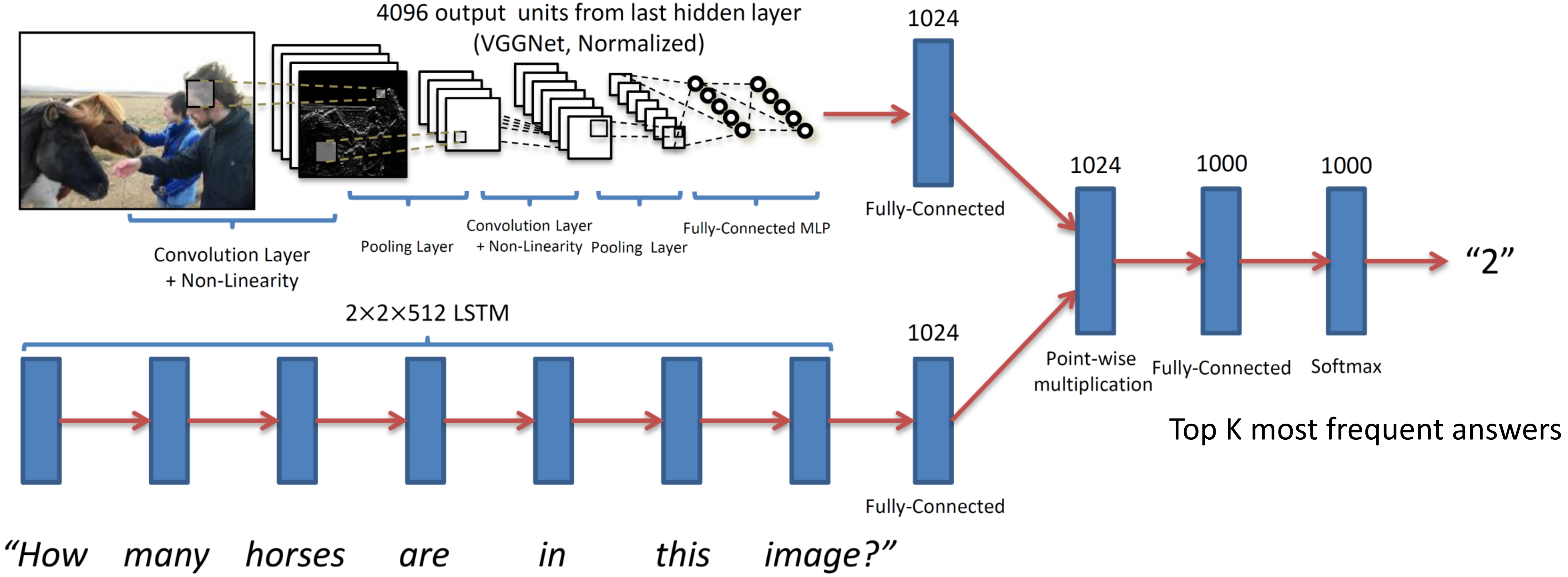
How many slices of pizza are there?
Is this a vegetarian pizza?

- Input
  - An image
  - A free-form, open-ended, natural language question
- Output
  - Case 1: open-ended answer
  - Case 2: multiple-choice task

$$\text{accuracy} = \min\left(\frac{\text{\# humans that provided that answer}}{3}, 1\right)$$

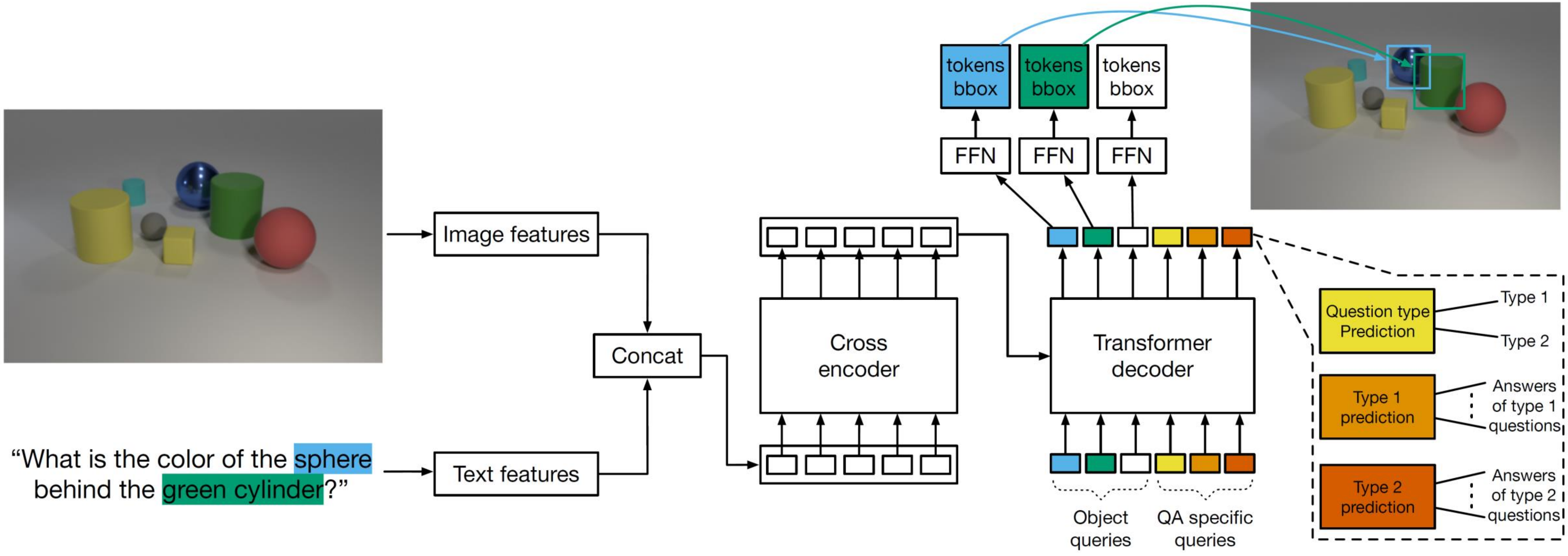
VQA: Visual Question Answering. Agrawal et al., ICCV, 2015

Visual Question Answering



VQA: Visual Question Answering. Agrawal et al., ICCV, 2015

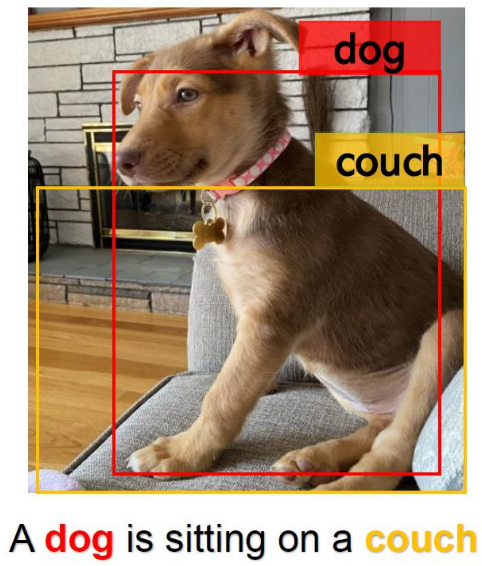
Visual Question Answering

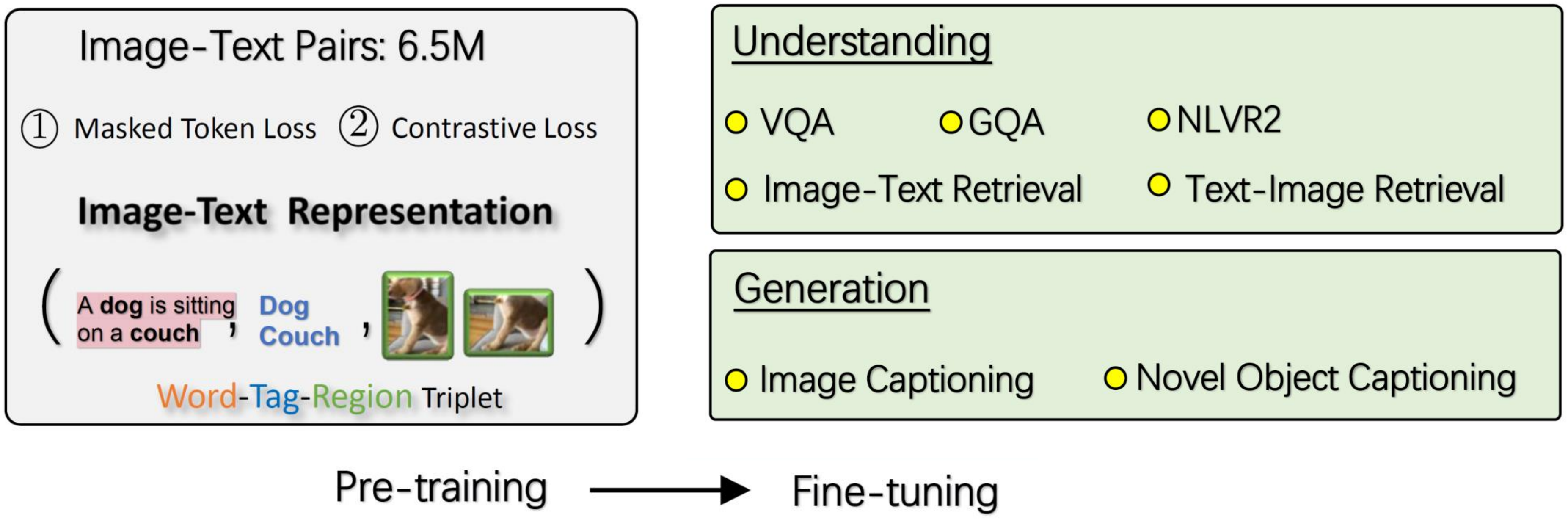


MDETR - Modulated Detection for End-to-End Multi-Modal Understanding. Kamath et al., 2021

Representation Learning

- Can we learn feature representations of images and text that can be useful for various vision-language tasks? (pre-training)

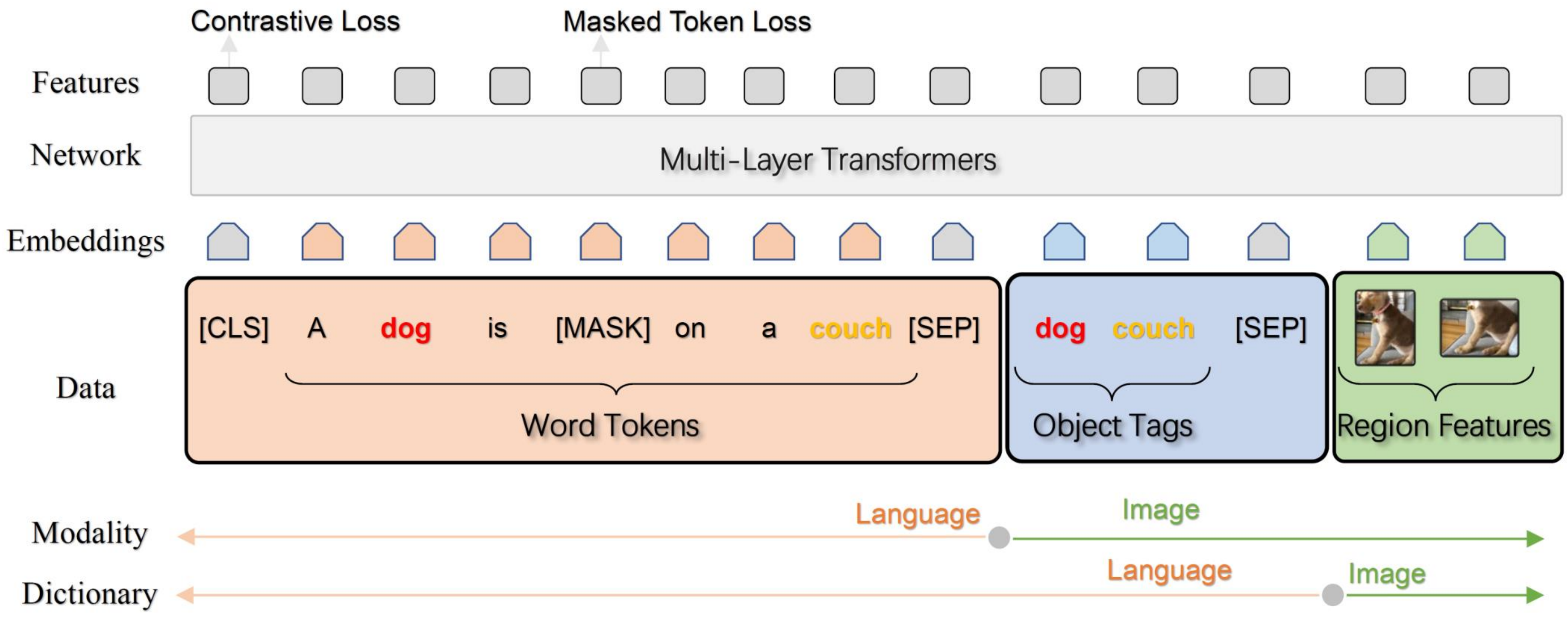




Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. Li et al., ECCV, 2020

Oscar: Object-Semantics Aligned Pre-training

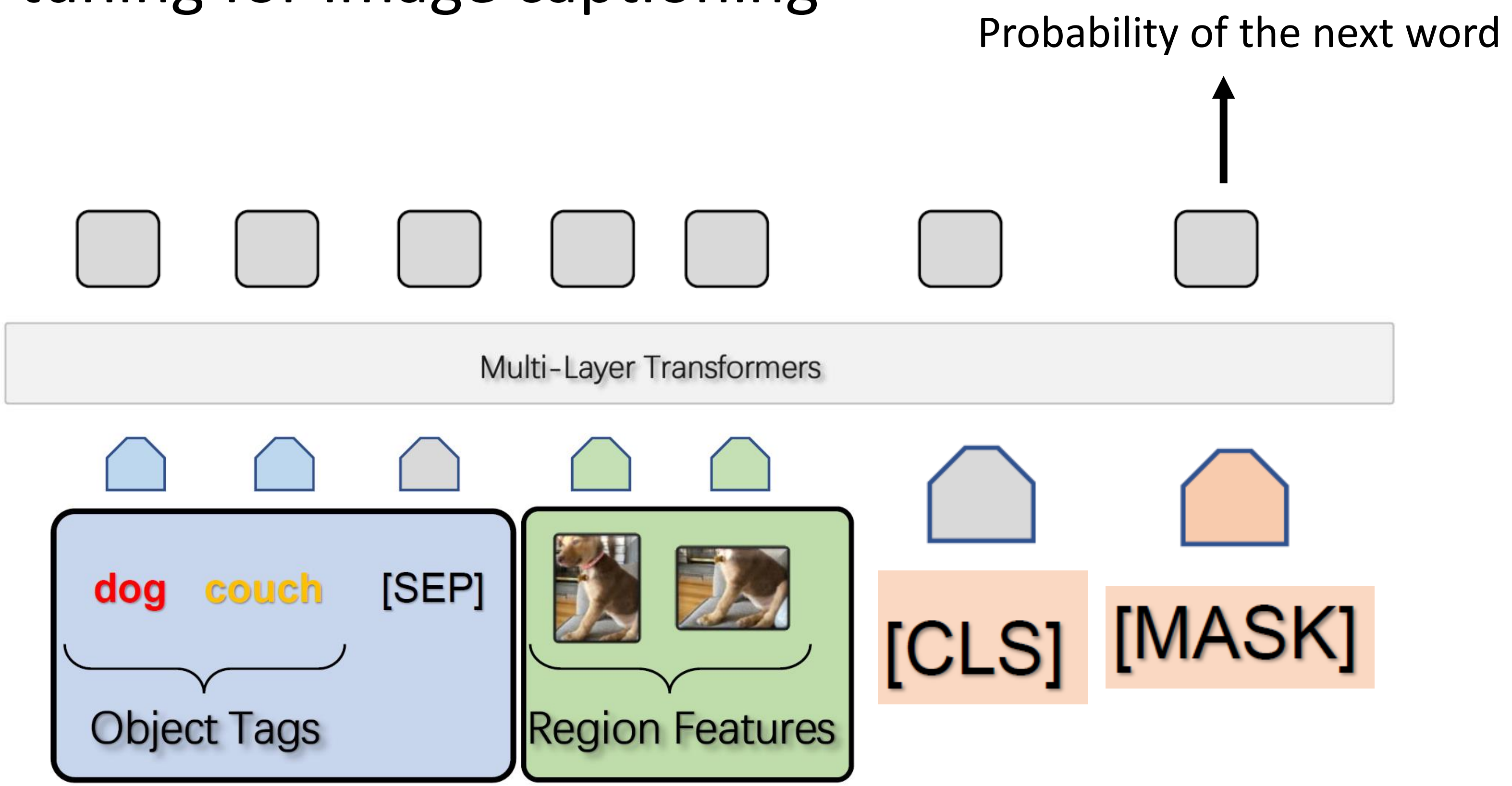
Classify "polluted" triplets with wrong tags



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

Oscar: Object-Semantics Aligned Pre-training

- Fine-tuning for image captioning

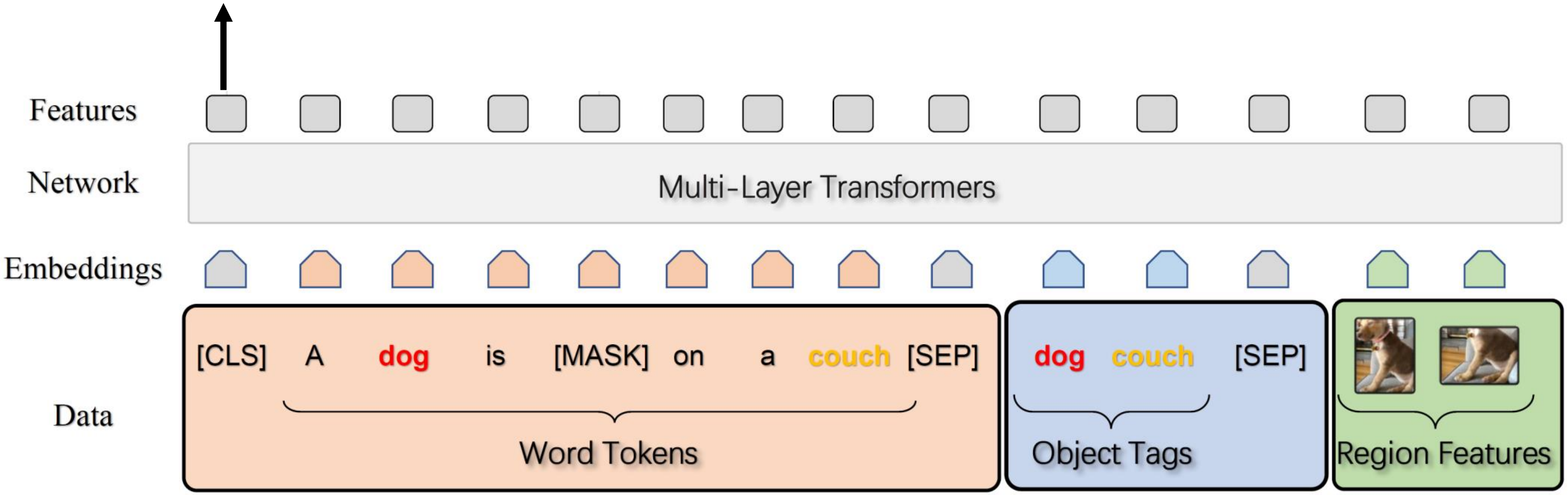


Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. Li et al., ECCV, 2020

Oscar: Object-Semantics Aligned Pre-training

- Fine-tuning for question answering

Classifier to answers (e.g., 3,129 answer set)

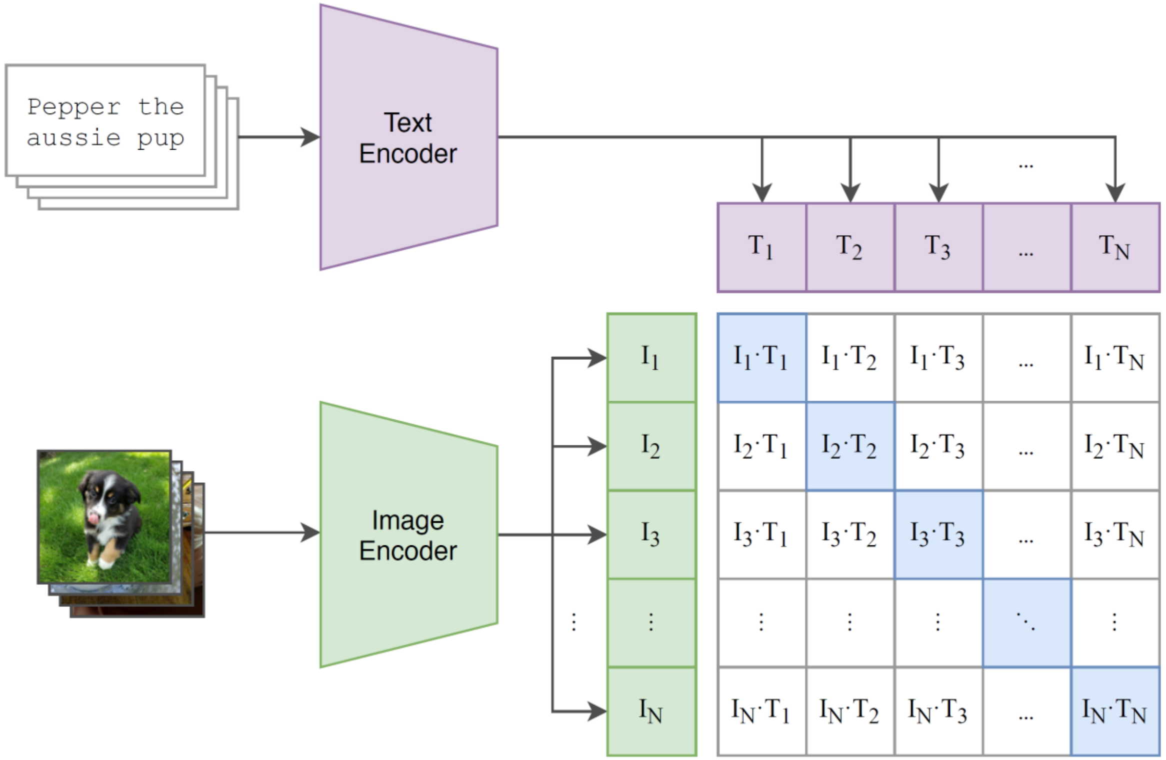


Replace with a Question

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

CLIP: Contrastive Language-Image Pre-Training

- Contrastive pre-training



- 400 million (image, text) pairs from Internet

Learning Transferable Visual Models From Natural Language Supervision. Radford, et al., 2021

CLIP: Contrastive Language-Image Pre-Training

- Contrastive pre-training

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Multi-class N-pair Loss

Softmax for multi-class classification

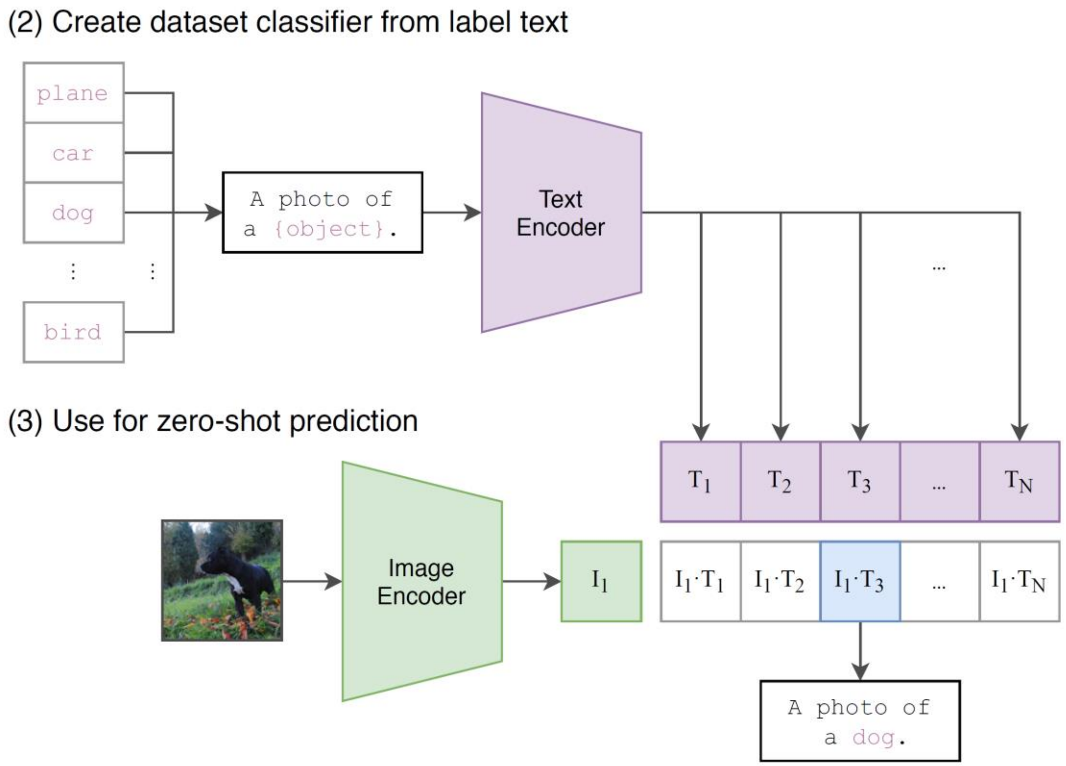
$$\mathcal{L}_{\text{N-pair}}(\mathbf{x}, \mathbf{x}^+, \{\mathbf{x}_i^-\}_{i=1}^{N-1}) = \log\left(1 + \sum_{i=1}^{N-1} \exp(f(\mathbf{x})^\top f(\mathbf{x}_i^-) - f(\mathbf{x})^\top f(\mathbf{x}^+))\right)$$

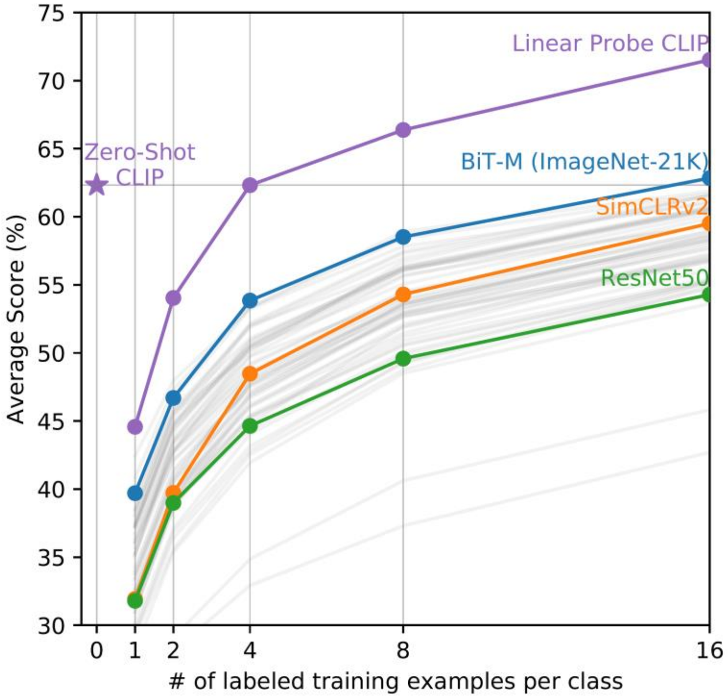
$$= -\log \frac{\exp(f(\mathbf{x})^\top f(\mathbf{x}^+))}{\exp(f(\mathbf{x})^\top f(\mathbf{x}^+)) + \sum_{i=1}^{N-1} \exp(f(\mathbf{x})^\top f(\mathbf{x}_i^-))}$$

Learning Transferable Visual Models From Natural Language Supervision. Radford, et al., 2021

CLIP: Contrastive Language-Image Pre-Training

- Zero-shot classification (no training on target datasets)





CLIP Linear Probe: logistic regression performed on CLIP encoded image features

Learning Transferable Visual Models From Natural Language Supervision. Radford, et al., 2021

Summary

- Vision + language tasks
  - Image captioning
  - Object/phase grounding
  - Visual question answering
  - Image-text retrieval

- Representation learning (Pre-training)
  - Learning image-text representations from large numbers (image, text) pairs
  - Fine-turning for downstream tasks

Further Reading

- Baby Talk: Understanding and Generating Image Descriptions, 2011 http://www.tamaraberg.com/papers/generation_cvpr11.pdf
- Deep Visual-Semantic Alignments for Generating Image Descriptions, 2015 https://arxiv.org/abs/1412.2306
- Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2015 https://arxiv.org/abs/1502.03044
- Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, 2018 https://arxiv.org/abs/1707.07998
- MDETR - Modulated Detection for End-to-End Multi-Modal Understanding, 2021 https://arxiv.org/abs/2104.12763
- VQA: Visual Question Answering, 2015 https://arxiv.org/abs/1505.00468
- Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, 2020 https://arxiv.org/abs/2004.06165
- Learning Transferable Visual Models From Natural Language Supervision, 2021 https://arxiv.org/abs/2103.00020