Self-supervised Learning for Vision-and-Language

Licheng Yu, Yen-Chun Chen, Linjie Li

Microsoft facebook

Nowadays Machine Learning

Nowadays Machine Learning

Datasets + Labels

Instructions:

- Describe all the important parts of the scene.
- Do not start the sentences with "There is".
- Do not describe unimportant details.
- Do not describe things that might have happened in the future or past.
- Do not describe what a person might say.
- Do not give people proper names.
- The sentence should contain at least 8 words.

Please describe the image:

Enter description here

prev next

- MS COCO’s Image Captioning:
  - 120,000 images
  - 5 sentences / image
  - 15 cents / sentence
  - +20% AWS processing fee

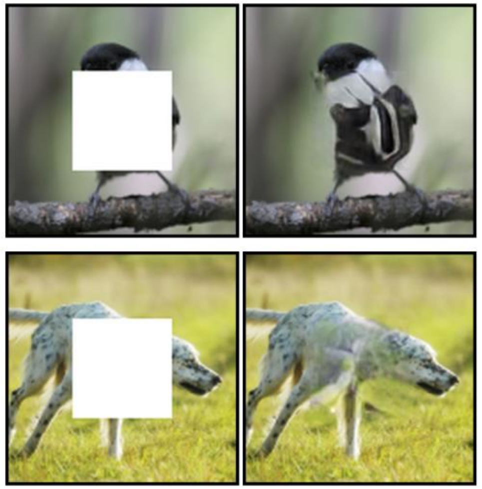
$108,000

Datasets ~~+ Labels~~: Self-Supervised Learning for Vision

Image Colorization

[Zhang et al. ECCV 2016]

Image Inpainting

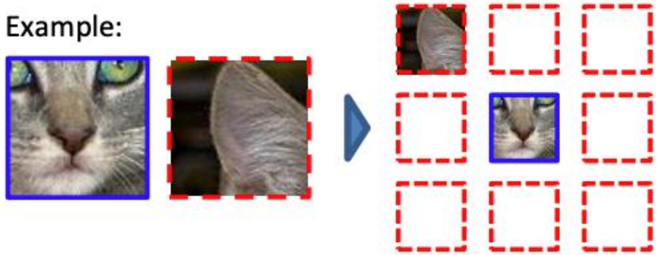


[Pathak et al. CVPR 2016]

Jigsaw puzzles

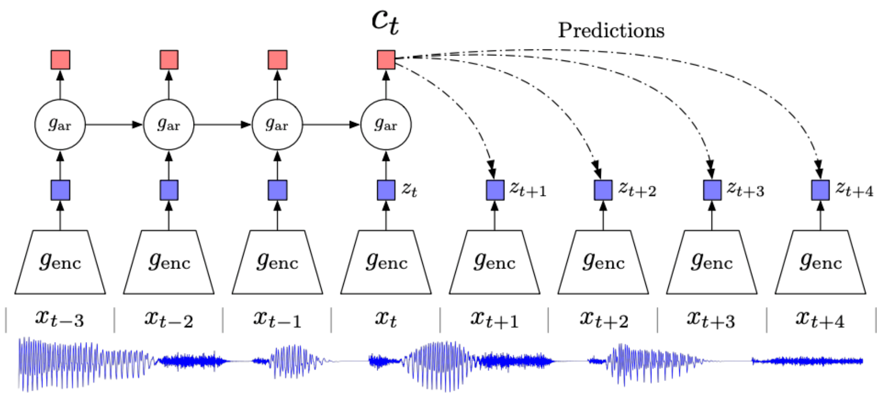
[Noroozi et al. ECCV 2016]

Relative Location Prediction

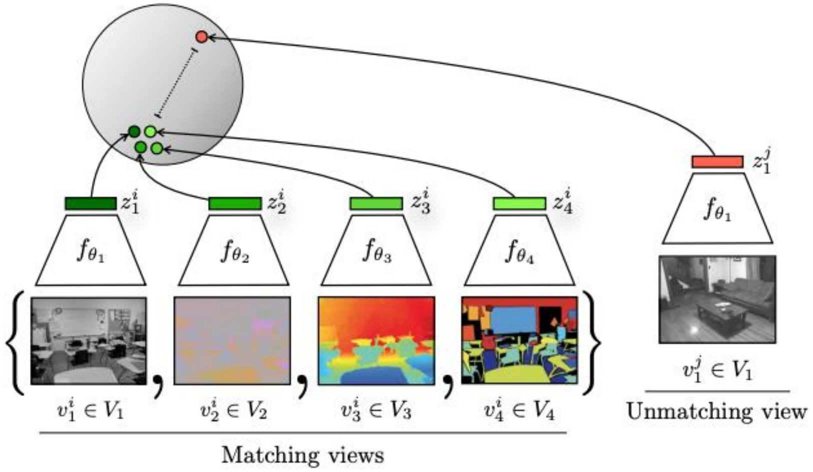


[Doersch et al. ICCV 2015]

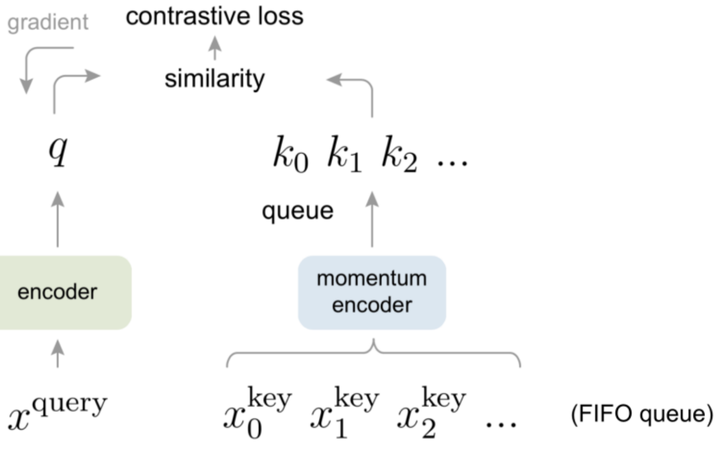
Datasets ~~+ Labels~~: Self-Supervised Learning for Vision



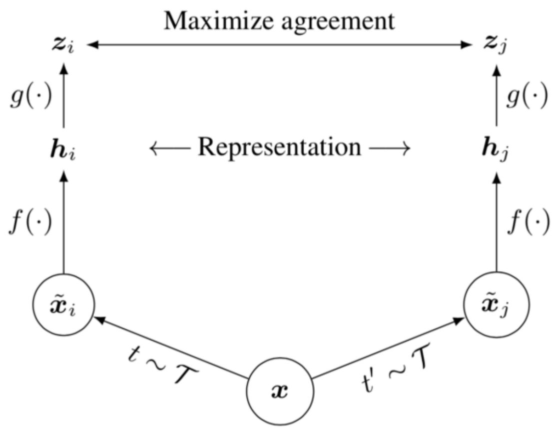
CPC; Ord et al, 2019



CMC; Tian et al, 2019

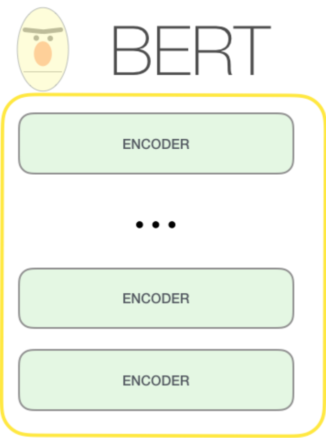


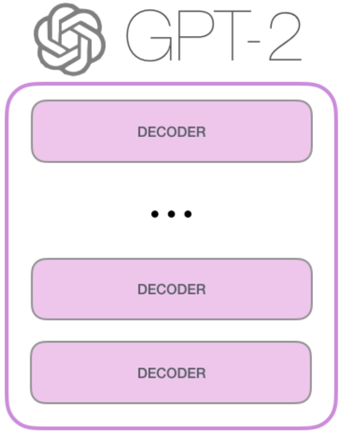
MOCO; He et al, 2019

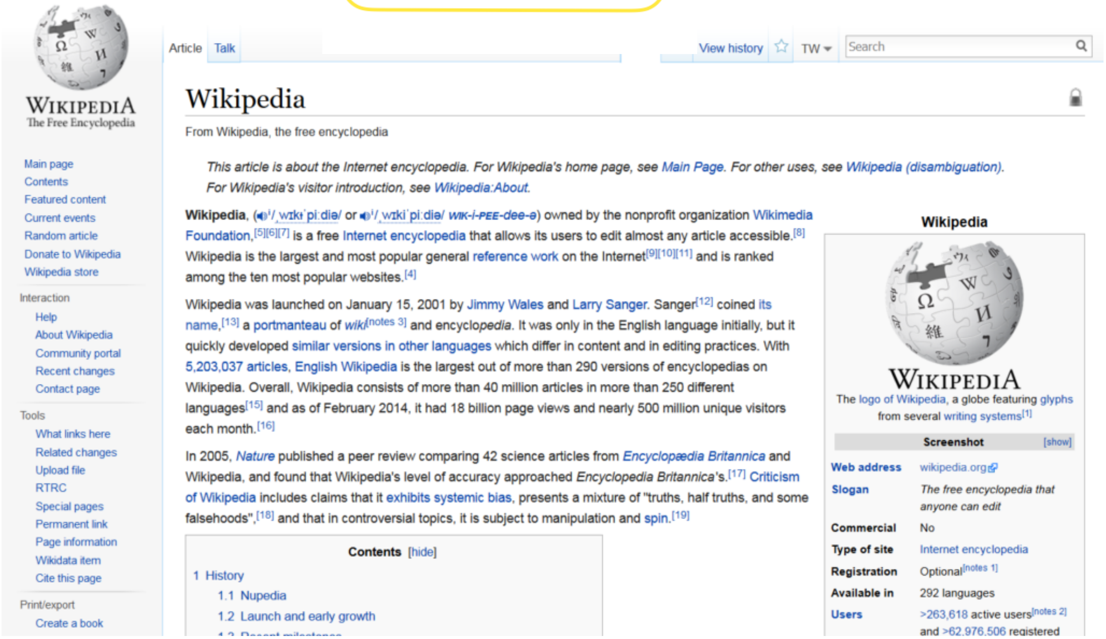


SimCLR; Chen et al, 2020

Datasets ~~+ Labels~~: Self-Supervised Learning for NLP



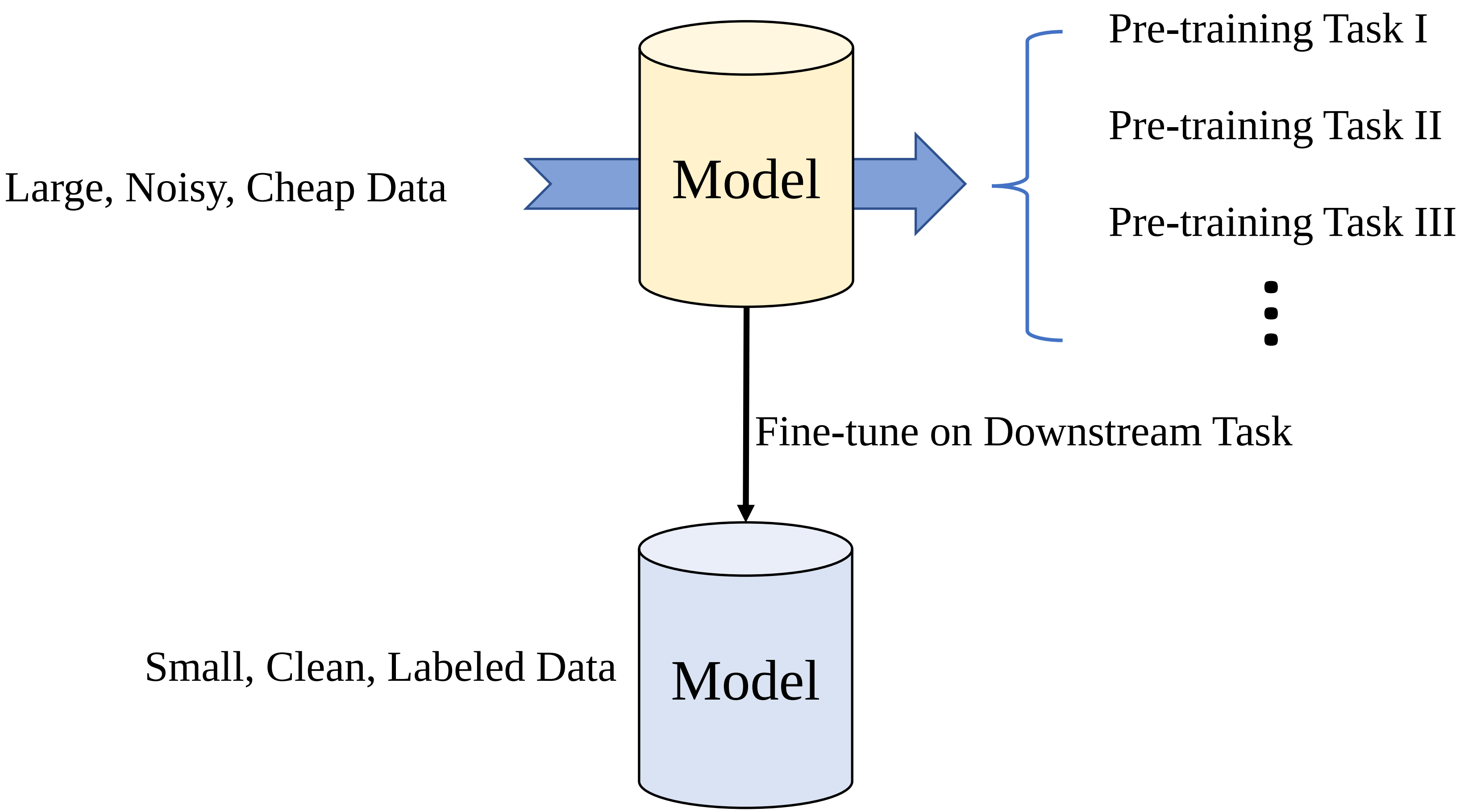




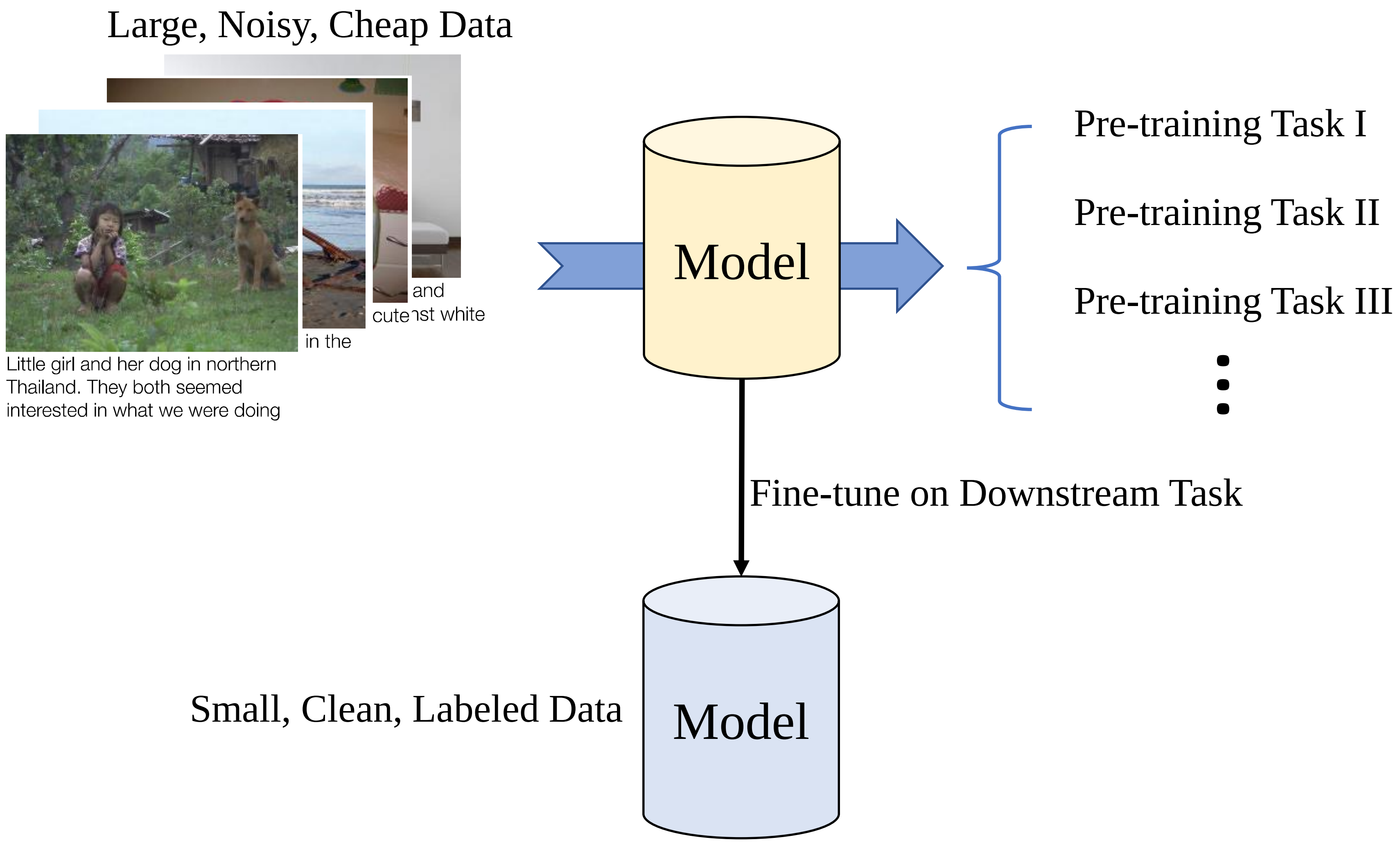
[Devlin et al. NAACL 2019]

[Radford et al. 2019]

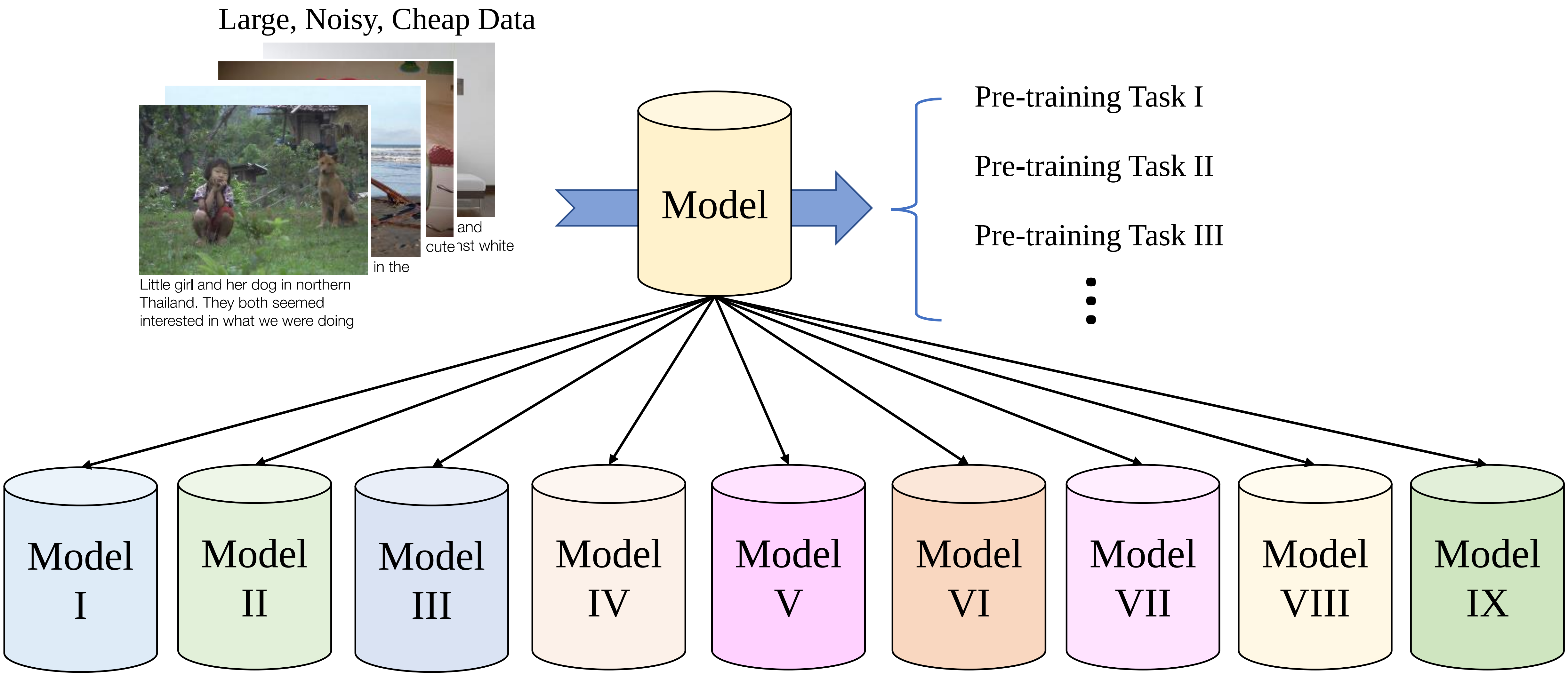
Pre-training + Finetuning

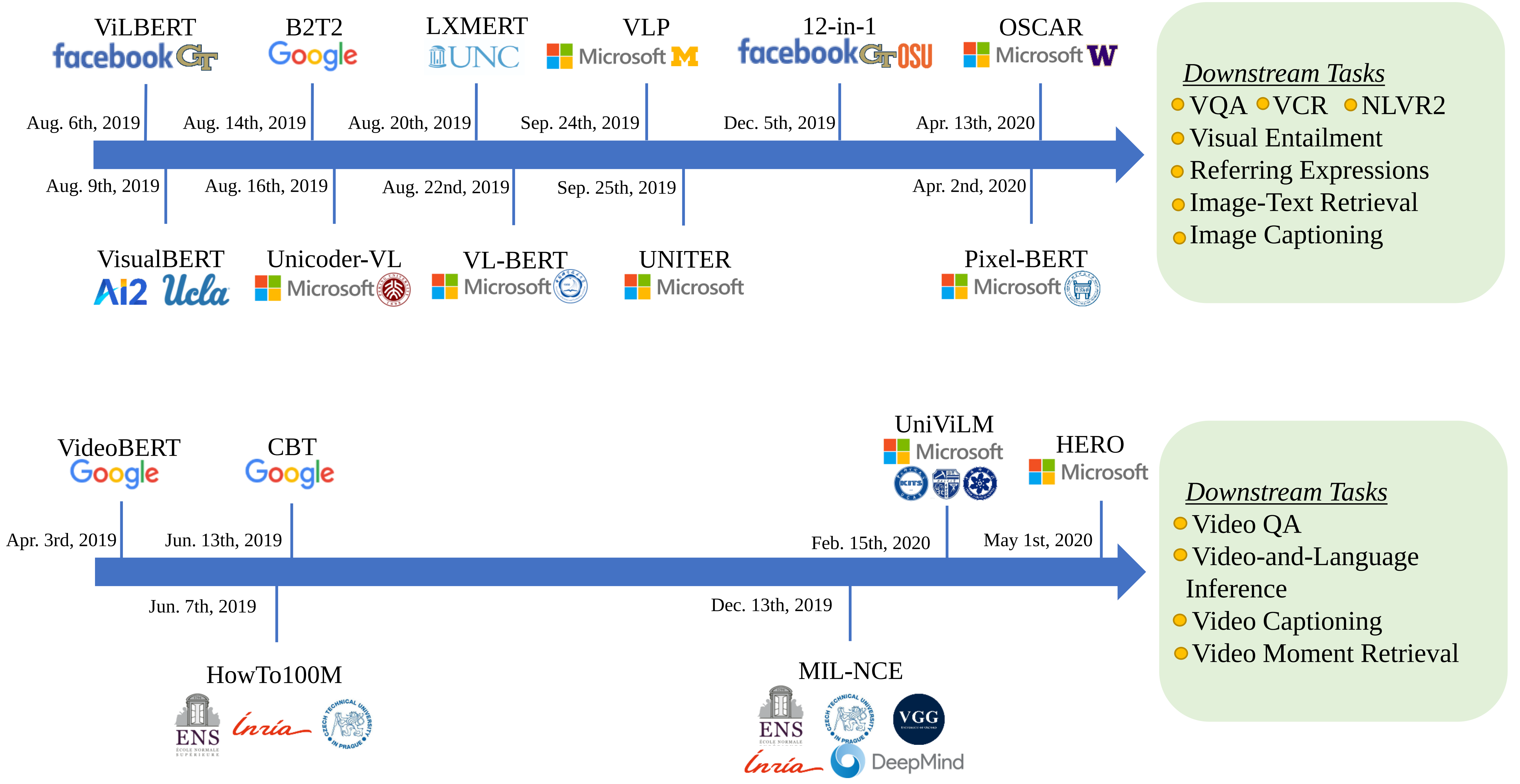


Two-Stage Training Pipeline

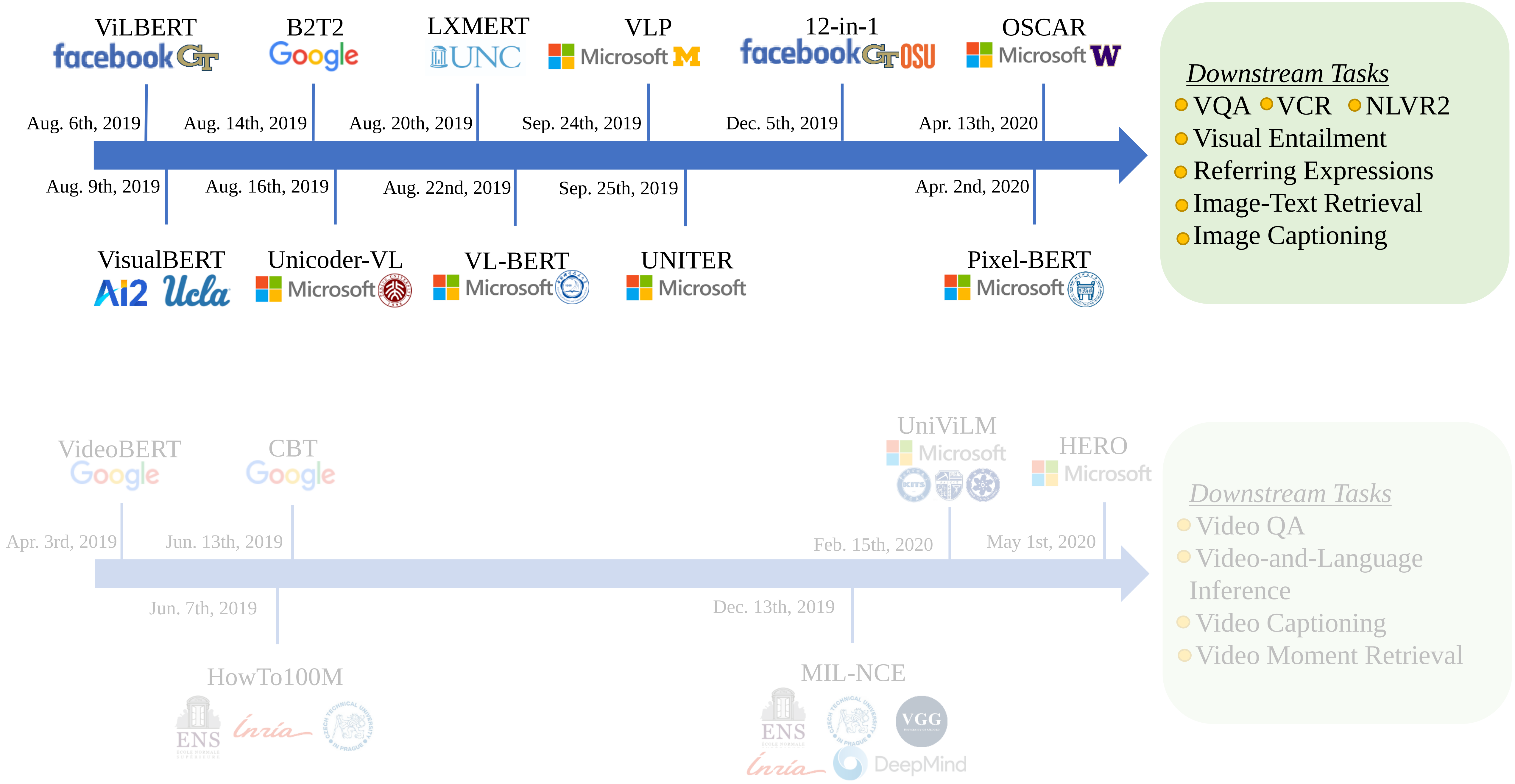


Generalization





ViLBERT
Aug. 6th, 2019
B2T2
Aug. 14th, 2019
LXMERT
Aug. 20th, 2019
VLP
Sep. 24th, 2019
12-in-1
Dec. 5th, 2019
OSCAR
Apr. 13th, 2020
Aug. 9th, 2019
VisualBERT
Aug. 16th, 2019
Unicoder-VL
Aug. 22nd, 2019
VL-BERT
Sep. 25th, 2019
UNITER
Apr. 2nd, 2020
Pixel-BERT
Downstream Tasks
VQA
VCR
NLVR2
Visual Entailment
Referring Expressions
Image-Text Retrieval
Image Captioning
VideoBERT
Apr. 3rd, 2019
CBT
Jun. 13th, 2019
UniViLM
Feb. 15th, 2020
HERO
May 1st, 2020
Jun. 7th, 2019
HowTo100M
Dec. 13th, 2019
MIL-NCE
Downstream Tasks
Video QA
Video-and-Language Inference
Video Captioning
Video Moment Retrieval


ViLBERT
Aug. 6th, 2019
B2T2
Aug. 14th, 2019
LXMERT
Aug. 20th, 2019
VLP
Sep. 24th, 2019
12-in-1
Dec. 5th, 2019
OSCAR
Apr. 13th, 2020
Aug. 9th, 2019
VisualBERT
Aug. 16th, 2019
Unicoder-VL
Aug. 22nd, 2019
VL-BERT
Sep. 25th, 2019
UNITER
Apr. 2nd, 2020
Pixel-BERT
Downstream Tasks
VQA
VCR
NLVR2
Visual Entailment
Referring Expressions
Image-Text Retrieval
Image Captioning
VideoBERT
Apr. 3rd, 2019
CBT
Jun. 13th, 2019
UniViLM
Feb. 15th, 2020
HERO
May 1st, 2020
Jun. 7th, 2019
HowTo100M
Dec. 13th, 2019
MIL-NCE
Downstream Tasks
Video QA
Video-and-Language Inference
Video Captioning
Video Moment Retrieval

Pre-training Data

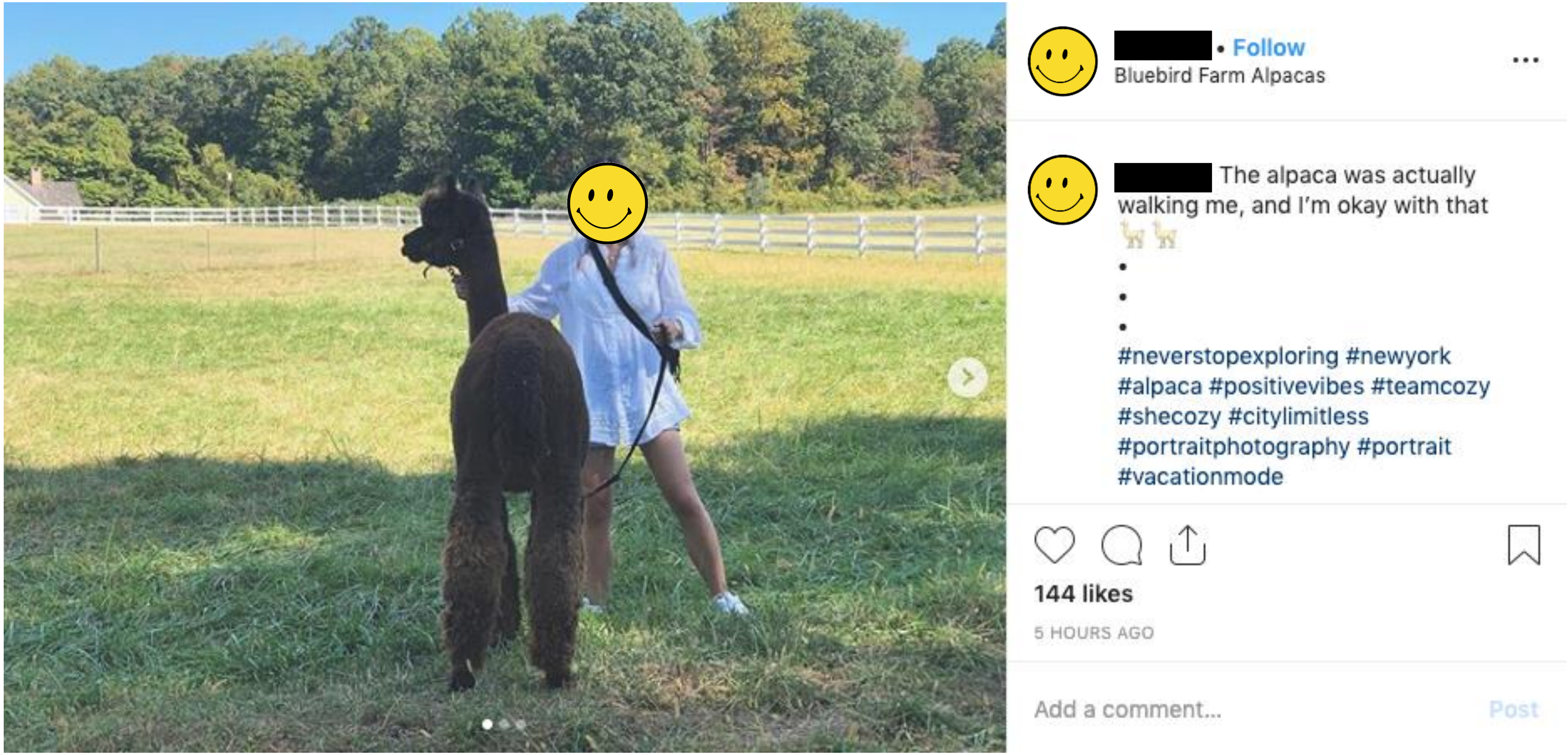
Pre-training Vision+Language Data

( , ‘man with his dog on a couch)

Free Data for Vision + Language

Free Data for Vision + Language

Free Data for Vision + Language



Common Pre-training Data for Vision + Language

	In-domain		Out-of-domain	
Split	COCO Captions	VG Dense Captions	Conceptual Captions	SBU Captions
train	533K (106K)	5.06M (101K)	3.0M (3.0M)	990K (990K)
val	25K (5K)	106K (2.1K)	14K (14K)	10K (10K)

Conceptual Caption

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.

SBU Caption

Little girl and her dog in northern Thailand. They both seemed interested in what we were doing

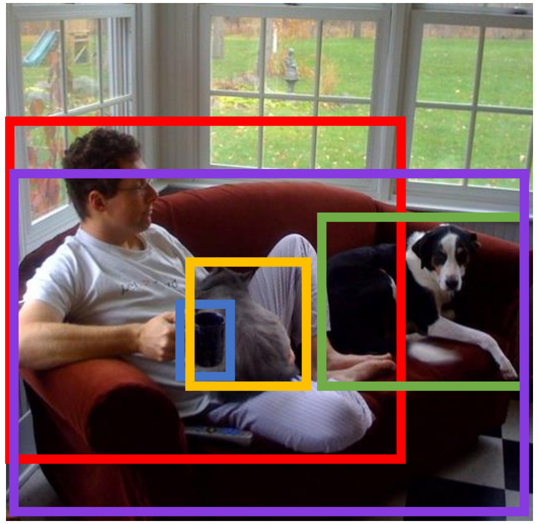
https://github.com/lichengunc/pretrain-vl-data

Feature Representations for Vision and Language

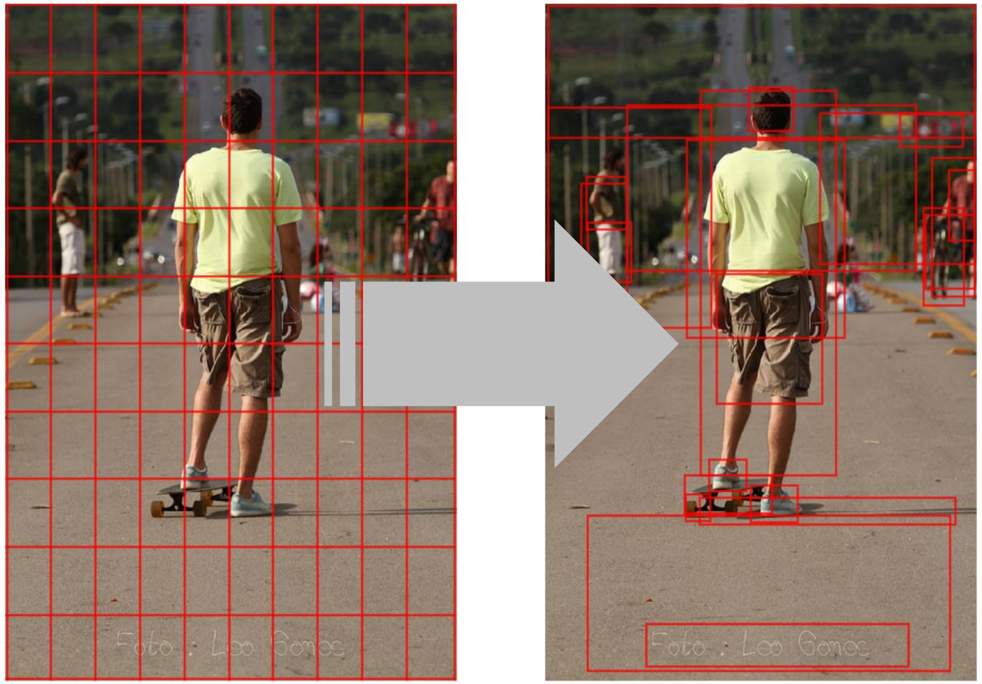
Visual and Language Features

( , "man with his dog on a couch")

Visual and Language Features

( , ‘man’ ‘with’ ‘his’ ‘dog’ ‘on’ ‘a’ ‘couch’)

Visual Features

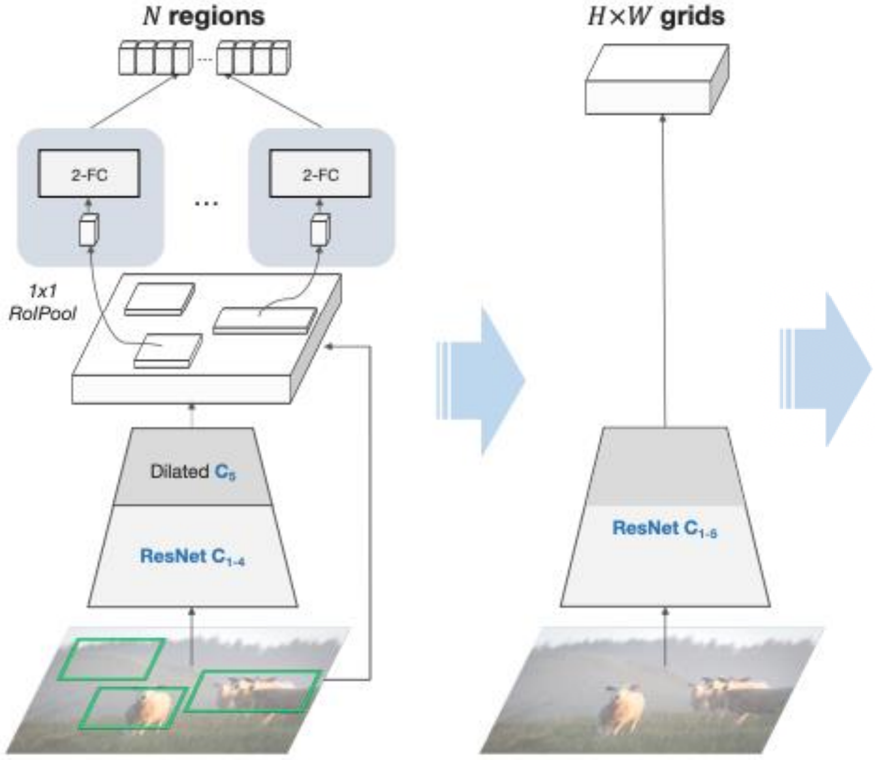


Pre-2017: grid feature maps

[Ren et al, NeurIPS 2015]

Post-2017: region features

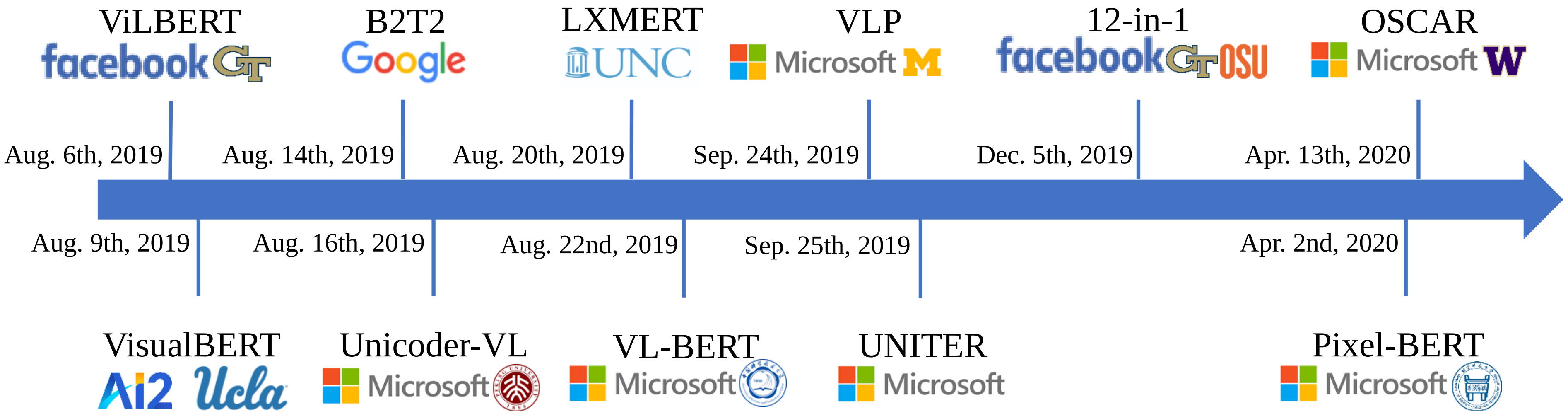
[Anderson et al, CVPR 2018]



Winner of
VQA Challenge 2020

[Jiang et al, CVPR 2020]

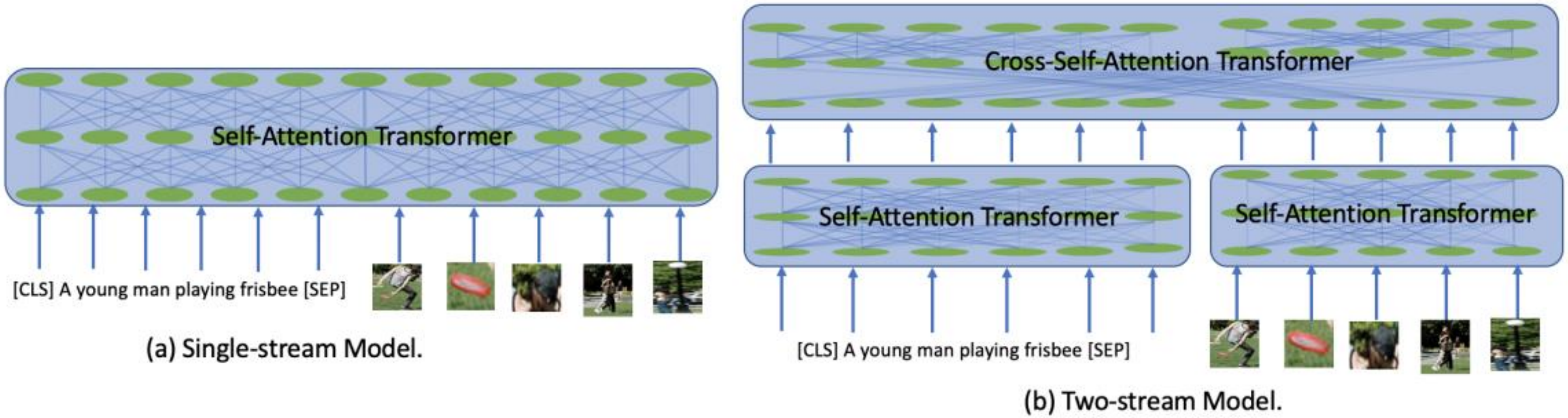
Model Architecture



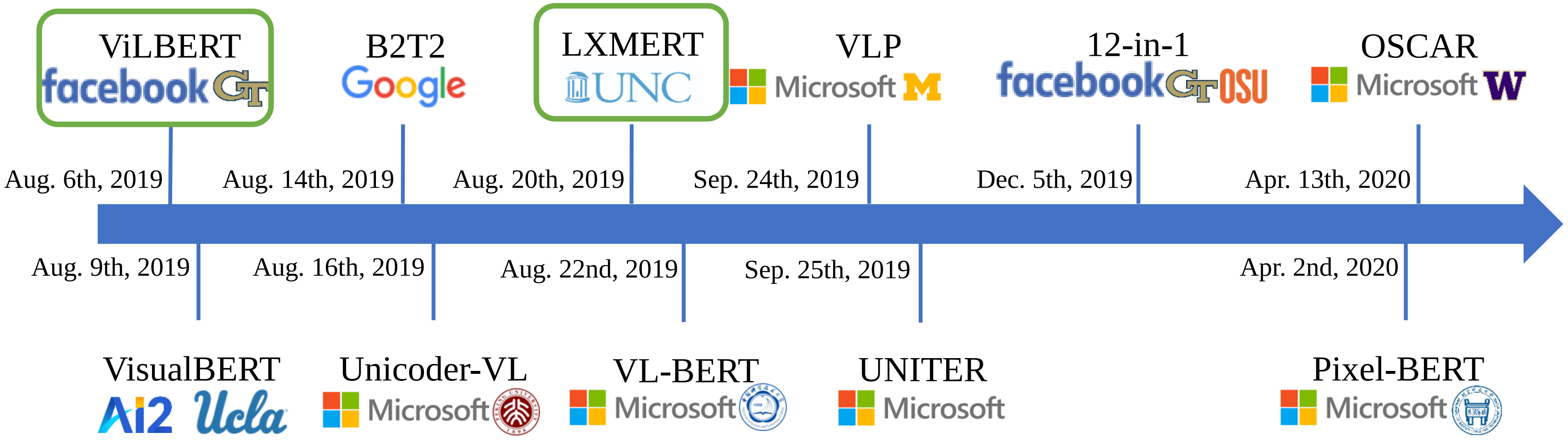
__Downstream Tasks__

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

Model Architecture:



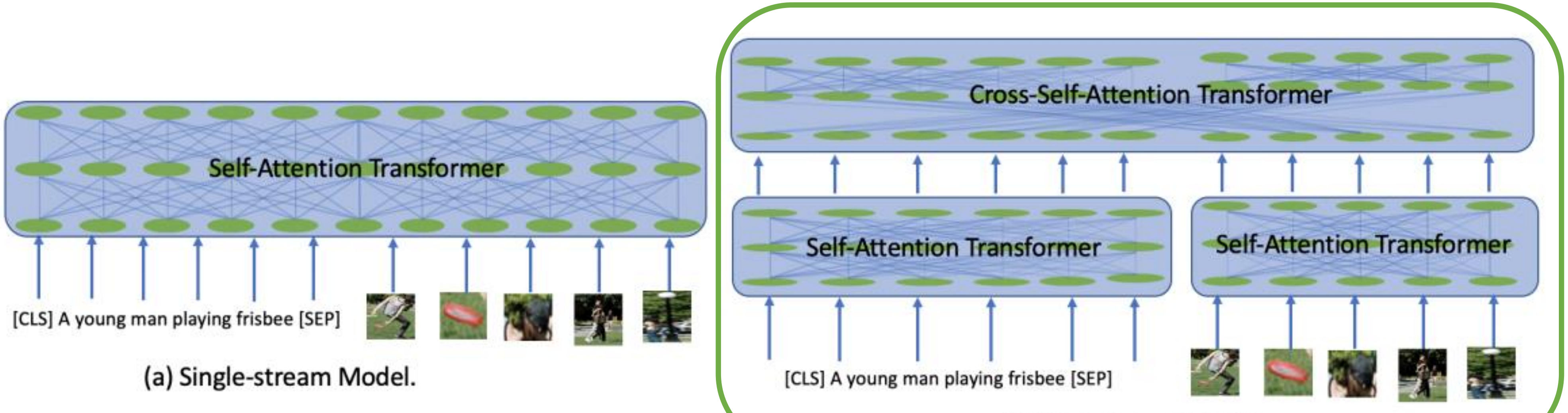
[Behand the Scene; Cao et al 2020]



Downstream Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

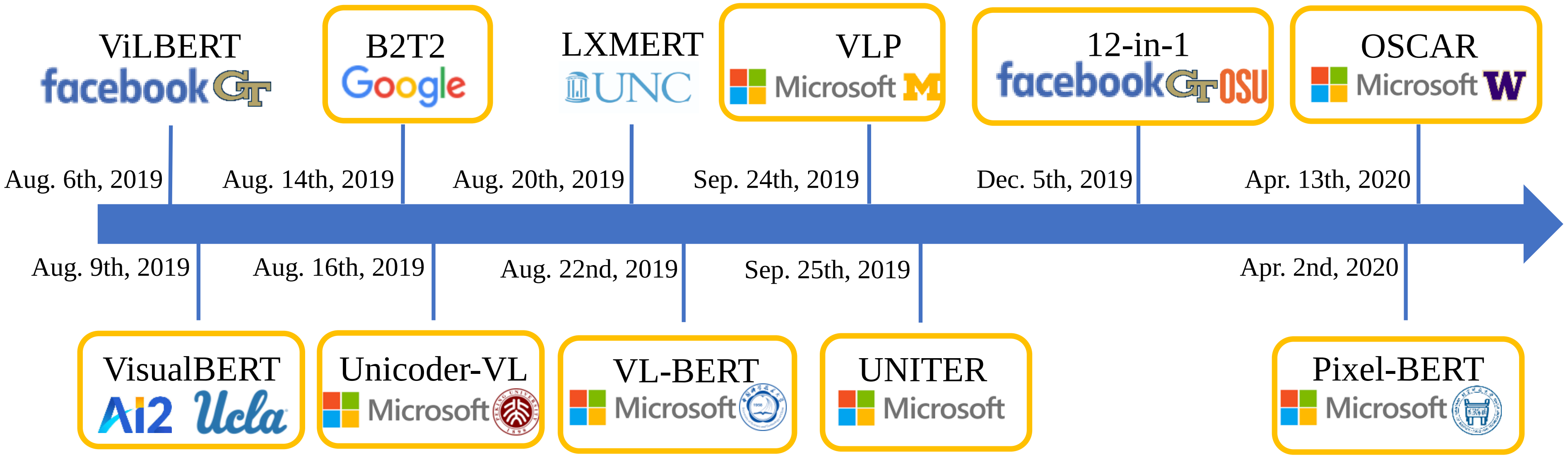
Model Architecture:



(a) Single-stream Model.

(b) Two-stream Model.

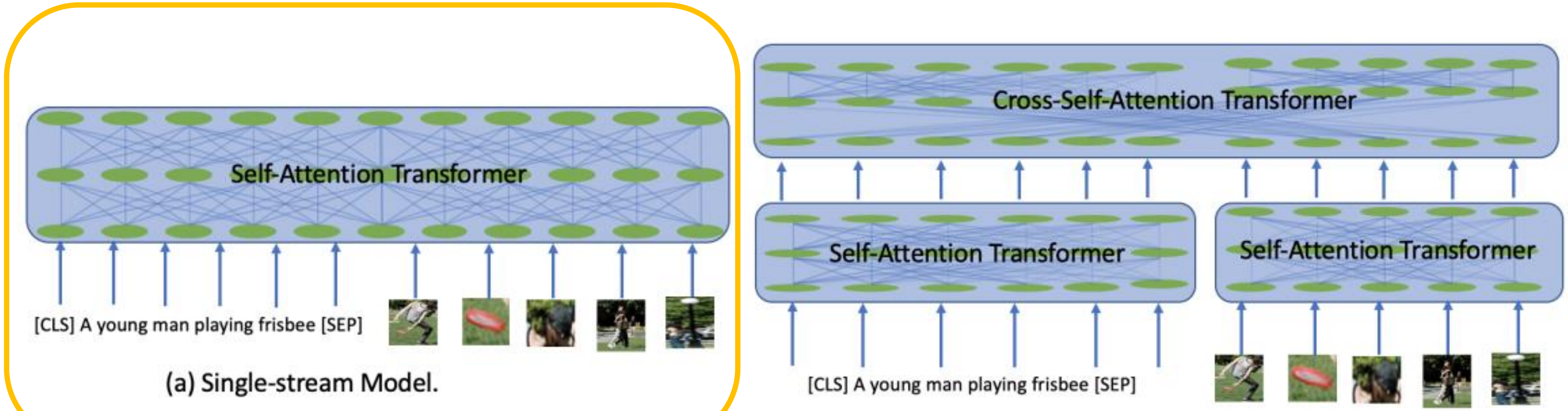
[Behand the Scene; Cao et al 2020]



Downstream Tasks

- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

Model Architecture:

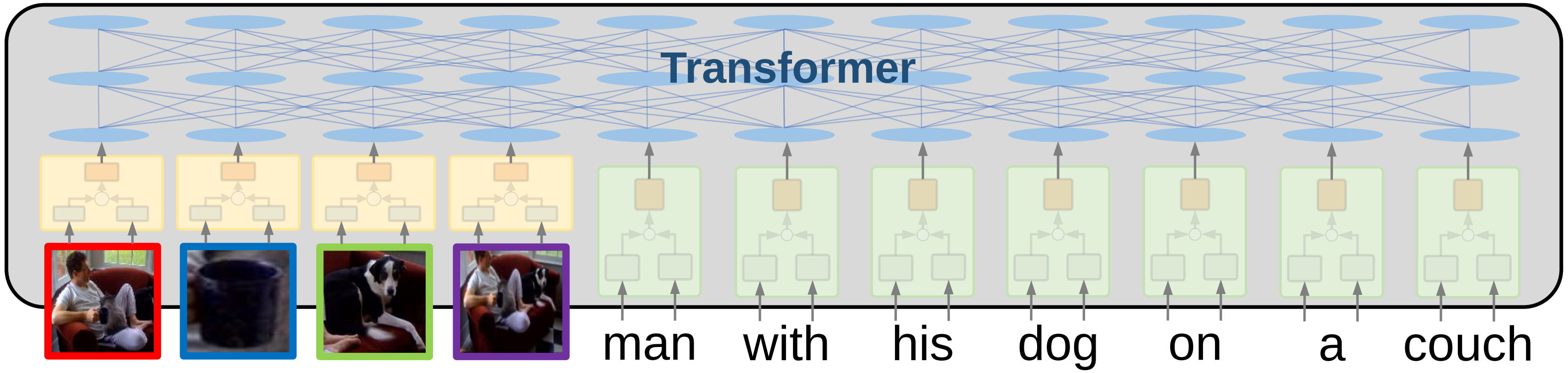


(a) Single-stream Model.

(b) Two-stream Model.

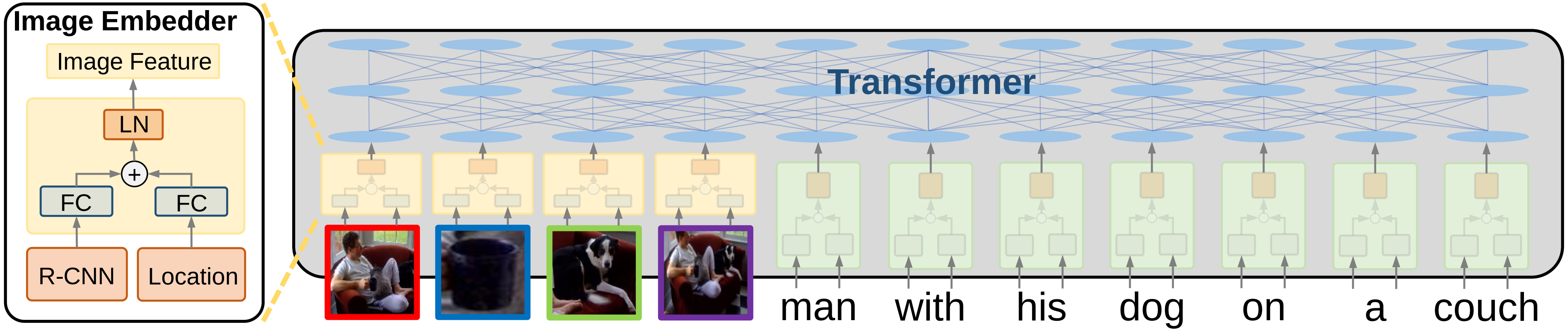
[Behand the Scene; Cao et al 2020]

Single-Stream Architecture



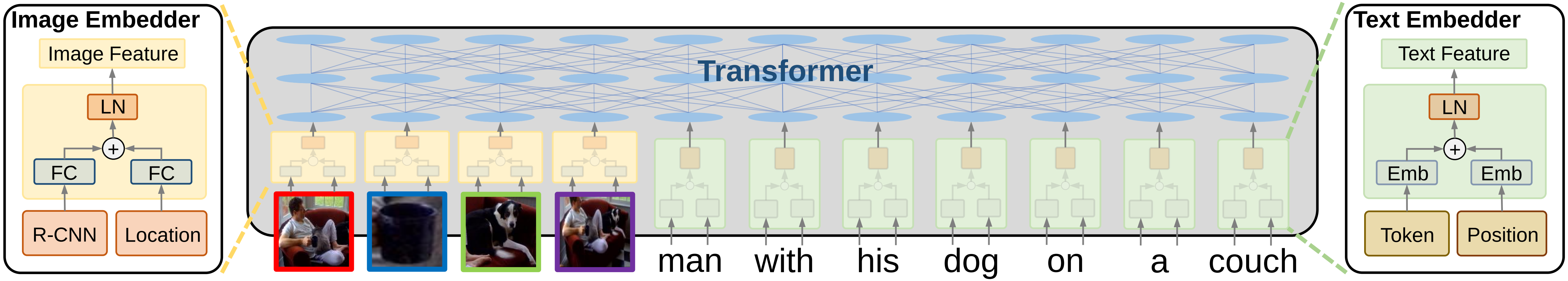
[UNITER; Chen et al2019]

Single-Stream Architecture



[UNITER; Chen et al2019]

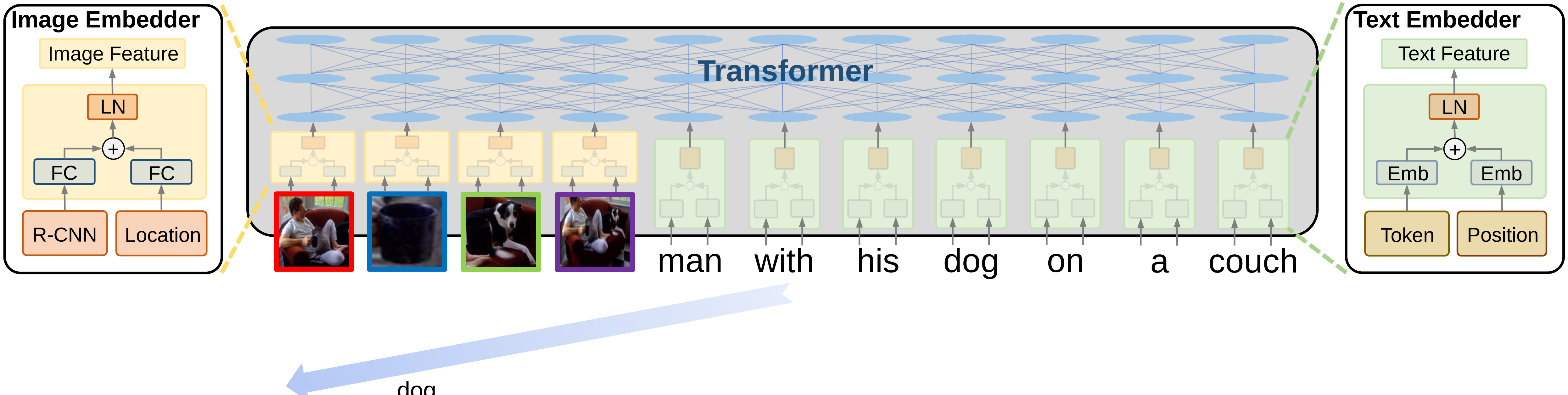
Single-Stream Architecture



[UNITER; Chen et al2019]

Pre-training Tasks

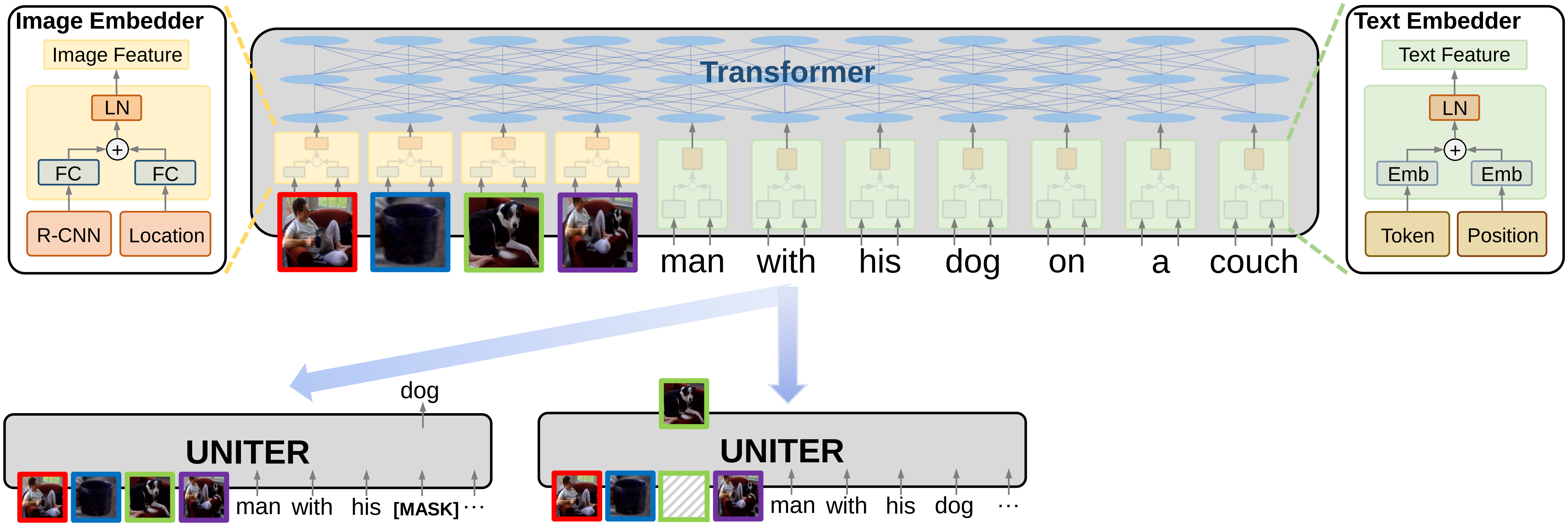
Pretraining Tasks



Masked Language Modeling (MLM)

[UNITER; Chen et al2019]

Pretraining Tasks

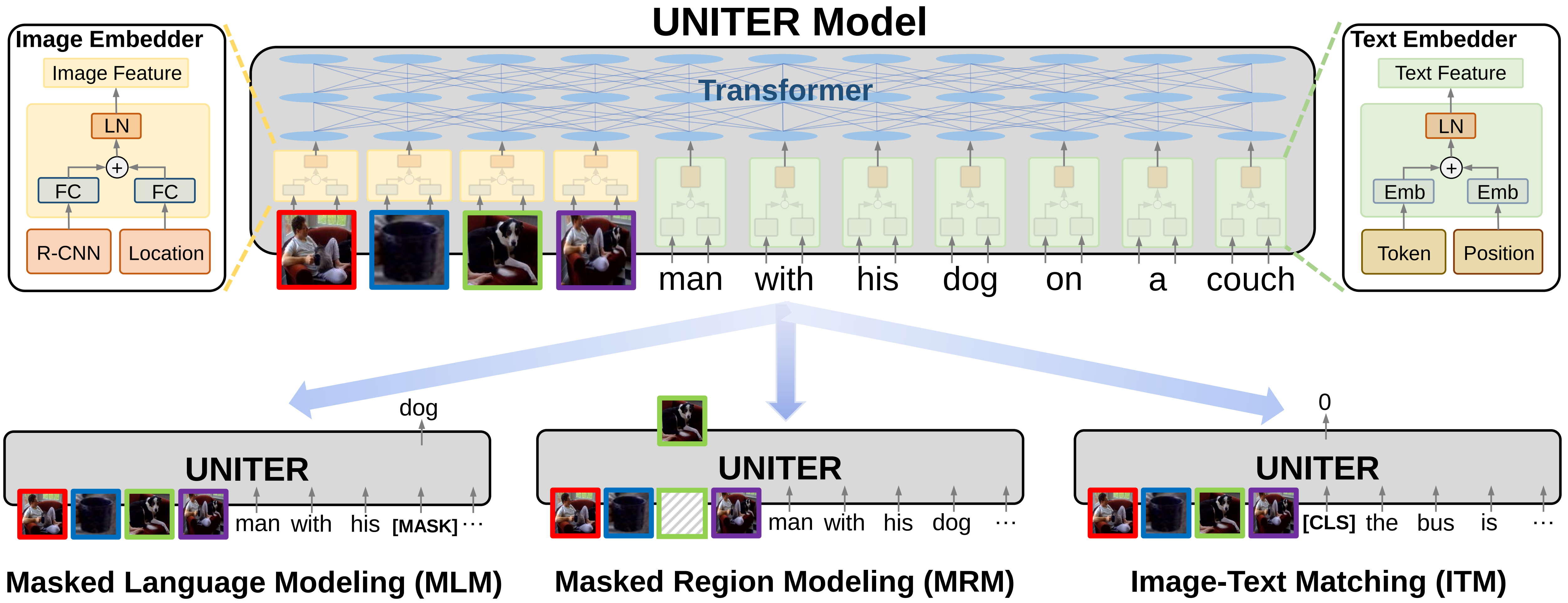


Masked Language Modeling (MLM)

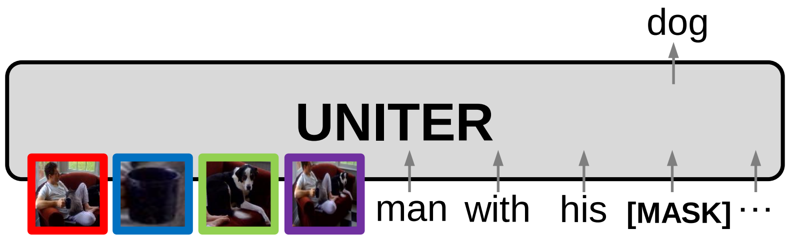
Masked Region Modeling (MRM)

[UNITER; Chen et al2019]

Pretraining Tasks



Pretraining Tasks



Masked Language Modeling (MLM)

Image Regions: $\mathbf{v} = \{v_1, ..., v_K\}$

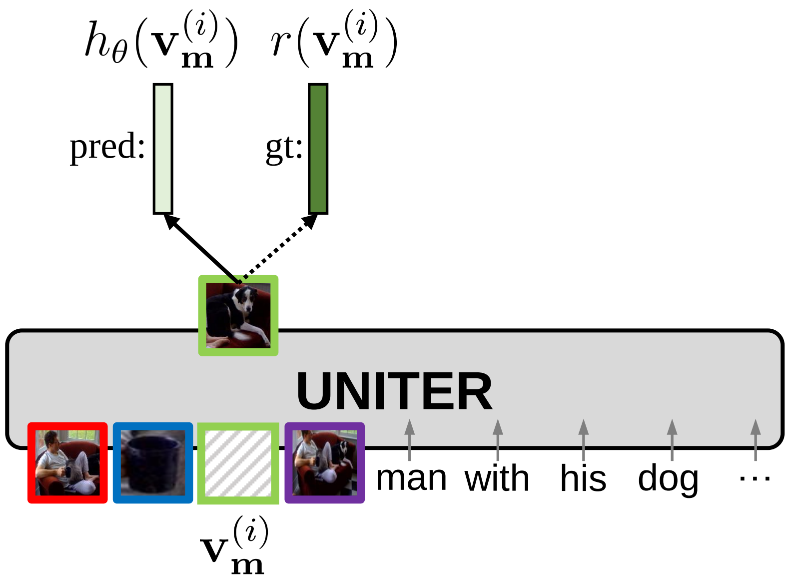
Sentence Tokens: $\mathbf{w} = \{w_1, ..., w_T\}$

Masking Indices: $\mathbf{m} \in \mathbb{N}^M$

Loss Function of Masked Language Modeling (MLM):

$$\mathcal{L}_{\text{MLM}}(\theta) = -E_{(\mathbf{w},\mathbf{v})\sim D} \log P_\theta(\mathbf{w_m}|\mathbf{w}_{\backslash\mathbf{m}}, \mathbf{v}).$$

Pretraining Tasks



Masked Region Modeling (MRM)

Image Regions: $\mathbf{v} = \{v_1, ..., v_K\}$

Sentence Tokens: $\mathbf{w} = \{w_1, ..., w_T\}$

Masking Indices: $\mathbf{m} \in \mathbb{N}^M$

Loss Function of Masked Region Modeling:

$$\mathcal{L}_{\text{MRM}}(\theta) = E_{(\mathbf{w},\mathbf{v})\sim D} f_\theta(\mathbf{v}_\mathbf{m}|\mathbf{v}_{\setminus\mathbf{m}}, \mathbf{w}).$$

1) Objective of **Masked Region Feature Regression (MRFR)**

$$f_\theta(\mathbf{v}_\mathbf{m}|\mathbf{v}_{\setminus\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} \|h_\theta(\mathbf{v}_\mathbf{m}^{(i)}) - r(\mathbf{v}_\mathbf{m}^{(i)})\|_2^2$$

Pretraining Tasks

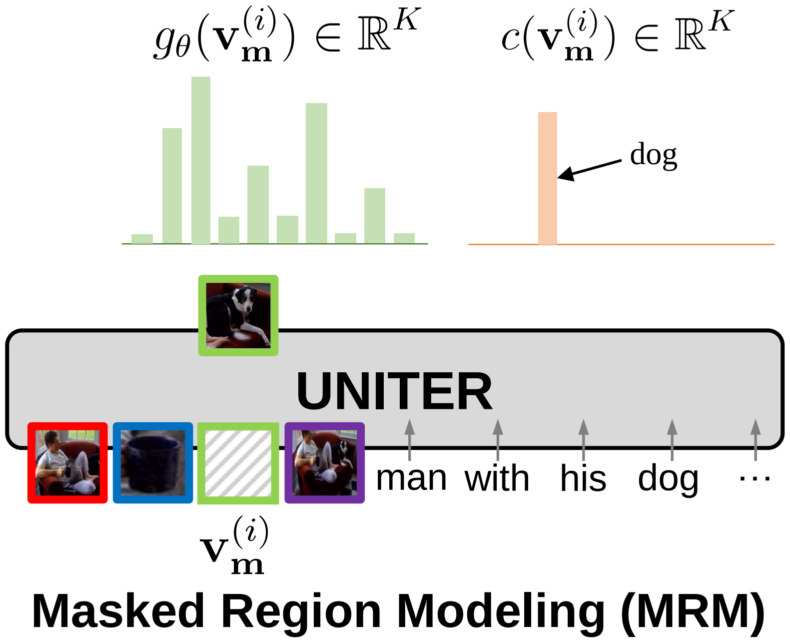


Image Regions: $\mathbf{v} = \{v_1, \ldots, v_K\}$

Sentence Tokens: $\mathbf{w} = \{w_1, \ldots, w_T\}$

Masking Indices: $\mathbf{m} \in \mathbb{N}^M$

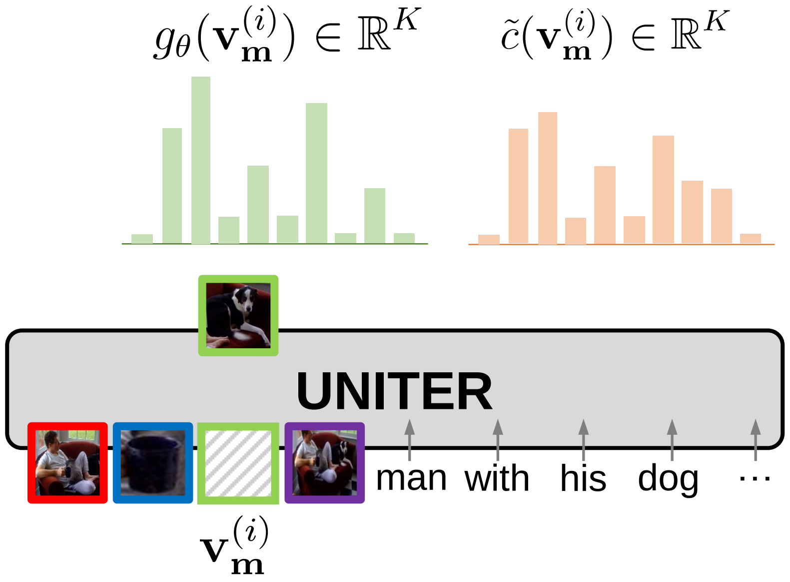
Loss Function of <u>Masked Region Modeling</u>:

$$\mathcal{L}_{\mathrm{MRM}}(\theta) = E_{(\mathbf{w},\mathbf{v})\sim D} f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash \mathbf{m}}, \mathbf{w}).$$

2) Objective of <u>**Masked Region Classification**</u> **(MRC)**

$$f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash \mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} \mathrm{CE}(c(\mathbf{v}_\mathbf{m}^{(i)}), g_\theta(\mathbf{v}_\mathbf{m}^{(i)}))$$

Pretraining Tasks



Masked Region Modeling (MRM)

Image Regions: $\mathbf{v} = \{v_1, ..., v_K\}$

Sentence Tokens: $\mathbf{w} = \{w_1, ..., w_T\}$

Masking Indices: $\mathbf{m} \in \mathbb{N}^M$

Loss Function of Masked Region Modeling:

$$\mathcal{L}_{\text{MRM}}(\theta) = E_{(\mathbf{w},\mathbf{v})\sim D} f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}).$$

3) Objective of **Masked Region Classification – KL Divergence (MRC-kl)**

$$f_\theta(\mathbf{v_m}|\mathbf{v}_{\backslash\mathbf{m}}, \mathbf{w}) = \sum_{i=1}^{M} D_{KL}(\tilde{c}(\mathbf{v}_\mathbf{m}^{(i)})||g_\theta(\mathbf{v}_\mathbf{m}^{(i)}))$$

Pretraining Tasks

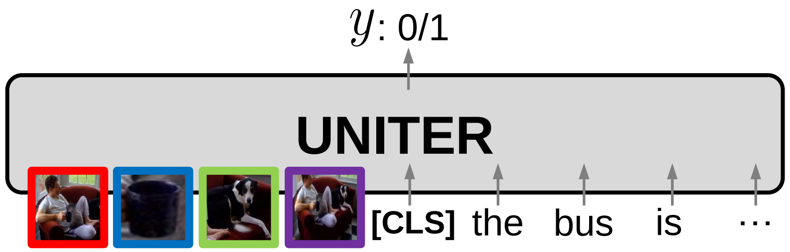


Image-Text Matching (ITM)

Image Regions: $\mathbf{v} = \{v_1, \ldots, v_K\}$

Sentence Tokens: $\mathbf{w} = \{w_1, \ldots, w_T\}$

Loss Function of **Image-Text Matching (ITM)**

$$\mathcal{L}_{\mathrm{ITM}}(\theta) = -E_{(\mathbf{w},\mathbf{v})\sim D}[y \log s_\theta(\mathbf{w},\mathbf{v}) + (1-y)\log(1-s_\theta(\mathbf{w},\mathbf{v}))]).$$

Pretraining Tasks

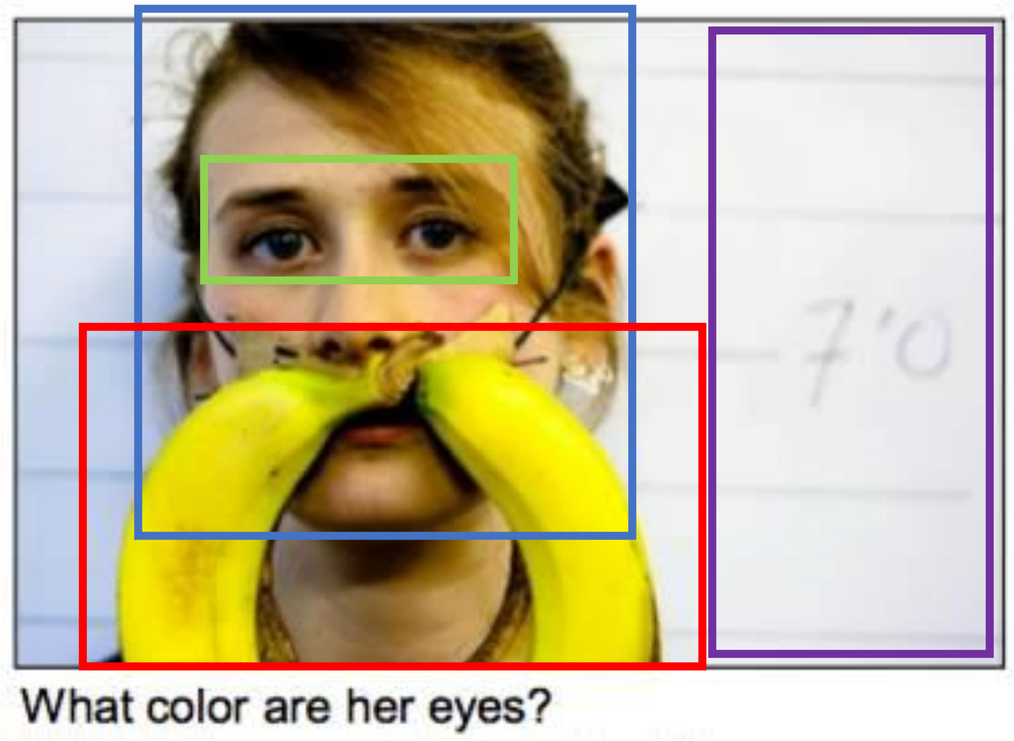
- UNITER: Word-Region Alignment
- VLP: Left-to-Right Language Modeling
- 12-in-1: Multi-task Learning
- LXMERT: Multi-task Learning
- OSCAR: Multi-View Alignment (tokens, tags, regions)
- …

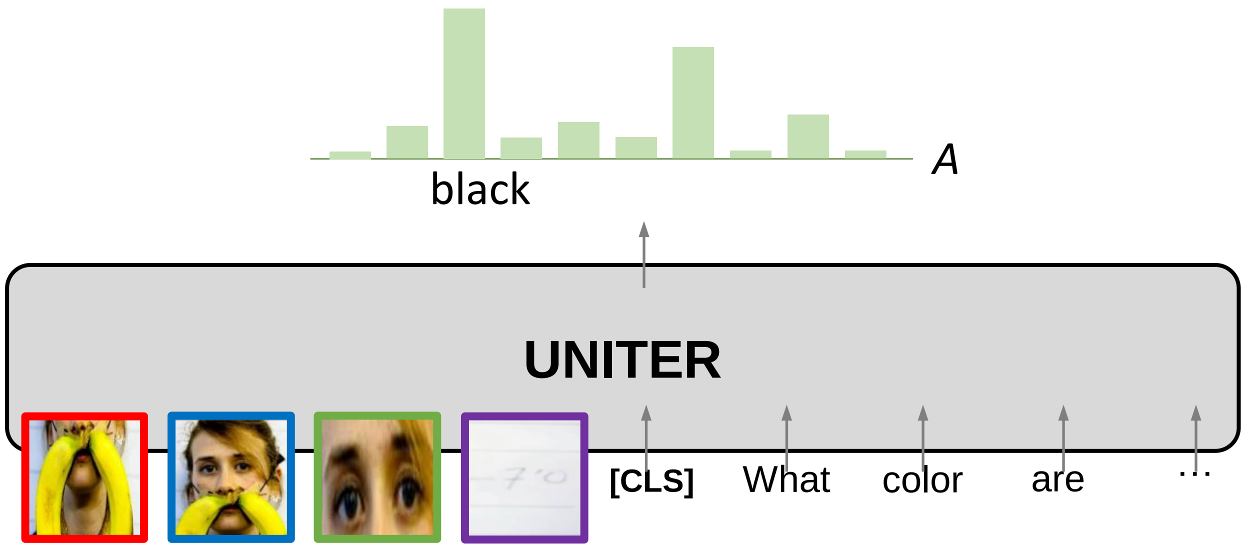
Downstream Tasks

Downstream Task 1: Visual Question Answering

[Antol et al., ICCV 2015]

Downstream Task 1: Visual Question Answering





Downstream Task 2: Visual Entailment

+

- *Two woman are holding packages.*
- *The sisters are hugging goodbye while holding to go packages after just eating lunch.*
- *The men are fighting outside a deli.*

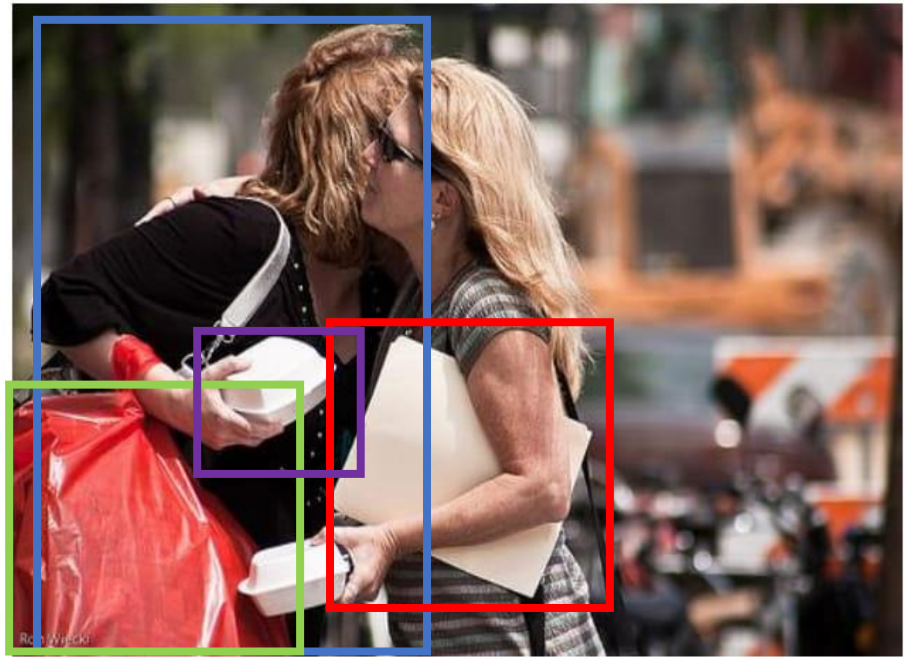
=

- *Entailment*
- *Neutral*
- *Contradiction*

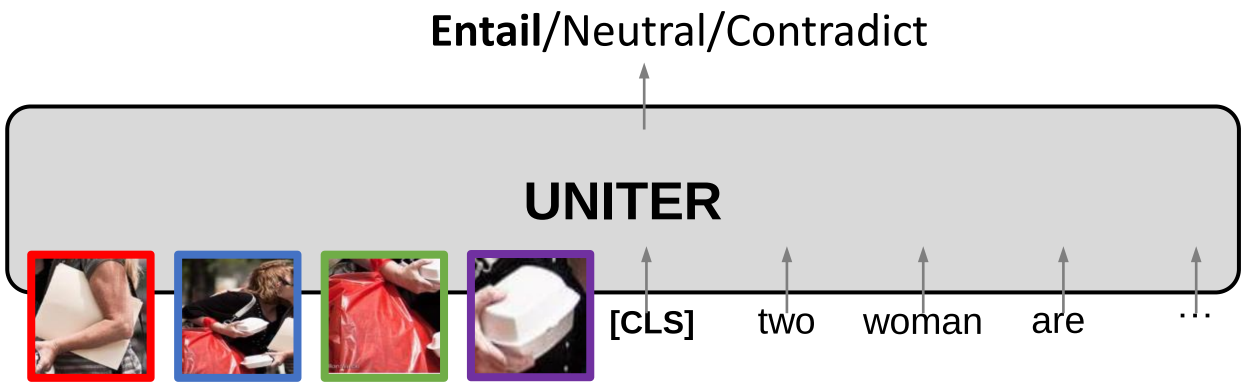
Premise *Hypothesis* *Answer*

[Xie et al., 2019]

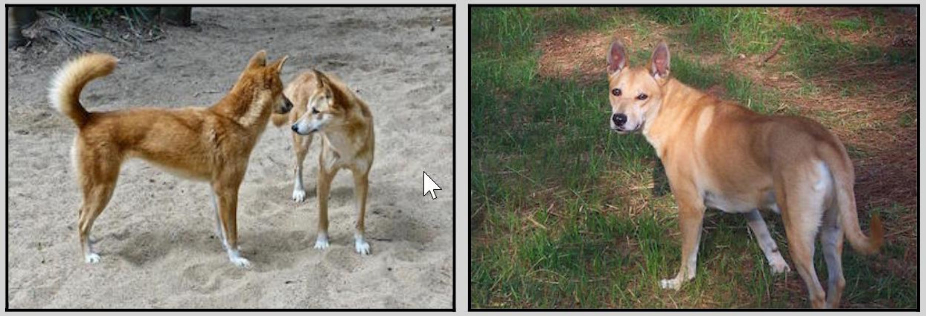
Downstream Task 2: Visual Entailment



Two woman are holding packages.



Downstream Task 3: Natural Language for Visual Reasoning



The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

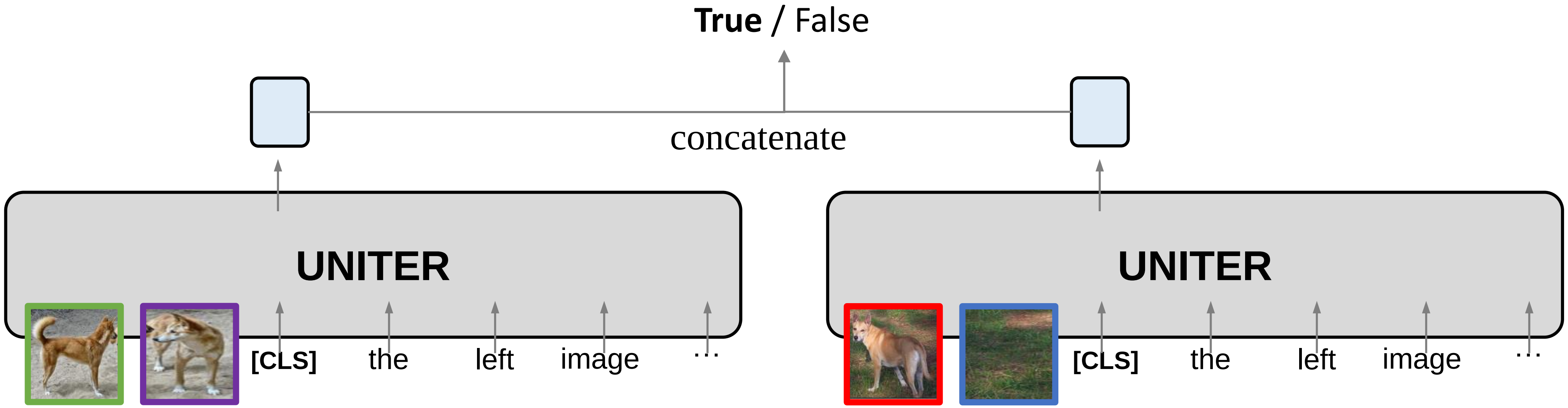
true

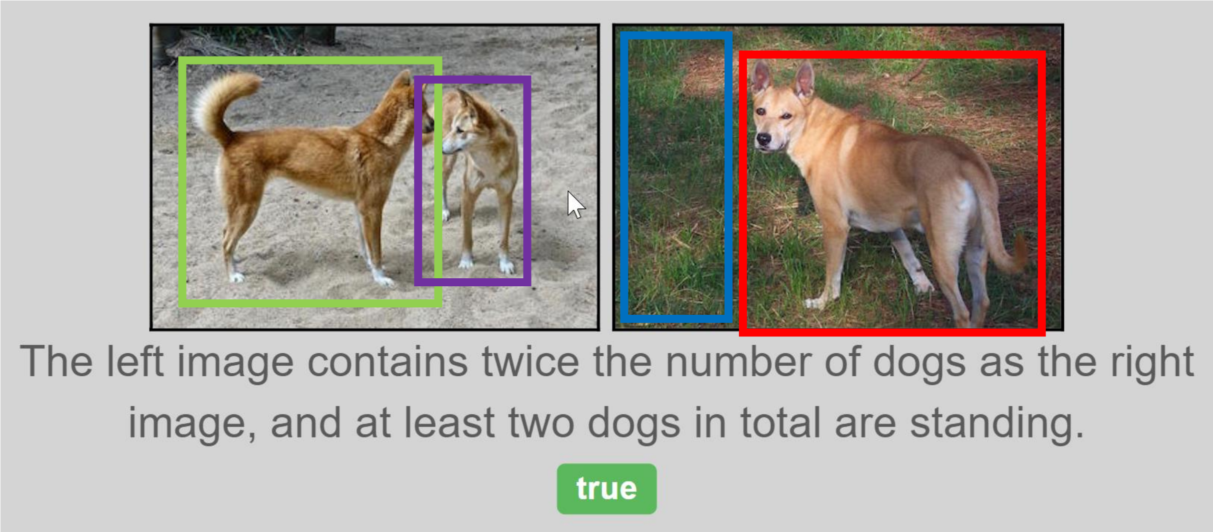
One image shows exactly two brown acorns in back-to-back caps on green foliage.

false

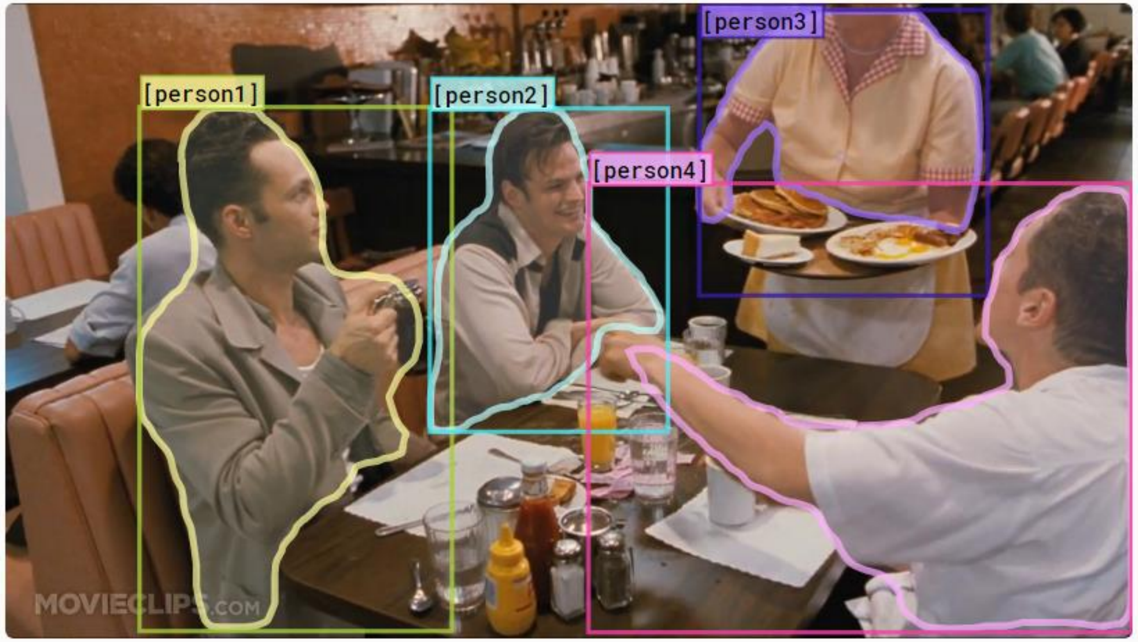
[Suhr et al., ACL 2019]

Downstream Task 3: Natural Language for Visual Reasoning





Downstream Task 4: Visual Commonsense Reasoning



Why is [person4] pointing at [person1]?

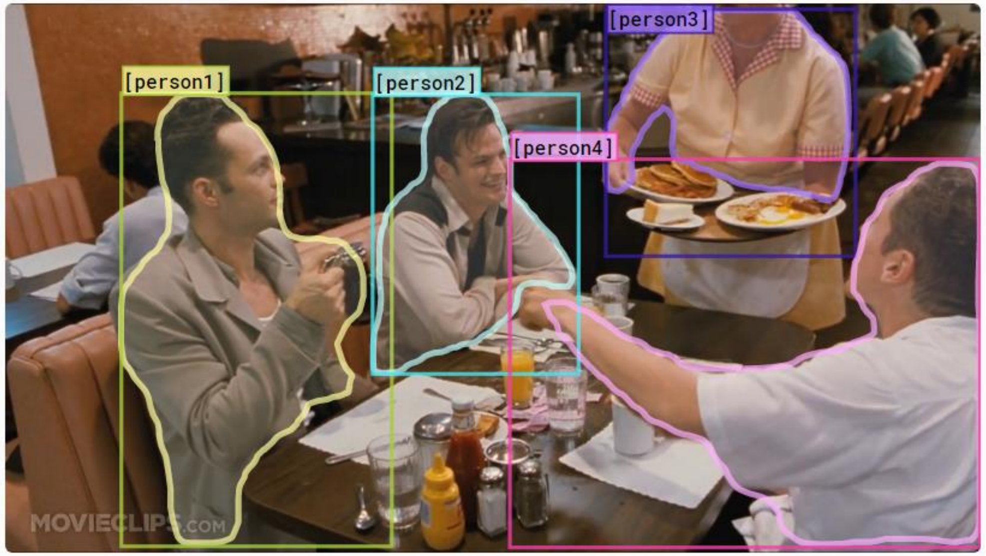
- **a) He is telling [person3] that [person1] ordered the pancakes.**
- b) He just told a joke.
- c) He is feeling accusatory towards [person1].
- d) He is giving [person1] directions.

I choose (a) 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 and both she and [person2] are smiling slightly.
- **d) [person3] is delivering food to the table, and she might not know whose order is whose.**

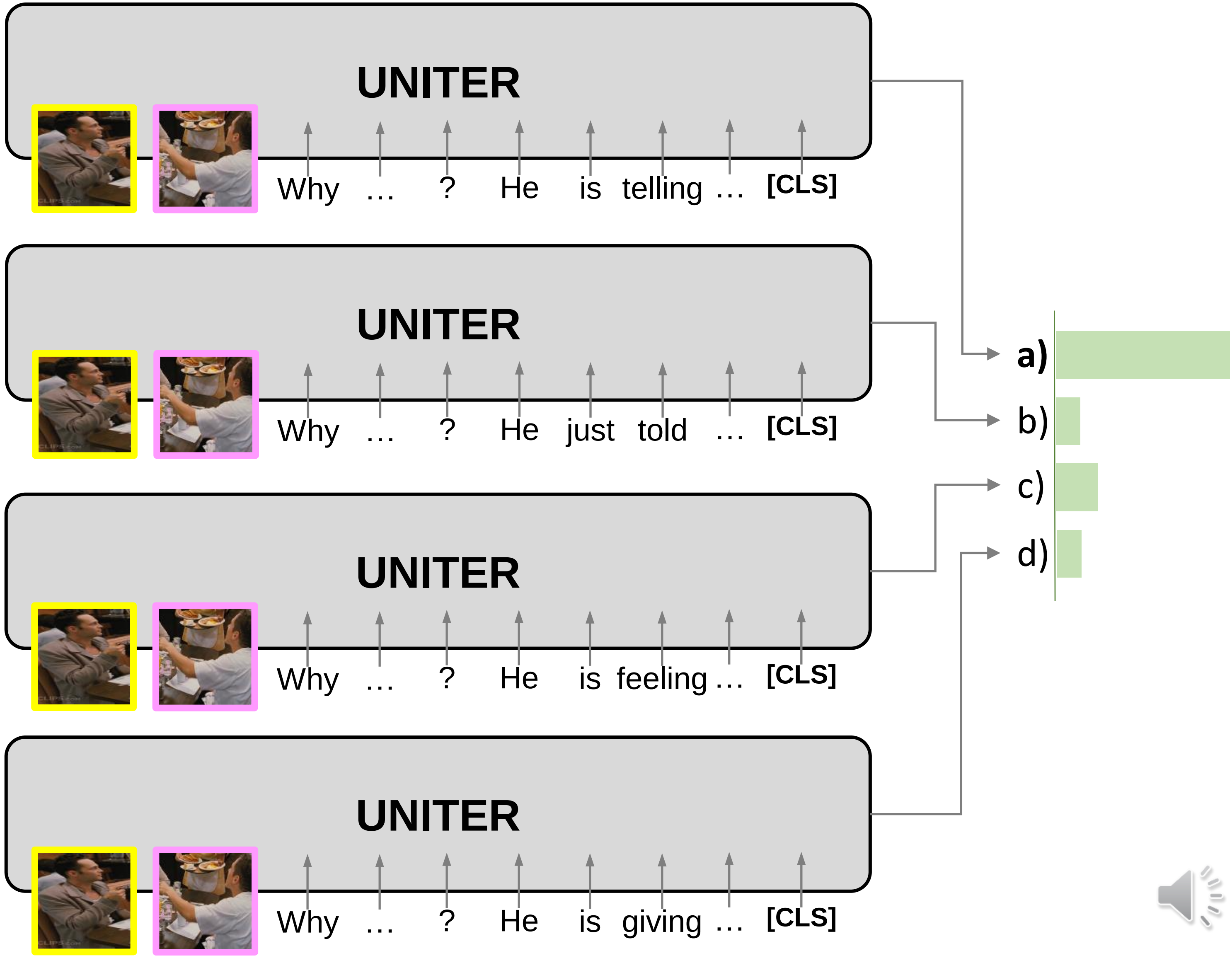
[Zellers et al., CVPR 2019]

Downstream Task 4: Visual Commonsense Reasoning



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.

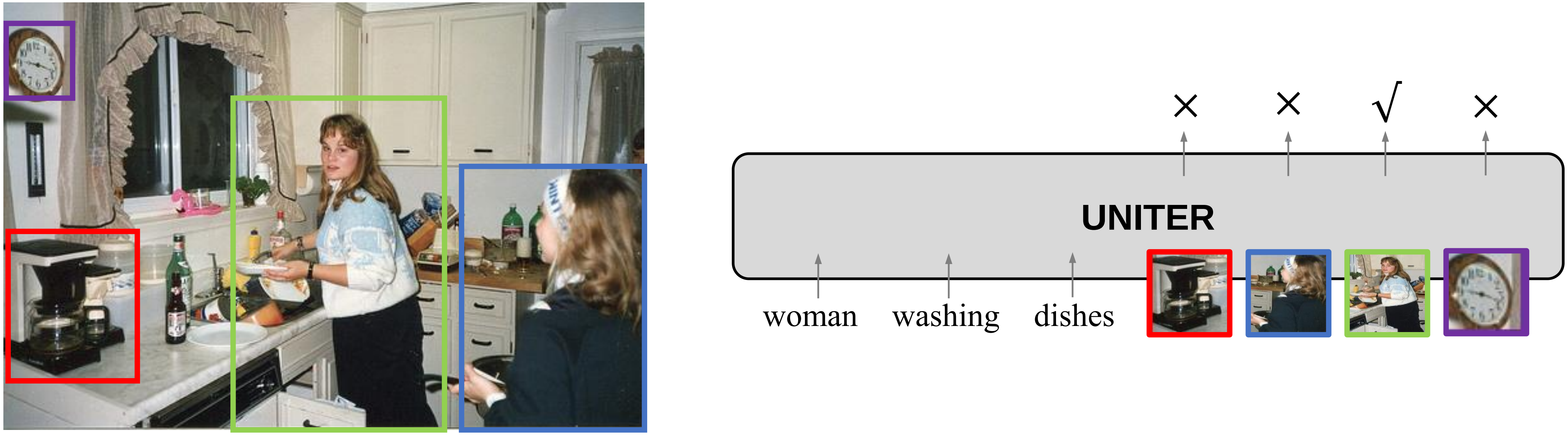


Downstream Task 5: Referring Expression Comprehension

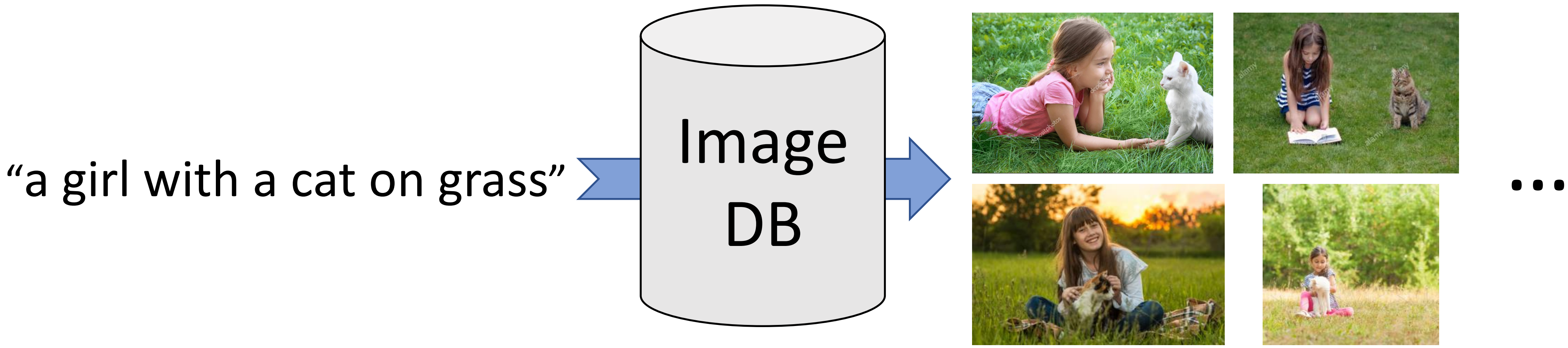
woman washing dishes

[Kazemzadeh et al., EMNLP 2014]

Downstream Task 5: Referring Expression Comprehension



Downstream Task 6: Image-Text Retrieval

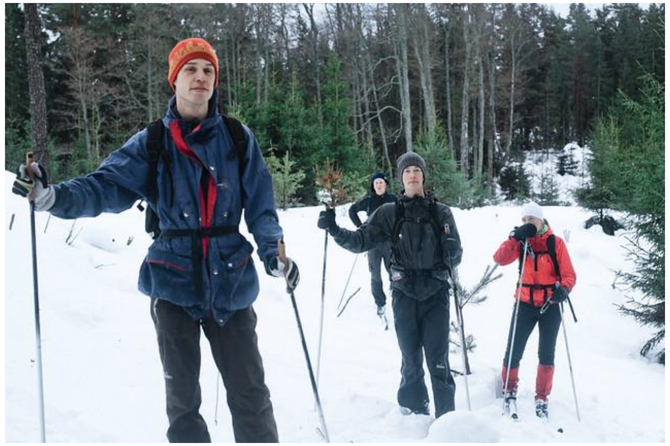


Downstream Task 6: Image-Text Retrieval

“a girl with a cat on grass”

Image DB

…

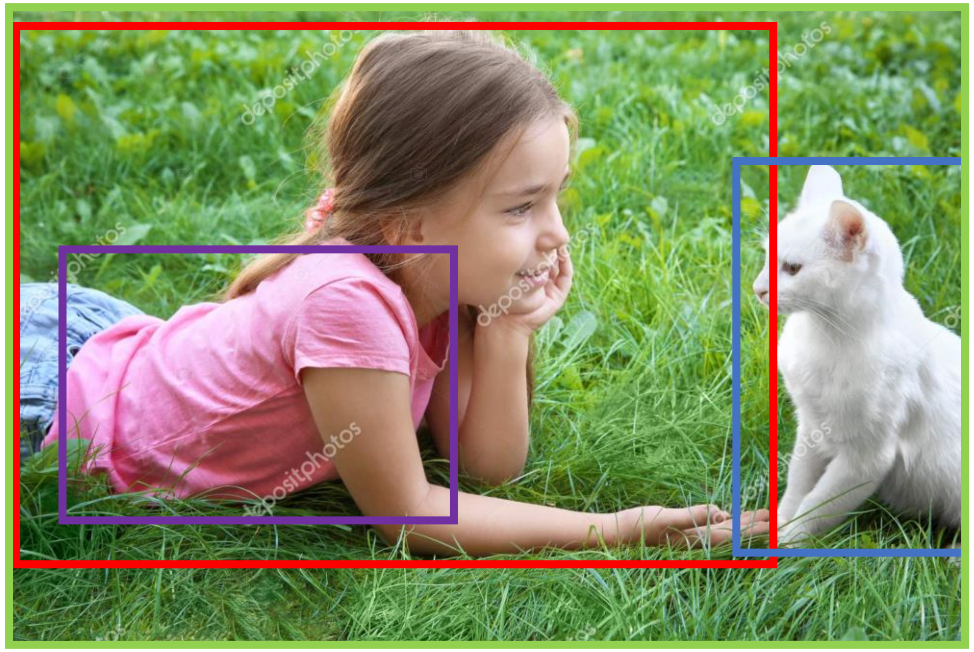


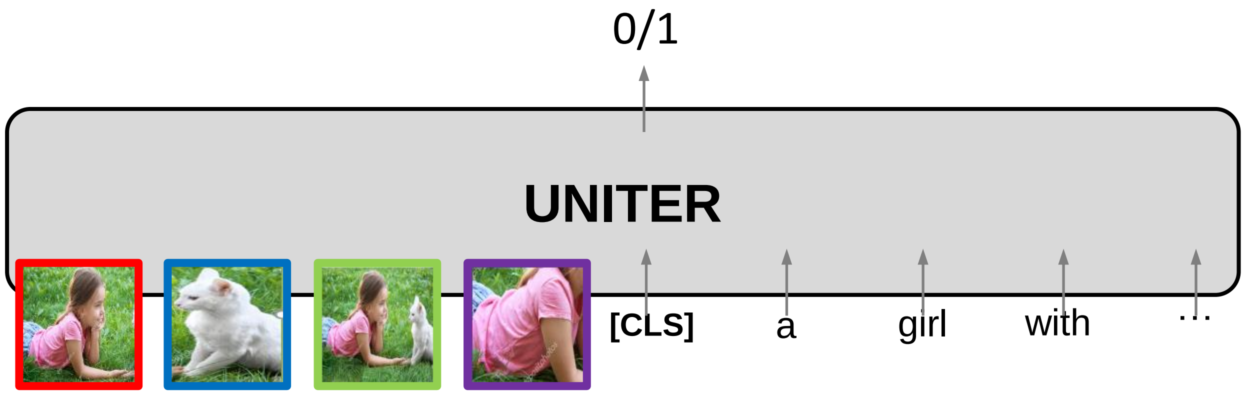
Text DB

“four people with ski poles in their hands in the snow”
“four skiers hold on to their poles in a snowy forest”
“a group of young men riding skis”
“skiers pose for a picture while outside in the woods”
“a group of people cross country skiing in the woods”

⋮

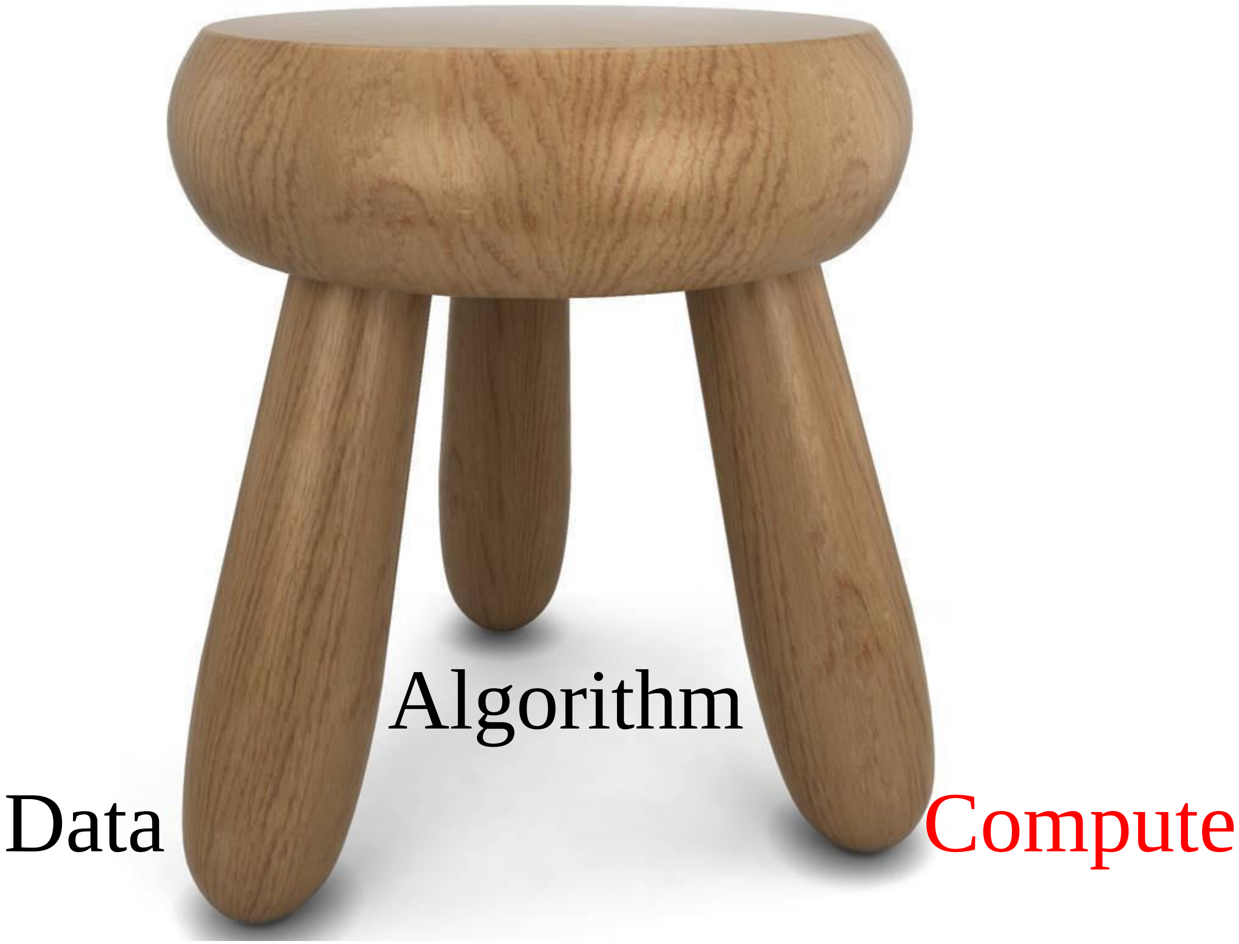
Downstream Task 6: Image-Text Retrieval





Lee et al., ECCV 2018

Self-Supervised Learning for Vision + Language

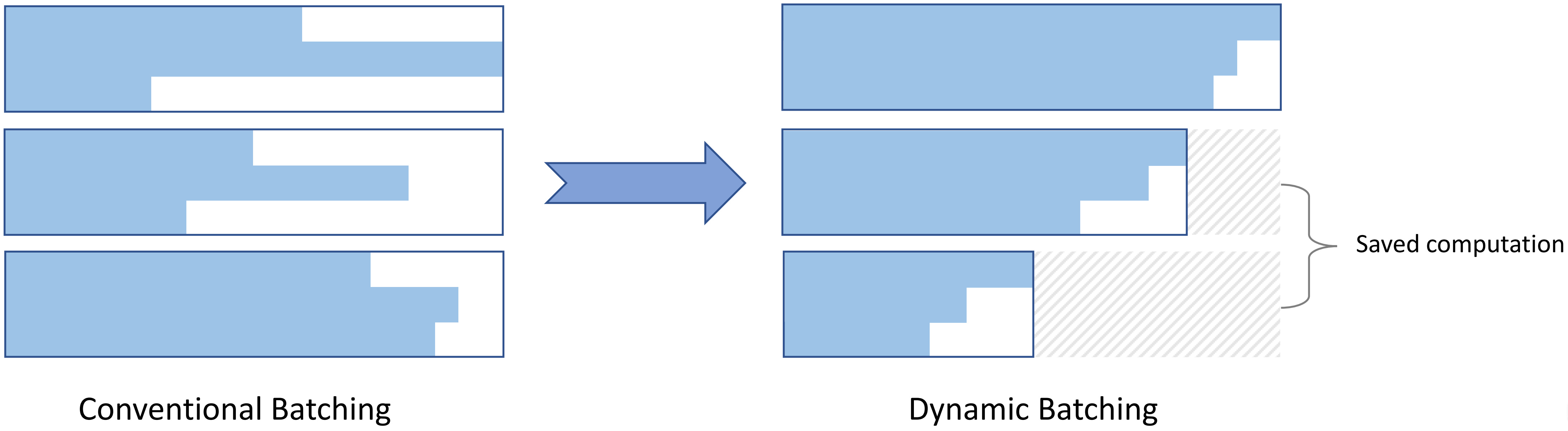


Optimization for Faster Training

- Dynamic Batching
- Gradient Accumulation
- Mixed-precision Training

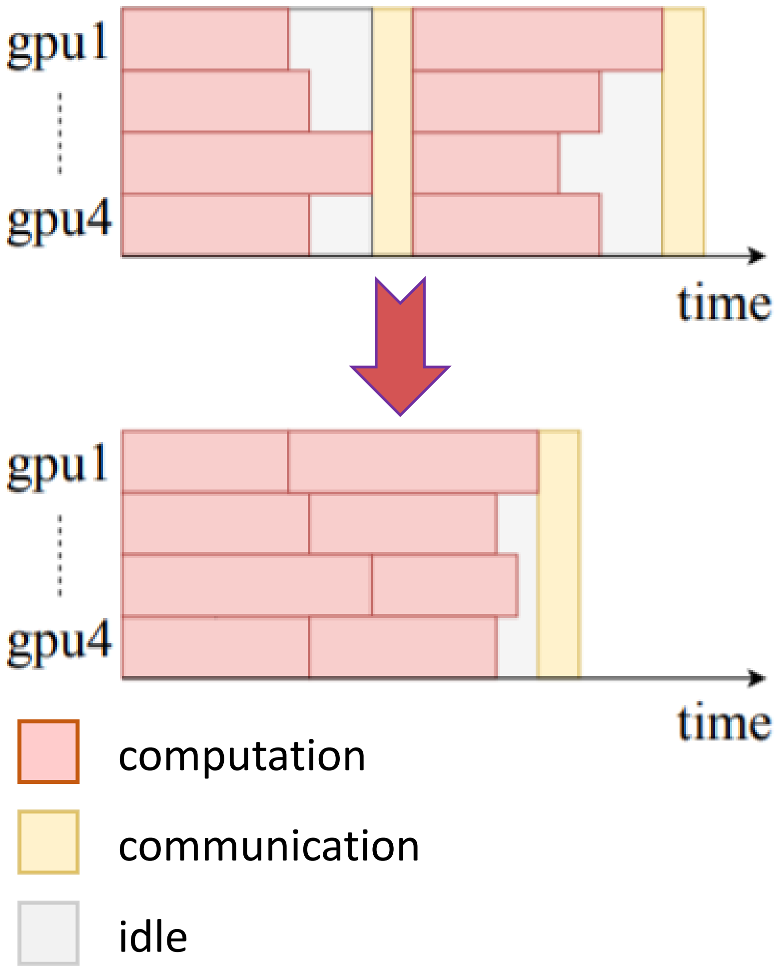
Optimization for Faster Training

- Dynamic Batching
  - Transformer (self-attention) is $O(L^2)$ (L: number of word + region)
  - Common practice: pad the input to the same maximum length (too long)
  - Our solution: batch data by similar length and only do minimum padding



Optimization for Faster Training

- Dynamic Batching
- Gradient Accumulation
  - For large models, the main training bottleneck is **network communication overhead** between nodes
  - We reduce the communication frequency, hence increase overall throughput



[Ott et al., WMT 2018]

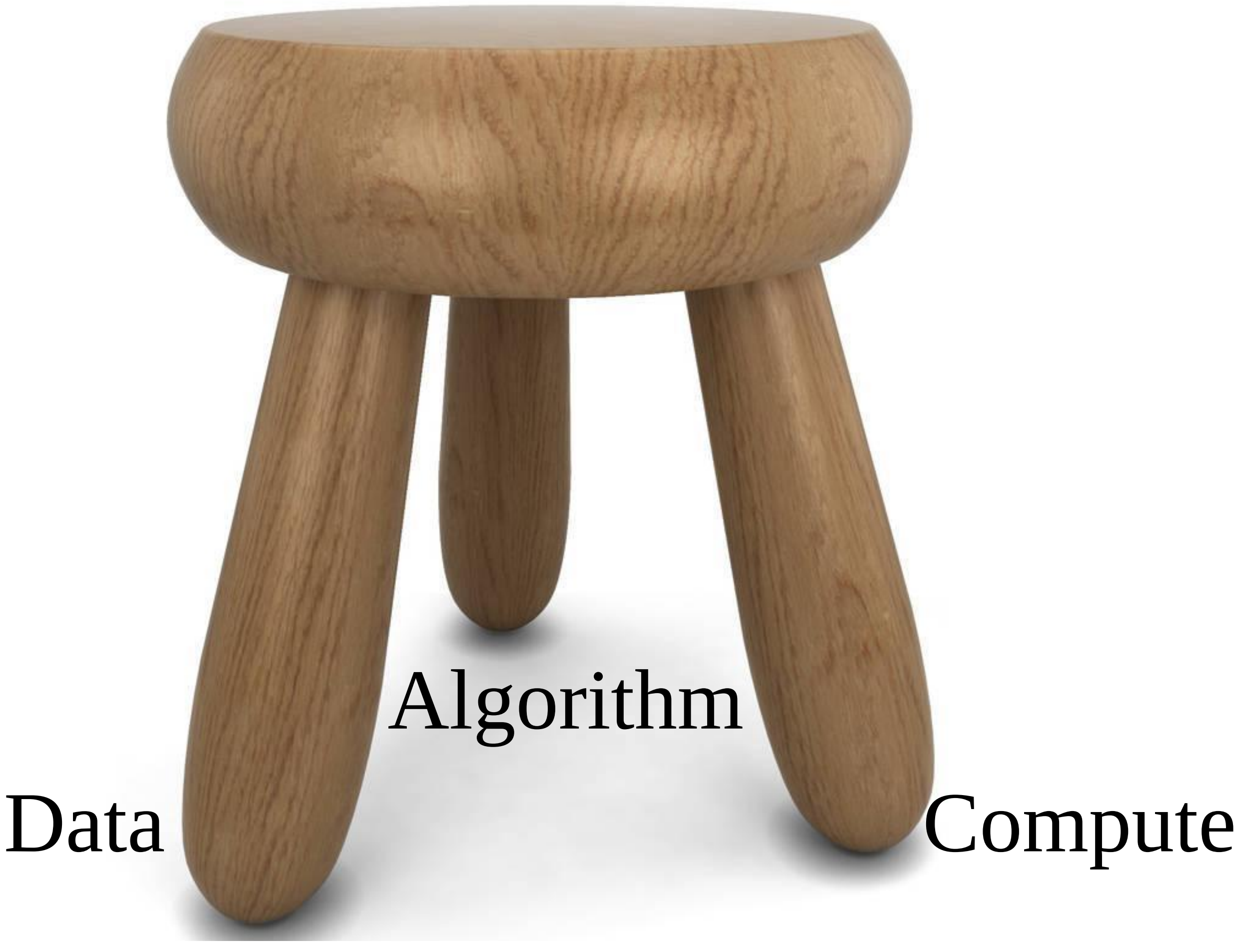
Optimization for Faster Training

- Dynamic Batching
- Gradient Accumulation
- Mixed-precision Training
  - Bring in the benefits from both worlds of 16-bit and 32-bit
  - **2x~4x speedup** compared to standard training

	Fp-16	Fp-32
Speed	**Fast**	Slow
Memory	**Low**	High
Numerical Stability	Bad	**Good**

apex (https://github.com/NVIDIA/apex)

Self-Supervised Learning for Vision + Language



SOTA of V+L Tasks
(Early 2020)

- VQA: UNITER
- VCR: UNITER
- GQA: NSM* [Hudson et al., NeurIPS 2019]
- NLVR2: UNITER
- Visual Entailment: UNITER
- Image-Text Retrieval: UNITER
- Image Captioning: VLP
- Referring Expressions: UNITER

*: without V+L pre-training

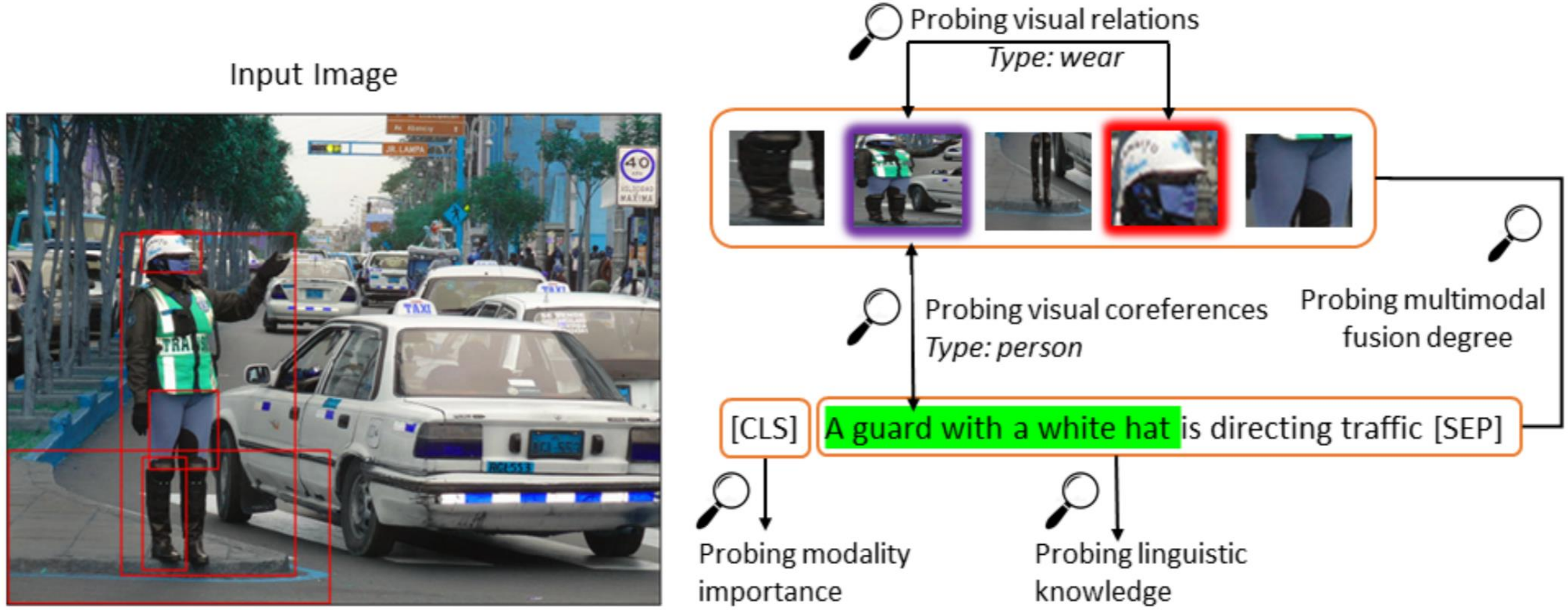
Tasks		SOTA	ViLBERT	VLBERT (Large)	Unicoder -VL	VisualBERT	LXMERT	UNITER Base	UNITER Large
VQA	test-dev	70.63	70.55	71.79	-	70.80	72.42	72.70	**73.82**
	test-std	70.90	70.92	72.22	-	71.00	72.54	72.91	**74.02**
VCR	Q→A	72.60	73.30	75.80	-	71.60	-	75.00	**77.30**
	QA→R	75.70	74.60	78.40	-	73.20	-	77.20	**80.80**
	Q→AR	55.00	54.80	59.70	-	52.40	-	58.20	**62.80**
NLVR^2	dev	54.80	-	-	-	67.40	74.90	77.18	**79.12**
	test-P	53.50	-	-	-	67.00	74.50	77.85	**79.98**
SNLI-VE	val	71.56	-	-	-	-	-	78.59	**79.39**
	test	71.16	-	-	-	-	-	78.28	**79.38**
ZS IR (Flickr)	R@1	-	31.86	-	48.40	-	-	66.16	**68.74**
	R@5	-	61.12	-	76.00	-	-	88.40	**89.20**
	R@10	-	72.80	-	85.20	-	-	92.94	**93.86**
IR (Flickr)	R@1	48.60	58.20	-	71.50	-	-	72.52	**75.56**
	R@5	77.70	84.90	-	91.20	-	-	92.36	**94.08**
	R@10	85.20	91.52	-	95.20	-	-	96.08	**96.76**
IR (COCO)	R@1	38.60	-	-	48.40	-	-	50.33	**52.93**
	R@5	69.30	-	-	76.70	-	-	78.52	**79.93**
	R@10	80.40	-	-	85.90	-	-	87.16	**87.95**
ZS TR (Flickr)	R@1	-	-	-	64.30	-	-	80.70	**83.60**
	R@5	-	-	-	85.80	-	-	**95.70**	**95.70**
	R@10	-	-	-	92.30	-	-	**98.00**	97.70
TR (Flickr)	R@1	67.90	-	-	86.20	-	-	85.90	**87.30**
	R@5	90.30	-	-	96.30	-	-	97.10	**98.00**
	R@10	95.80	-	-	99.00	-	-	98.80	**99.20**
TR (COCO)	R@1	50.40	-	-	62.30	-	-	64.40	**65.68**
	R@5	82.20	-	-	87.10	-	-	87.40	**88.56**
	R@10	90.00	-	-	92.80	-	-	93.08	**93.76**
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	-	80.31	-	-	-	83.66	**84.25**
	testA	80.04	-	83.62	-	-	-	86.19	**86.34**
	testB	69.30	-	75.45	-	-	-	78.89	**79.75**
	val^d	68.19	72.34	72.59	-	-	-	75.31	**75.90**
	testA^d	75.97	78.52	78.57	-	-	-	81.30	**81.45**
	testB^d	57.52	62.61	62.30	-	-	-	65.58	**66.70**
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**

Moving Forward…

- Interpretability of VLP models
  - VALUE [Cao et al., 2020]
- Better visual features
  - Pixel-BERT [Huang et al., 2020]
  - OSCAR [Li et al., 2020]
- Adversarial (pre-)training for V+L
  - VILLA [Gan et al., 2020]

What do V+L pretrained models learn?

VALUE: **V**ision-**A**nd-**L**anguage **U**nderstanding **E**valuation



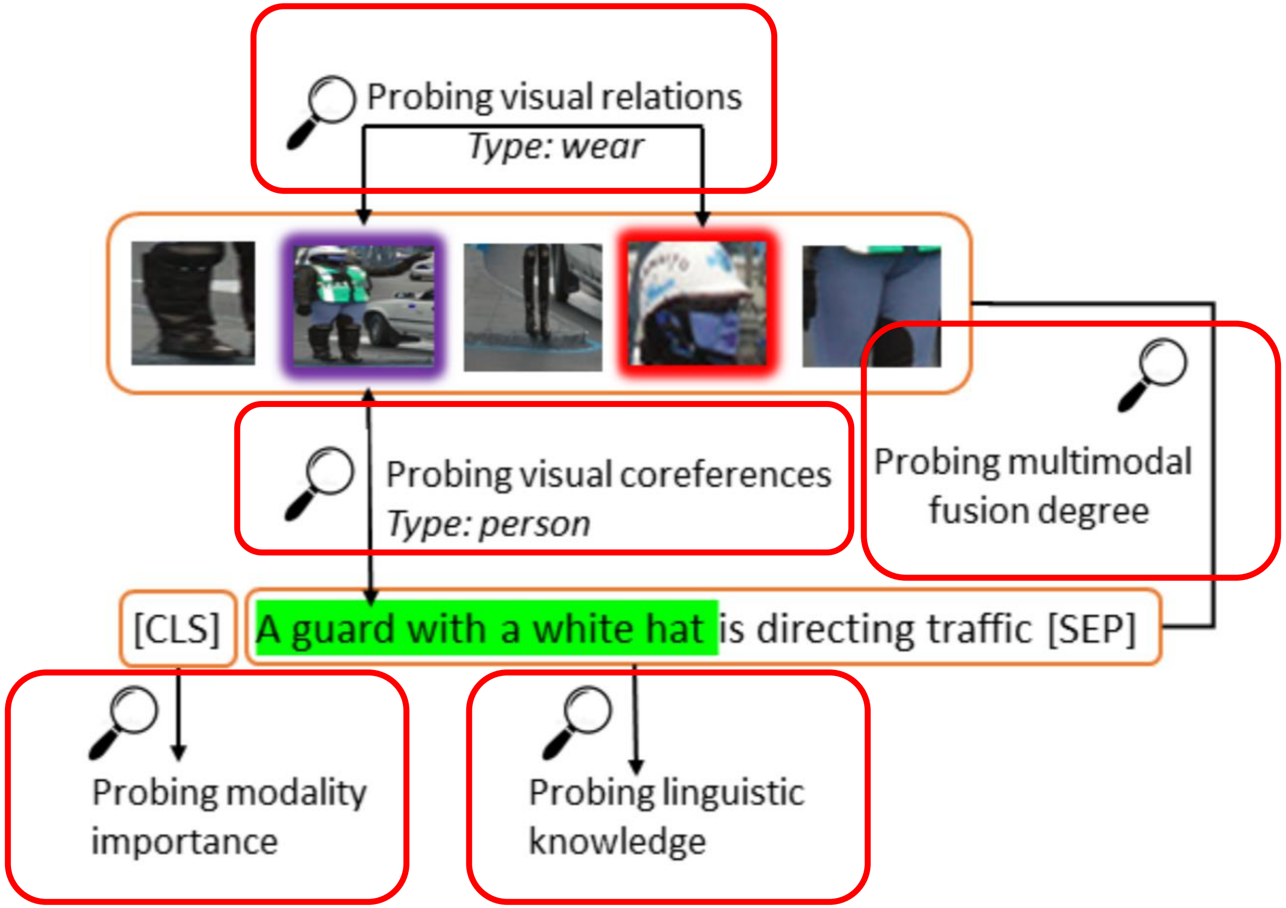
[Value, Cao et al., 2020]

Probing Pre-Trained Models

- Single-stream vs. two-stream
- Attention weight probing
  - 12 layers x 12 heads = 144 attention weight matrices
- Embedding probing
  - 768-dim x 12 layers

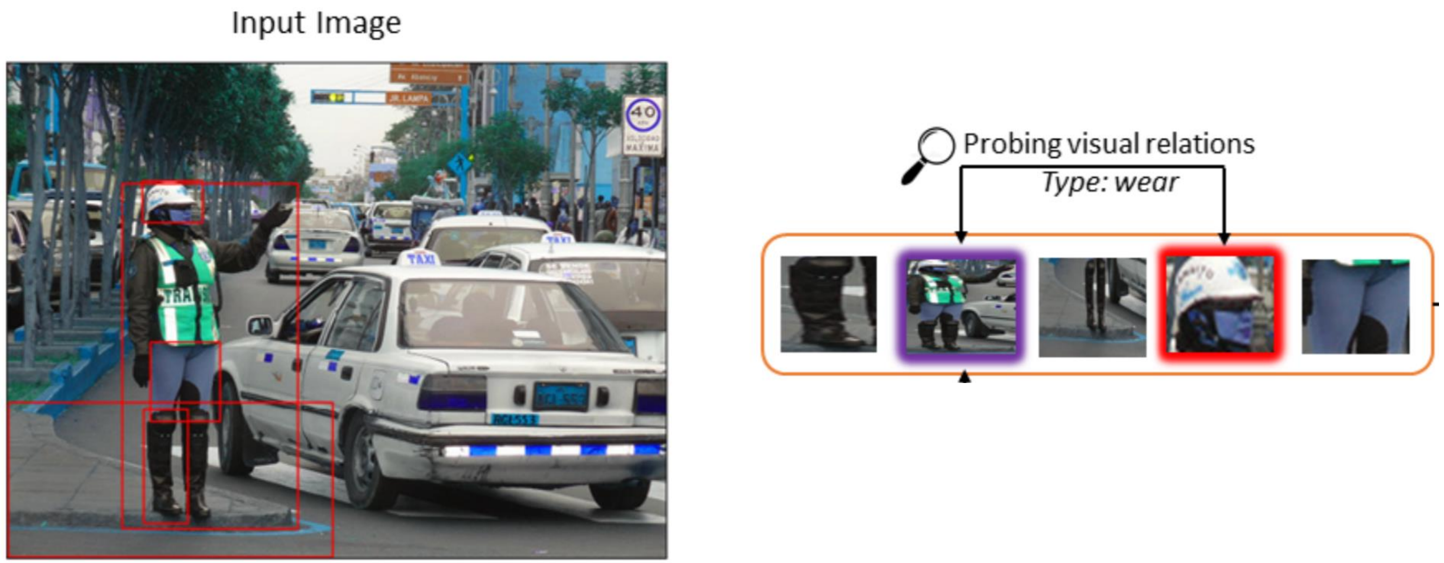
Modality Probing

- Visual Probing
- Linguistic Probing
- Cross-Modality Probing



Modality Probing

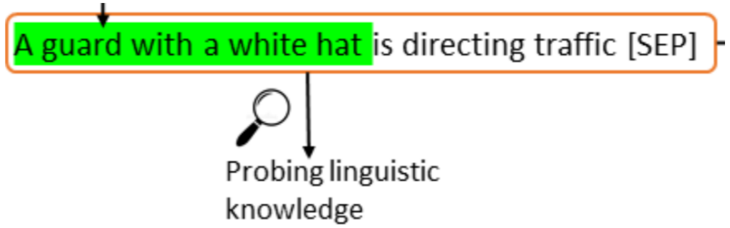
- Visual Probing
  - Visual relation detection (existence, type)
  - VG dataset; top-32 frequent relations



Modality Probing

- Visual Probing
- Linguistic Probing
  - Surface tasks (sentence length)
  - Syntactic tasks (syntax tree, top constituents, …)
  - Semantic tasks (tense, subject/object, …)

Input Image



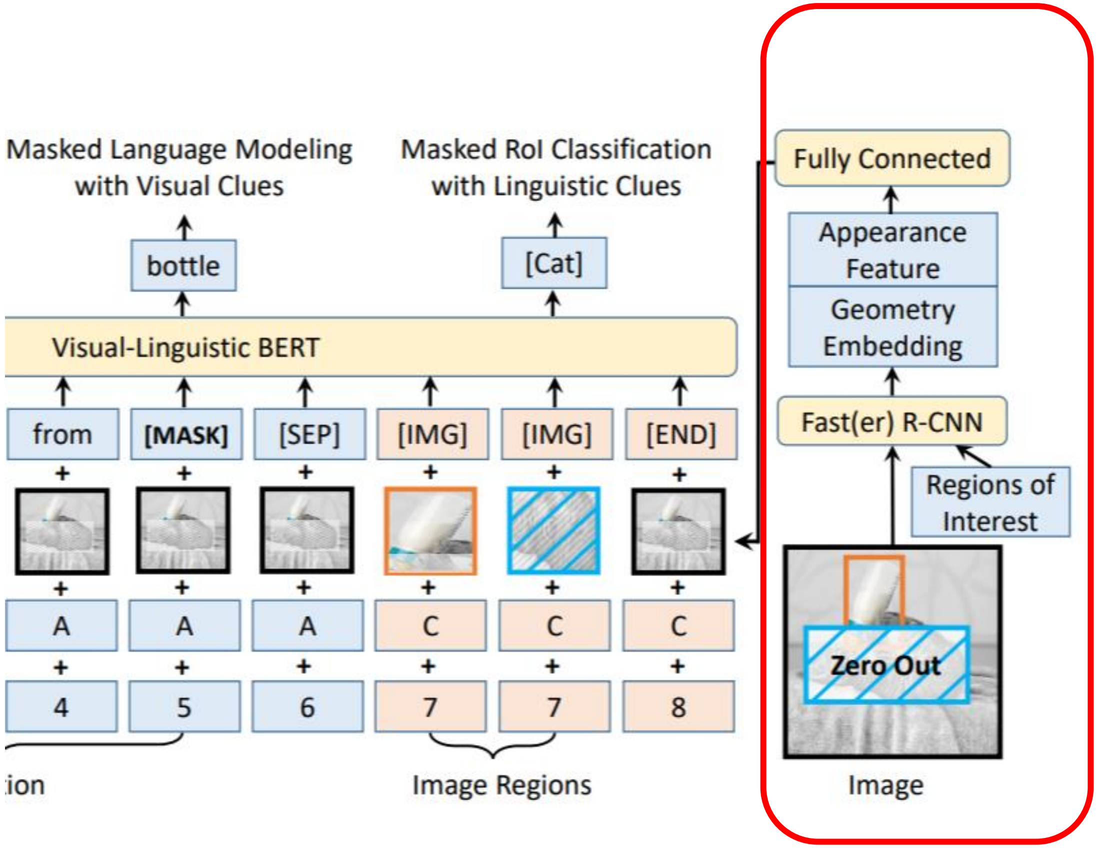
Modality Probing

- Visual Probing
- Linguistic Probing
- Cross-Modality Probing
  - Multimodal fusion degree
  - Modality importance
  - Visual coreference

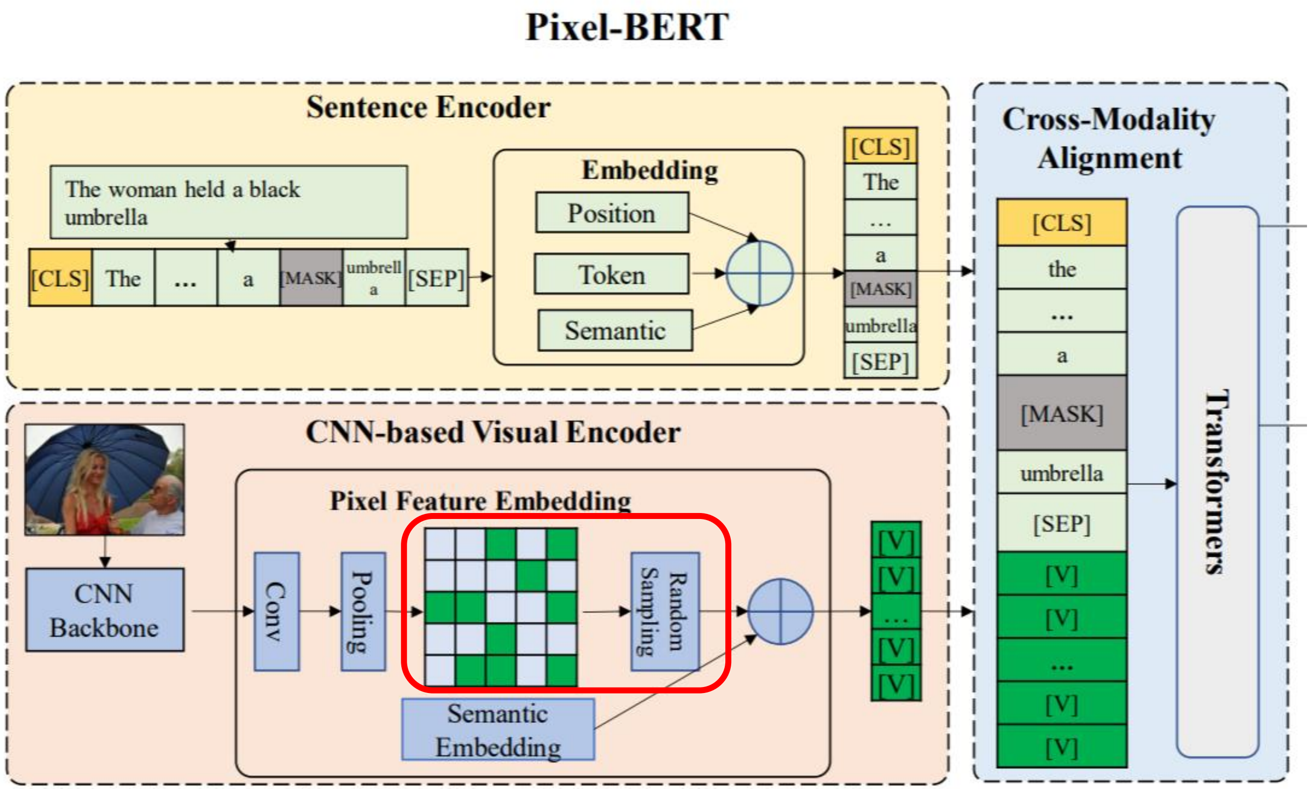
VALUE:
Vision-**A**nd-**L**anguage **U**nderstanding **E**valuation

1. Cross-modal fusion:
   a. In single-stream model (UNITER), deeper layers have more cross-modal fusion.
   b. The opposite for two-stream model (LXMERT).
2. Text modality is more important than image.
3. In single-stream model, some heads only focus on cross-modal interaction.
4. Visual relations are learned in pre-training.
5. Linguistic knowledge can be found.

From Region Features to Grid Features



[VL-BERT; Su et al., ICLR 2020]

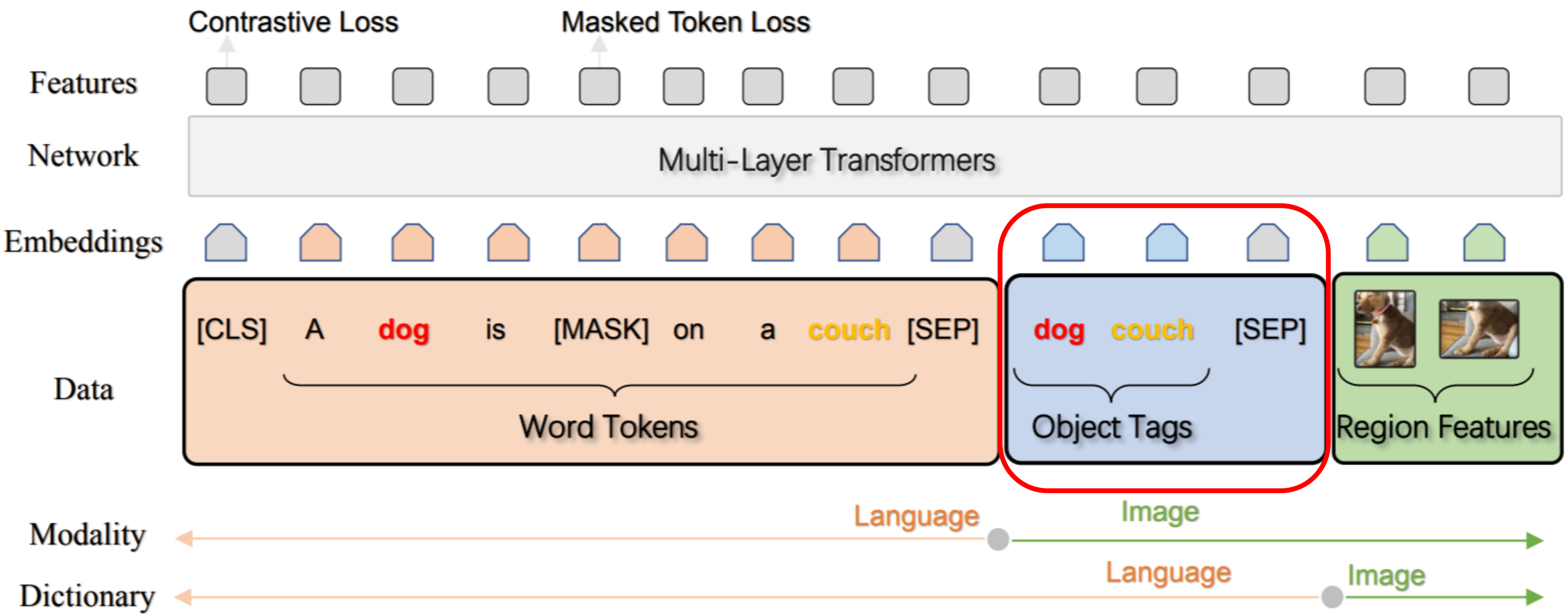


[Pixel-BERT; Huang et al., 2020]

Object Tags as Input Features

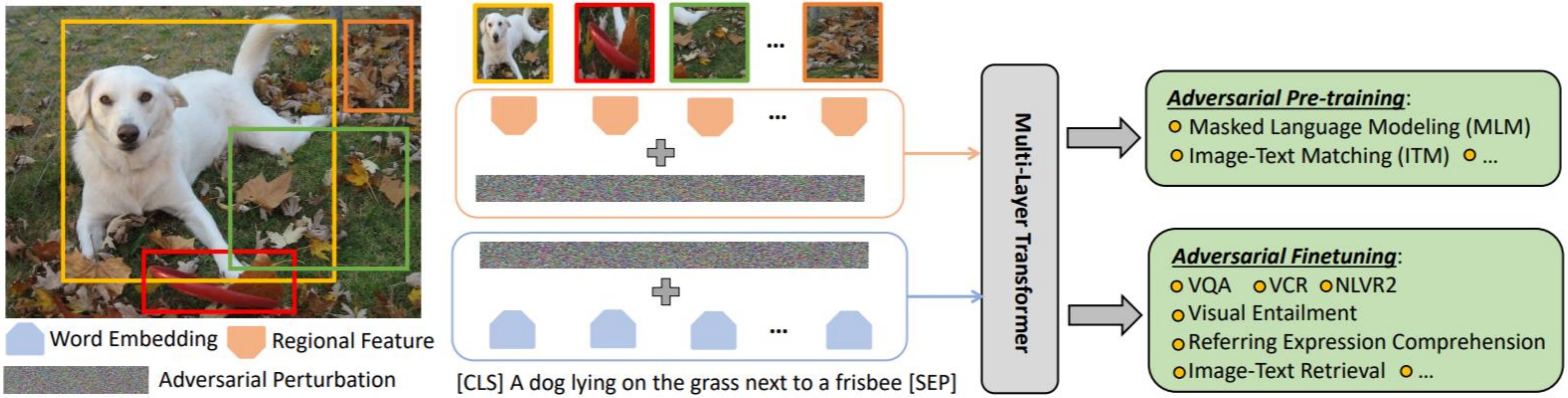
OSCAR: **O**bject-**S**emanti**c**s **A**ligned P**r**e-training

$$\boldsymbol{x} \triangleq [\underbrace{\boldsymbol{w}}_{\text{language}}, \underbrace{\boldsymbol{q}, \boldsymbol{v}}_{\text{image}}] = [\underbrace{\boldsymbol{w}, \boldsymbol{q}}_{\text{language}}, \underbrace{\boldsymbol{v}}_{\text{image}}] \triangleq \boldsymbol{x}'$$



[OSCAR; Li et al., 2020]

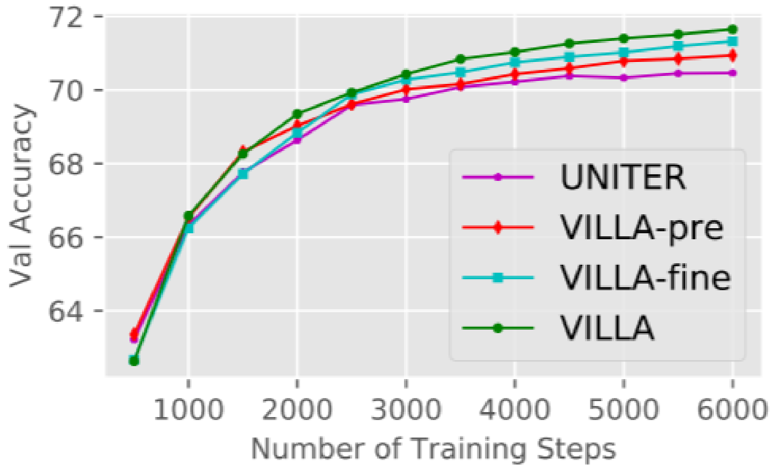
VILLA:
Vision-and-**L**anguage **L**arge-scale **A**dversarial training



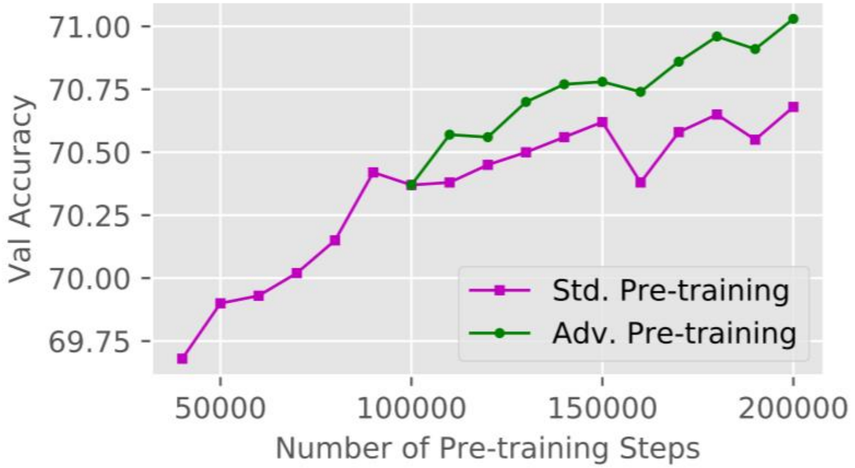
[VILLA; Gan et al., 2020]

VILLA:
Vision-and-Language Large-scale Adversarial training

1. Task-agnostic adversarial pre-training
2. Task-specific adversarial finetuning
3. "Free" adversarial training
   - FreeLB [Zhu et al., ICLR 2020]
   - KL-constraint
4. Improved generalization
   - No trade-off between accuracy and robustness.



Method	VQA		VCR			$NLVR^2$		SNLI-VE	
	test-dev	test-std	Q→A	QA→R	Q→AR	dev	test-P	val	test
$\text{VL-BERT}_{\text{LARGE}}$	71.79	72.22	75.5 (75.8)	77.9 (78.4)	58.9 (59.7)	-	-	-	-
$\text{Oscar}_{\text{LARGE}}$	73.61	73.82	-	-	-	79.12	80.37	-	-
$\text{UNITER}_{\text{LARGE}}$	73.82	74.02	77.22 (77.3)	80.49 (80.8)	62.59 (62.8)	79.12	79.98	79.39	79.38
$\text{VILLA}_{\text{LARGE}}$	**74.69**	**74.87**	**78.45 (78.9)**	**82.57 (82.8)**	**65.18 (65.7)**	**79.76**	**81.47**	**80.18**	**80.02**



(a) Standard vs. adversarial pre-training.

SOTA of V+L Tasks

- VQA: UNITER
- VCR: UNITER
- GQA: NSM* [Hudson et al., NeurIPS 2019]
- NLVR2: UNITER
- Visual Entailment: UNITER
- Image-Text Retrieval: UNITER
- Image Captioning: VLP
- Referring Expressions: UNITER

*: without V+L pre-training

SOTA of V+L Tasks

- VQA: VILLA (single), GridFeat+MoVie* (ensemble)
- VCR: VILLA
- GQA: HAN* [Kim et al., CVPR 2020]
- NLVR2: VILLA
- Visual Entailment: VILLA
- Image-Text Retrieval: OSCAR
- Image Captioning: OSCAR
- Referring Expressions: VILLA

[GridFeat; Jiang et al., CVPR 2020]
[MoVie; Nguyen et al., 2020]

*: without V+L pre-training

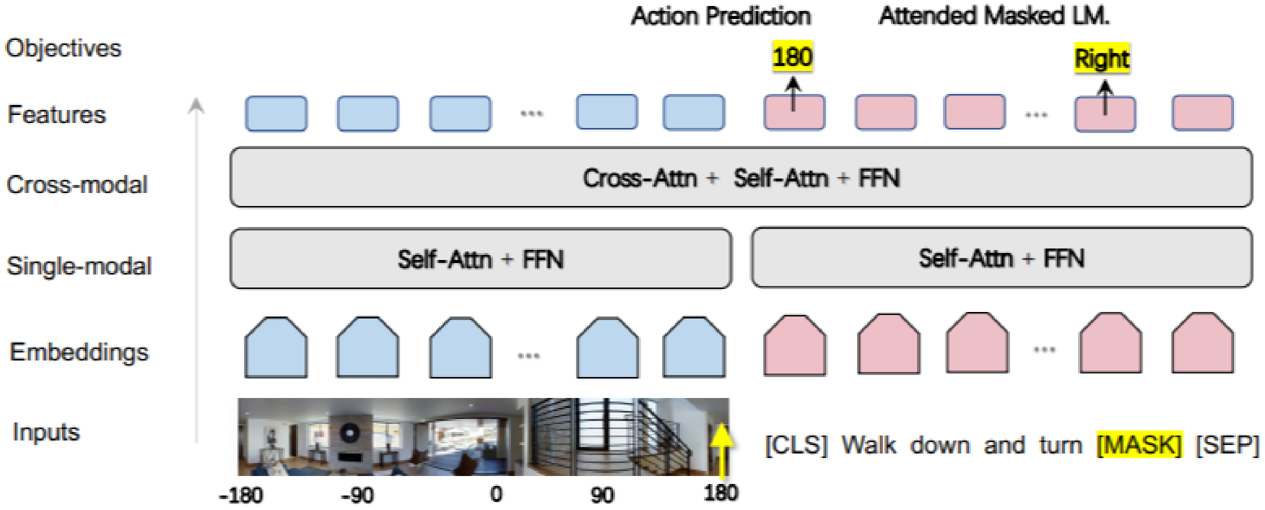
Take-away

- SOTA pre-training for V+L
    - Available datasets
    - Model architecture
    - Pre-training tasks
- Future directions
    - Study the representation learned by pre-training → pruning/compression
    - Better visual features → end-to-end training of CNN
    - Reasoning tasks (GQA)

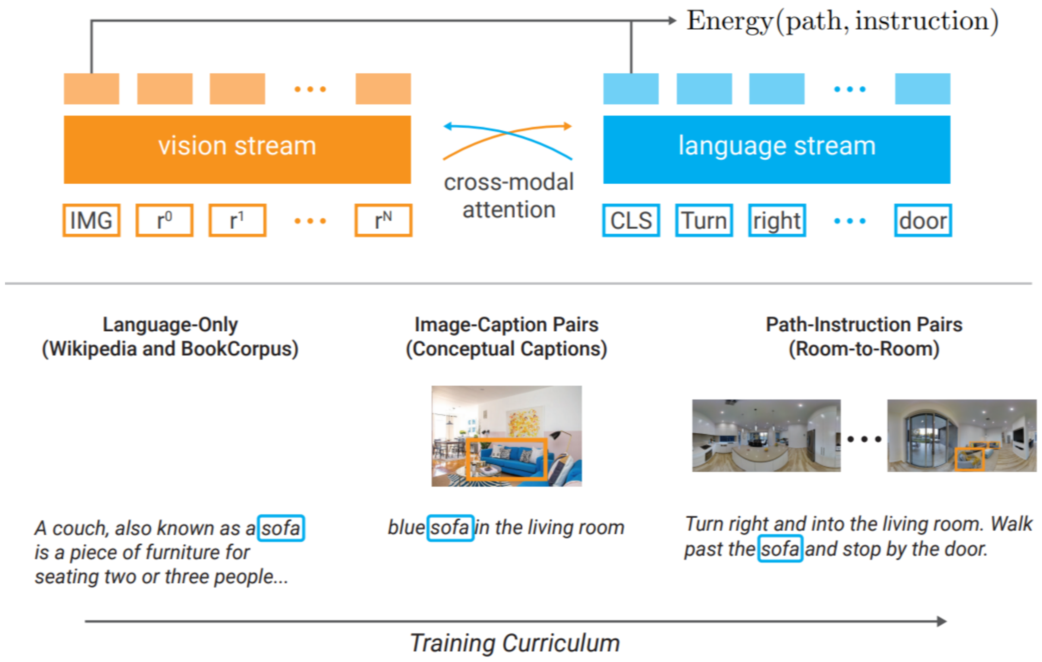
Beyond Image+Text Pre-Training

- Self-supervised learning for vision-and-language navigation (VLN)
  - PREVALENT [Hao et al., CVPR 2020]
  - VLN-BERT [Majumdar et al., 2020]
- Video+Language Pre-training

Self-Supervised Learning for VLN

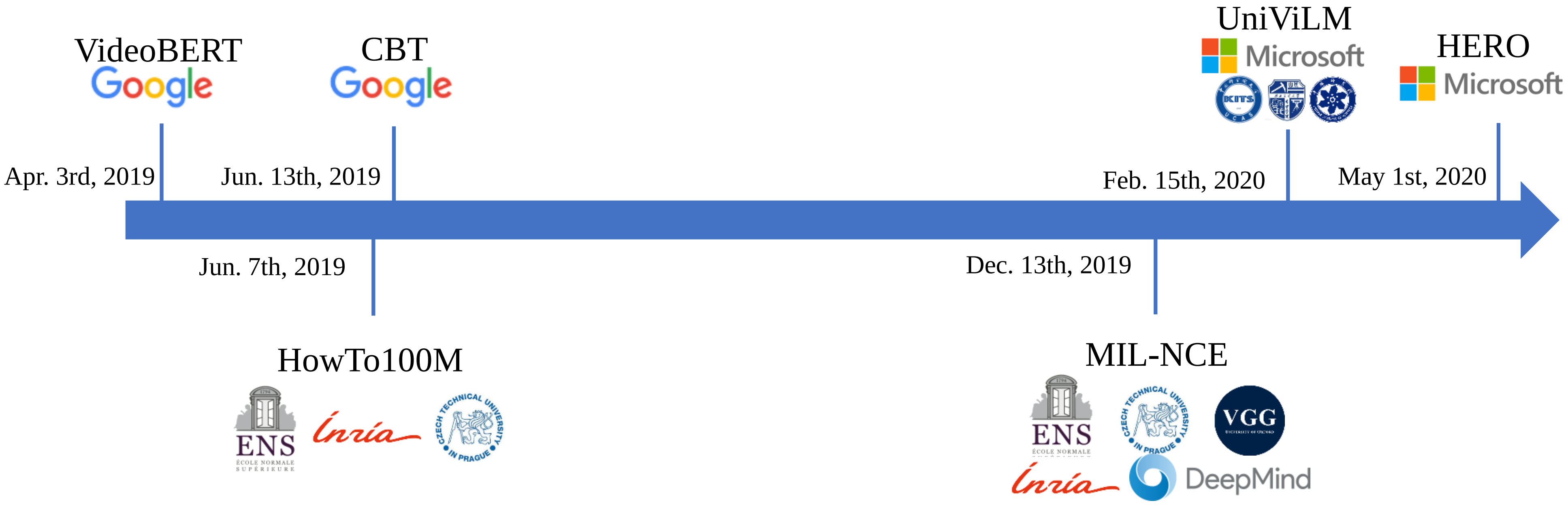


[PREVALENT; Hao et al., CVPR 2020]



[VLN-BERT; Majumdar et al., 2020]

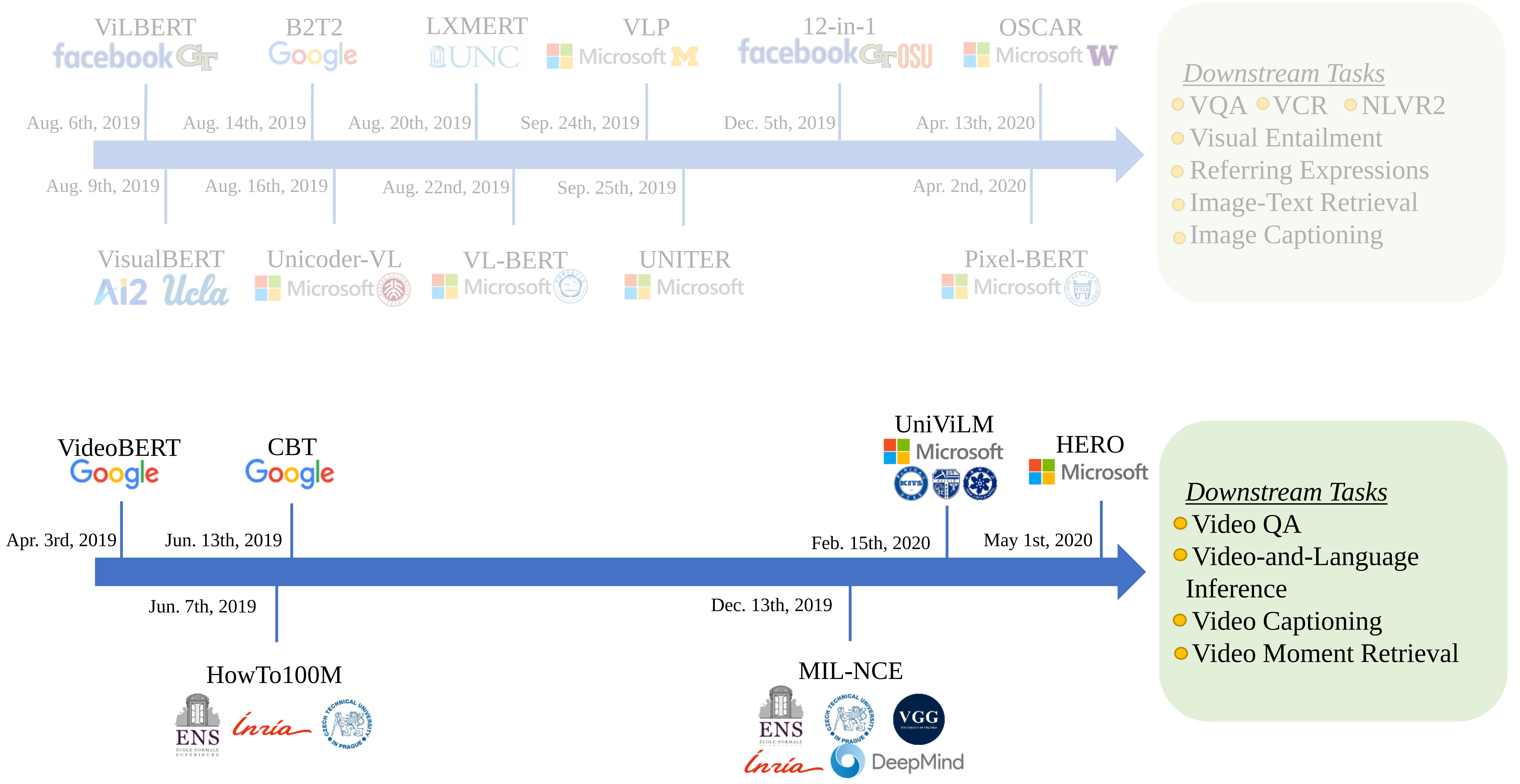
Video+Language Pre-Training



Downstream Tasks

- Video QA
- Video-and-Language Inference
- Video Captioning
- Video Moment Retrieval

Self-supervised Learning for Video-and-Language



ViLBERT
Aug. 6th, 2019
B2T2
Aug. 14th, 2019
LXMERT
UNC
Aug. 20th, 2019
VLP
Microsoft
Sep. 24th, 2019
12-in-1
OSU
Dec. 5th, 2019
OSCAR
Microsoft
Apr. 13th, 2020
Aug. 9th, 2019
VisualBERT
Aug. 16th, 2019
Unicoder-VL
Microsoft
Aug. 22nd, 2019
VL-BERT
Microsoft
Sep. 25th, 2019
UNITER
Microsoft
Apr. 2nd, 2020
Pixel-BERT
Microsoft
Downstream Tasks
VQA
VCR
NLVR2
Visual Entailment
Referring Expressions
Image-Text Retrieval
Image Captioning
VideoBERT
Apr. 3rd, 2019
CBT
Jun. 13th, 2019
UniViLM
Microsoft
Feb. 15th, 2020
HERO
Microsoft
May 1st, 2020
Jun. 7th, 2019
HowTo100M
ENS
Dec. 13th, 2019
MIL-NCE
ENS
VGG
DeepMind
Downstream Tasks
Video QA
Video-and-Language Inference
Video Captioning
Video Moment Retrieval

Video + Language Pre-training

Keep rolling tight and squeeze the air out to its side and you can kind of pull a little bit.

Image credits: https://ai.googleblog.com/2019/09/learning-cross-modal-temporal.html

Video + Language Pre-training

Video: Sequence of image frames
Language: Subtitles/Narrations

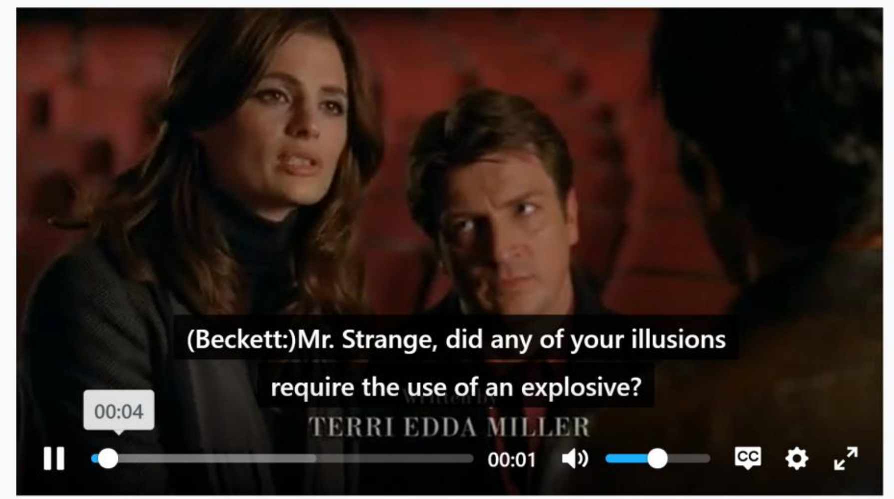
Keep rolling tight and squeeze the air out to its side and you can kind of pull a little bit.

Image credits: https://ai.googleblog.com/2019/09/learning-cross-modal-temporal.html

Pre-training Data for Video + Language

TV Dataset

[Lei et al. EMNLP 2018]



- 22K video clips from 6 popular TV shows
- Each video clip is 60-90 seconds long
- Dialogue (“character name: subtitle”) is provided

HowTo100M Dataset

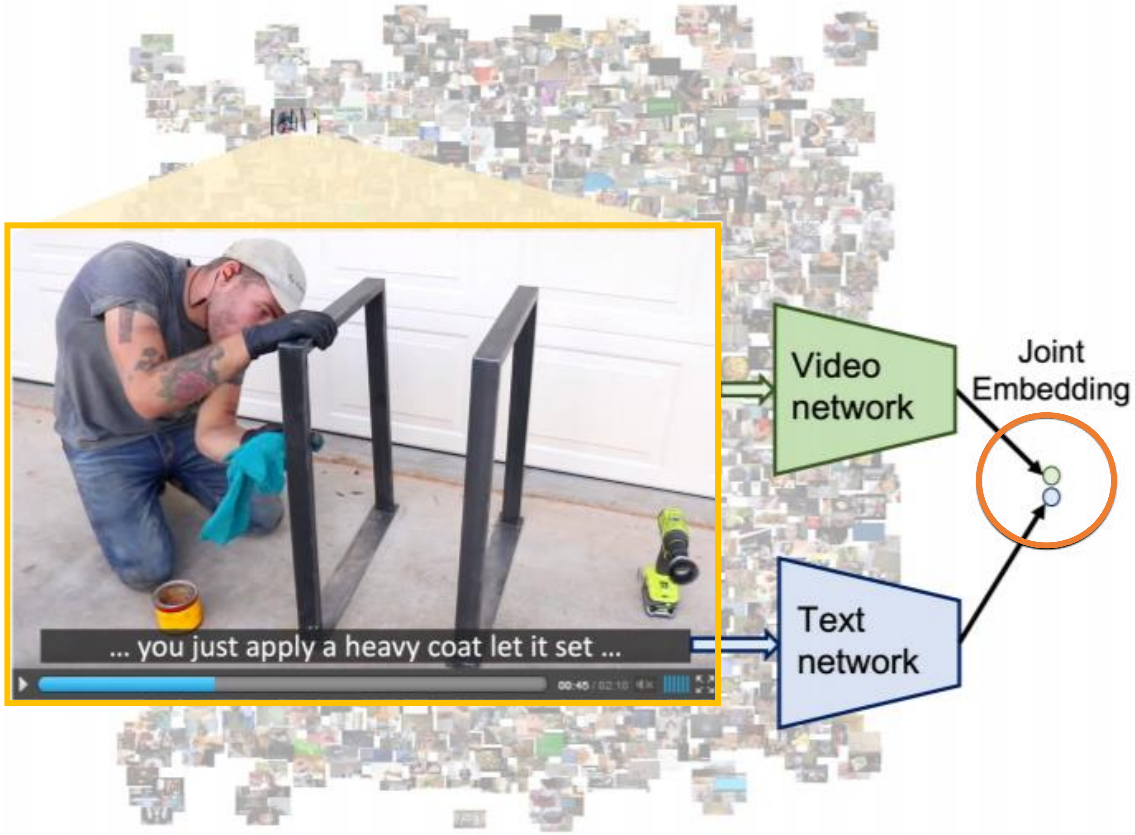
[Miech et al. ICCV 2019]

- 1.22M instructional videos from YouTube
- Each video is 6 minutes long on average
- Narrations in different languages

Image credits: from the original papers

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Pre-training



[Miech et al, ICCV 2019]

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Pre-training

Large-scale Pre-training Dataset

- 136M video clips with narrations from 1.2M YouTube videos spanning 23K activities

[Miech et al, ICCV 2019]

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Pre-training

[Miech et al, ICCV 2019]

Large-scale Pre-training Dataset

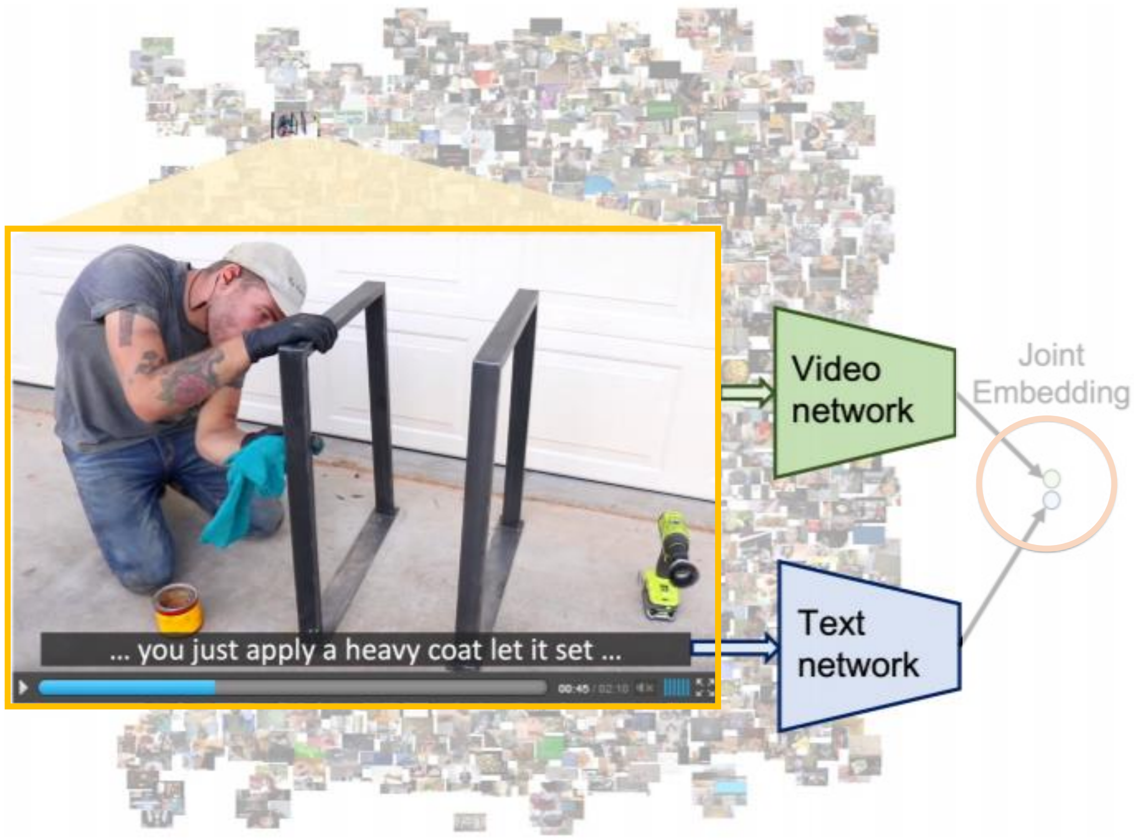
- 136M video clips with narrations from 1.2M YouTube videos spanning 23K activities

Video Representations

- 2D features from ImageNet pretrained ResNet-152
- 3D features from Kinetics pretrained ResNeXt-101

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Pre-training



[Miech et al, ICCV 2019]

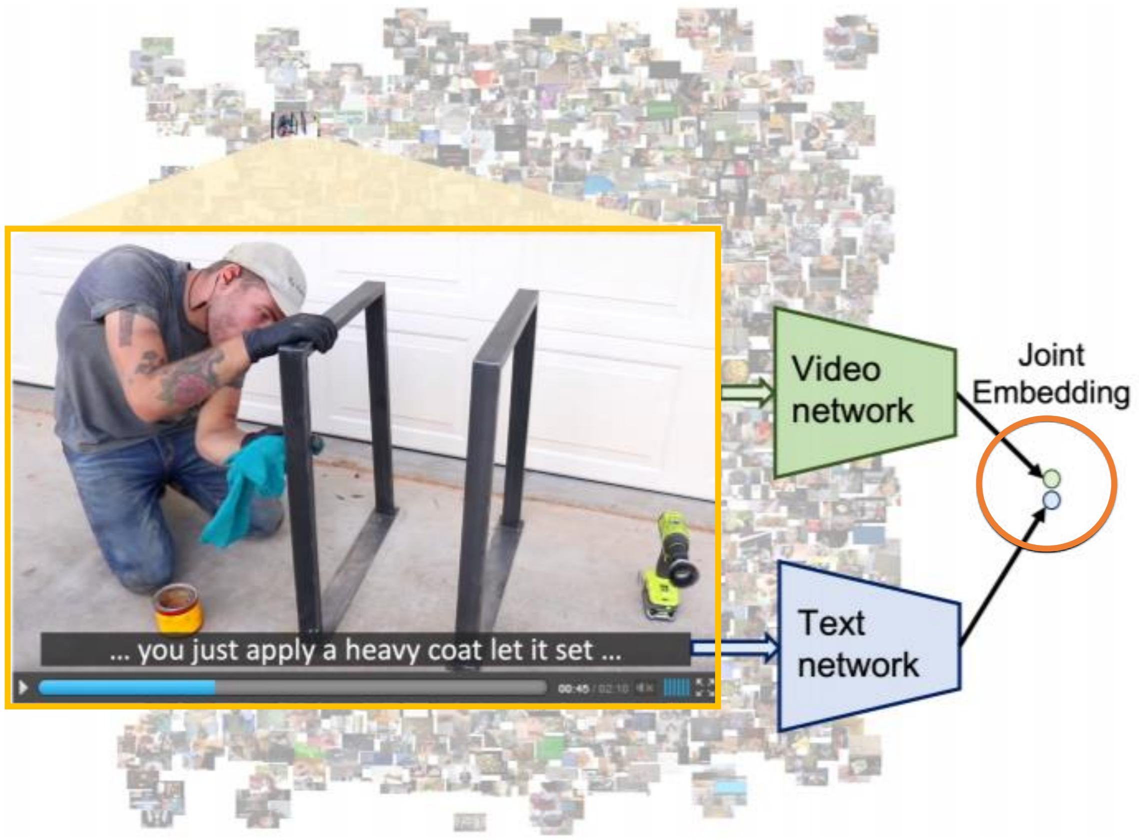
Large-scale Pre-training Dataset
- 136M video clips with narrations from 1.2M YouTube videos spanning 23K activities

Video Representations
- 2D features from ImageNet pretrained ResNet-152
- 3D features from Kinetics pretrained ResNeXt-101

Text Representations
- GoogleNews pre-trained word2vec embedding models

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Pre-training



[Miech et al, ICCV 2019]

Large-scale Pre-training Dataset
- 136M video clips with narrations from 1.2M YouTube videos spanning 23K activities

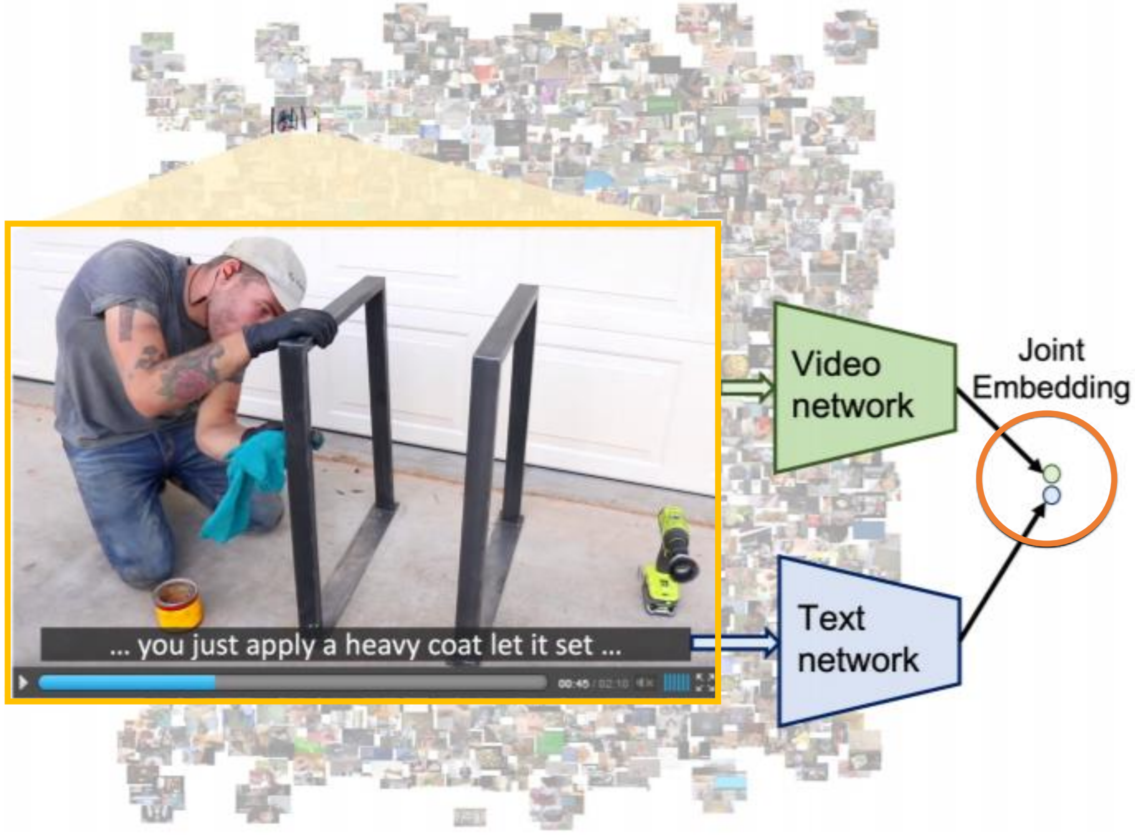
Video Representations
- 2D features from ImageNet pretrained ResNet-152
- 3D features from Kinetics pretrained ResNeXt-101

Text Representations
- GoogleNews pre-trained word2vec embeddings

Pre-training Joint Embedding
- Non-linear functions to embed both modalities to a common embedding space
- Supervise the training with max-margin ranking loss

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Pre-training



Downstream Tasks

Weakly Supervised Step Localization

Step #1
Apply the jam

Step #2
Assemble the sandwich

Retrieval

Query: Toast the bread slices in the toaster

[Miech et al, ICCV 2019]

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Model	**CrossTask (Averaged Recall)**
Fully-supervised Upper-bound [1]	31.6
HowTo100M PT only (weakly supervised)	<u>33.6</u>

<u>Step Localization</u>

- ❖ HowTo100M PT is better than training a fully supervised model on a small training set

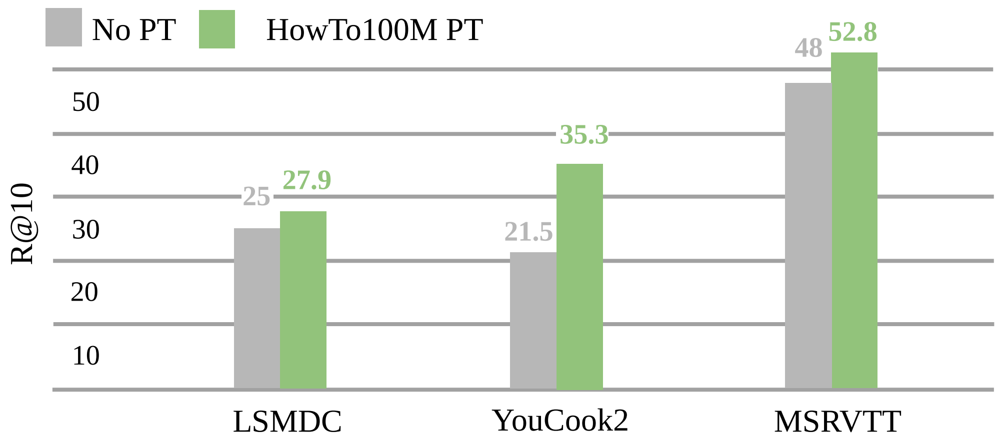
[1] Zhukov, Dimitri, et al. "Cross-task weakly supervised learning from instructional videos." CVPR 2019

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Model	CrossTask (Averaged Recall)
Fully-supervised Upper-bound [1]	31.6
HowTo100M PT only (weakly supervised)	33.6

Step Localization

- ❖ HowTo100M PT is better than training a fully supervised model on a small training set



Clip Retrieval

- ❖ HowTo100M PT largely boosts model performance despite the domain differences

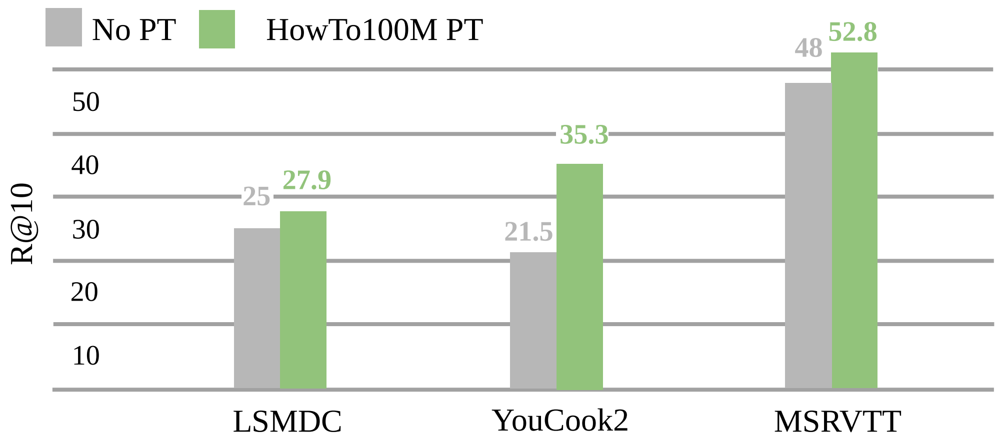
[1] Zhukov, Dimitri, et al. "Cross-task weakly supervised learning from instructional videos." CVPR 2019

HowTo100M: Learning a Text-Video Embedding from Watching Hundred Million Narrated Video Clips

Model	CrossTask (Averaged Recall)
Fully-supervised Upper-bound [1]	31.6
HowTo100M PT only (weakly supervised)	<u>33.6</u>

<u>Step Localization</u>

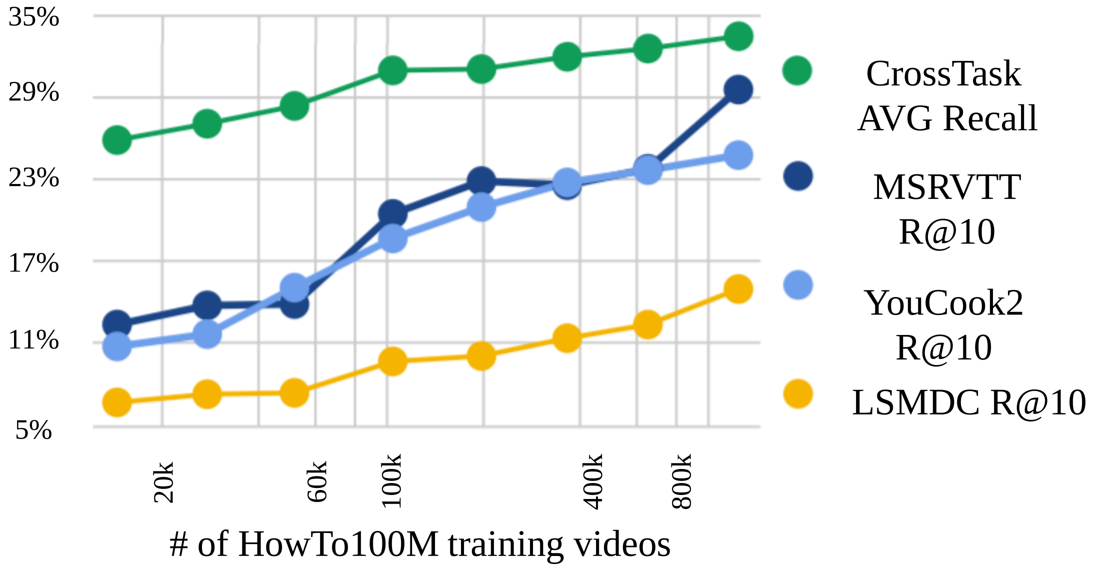
❖ HowTo100M PT is better than training a fully supervised model on a small training set



<u>Clip Retrieval</u>

❖ HowTo100M PT largely boosts model performance despite the domain differences

Downstream Performance vs. Pre-training Data Size

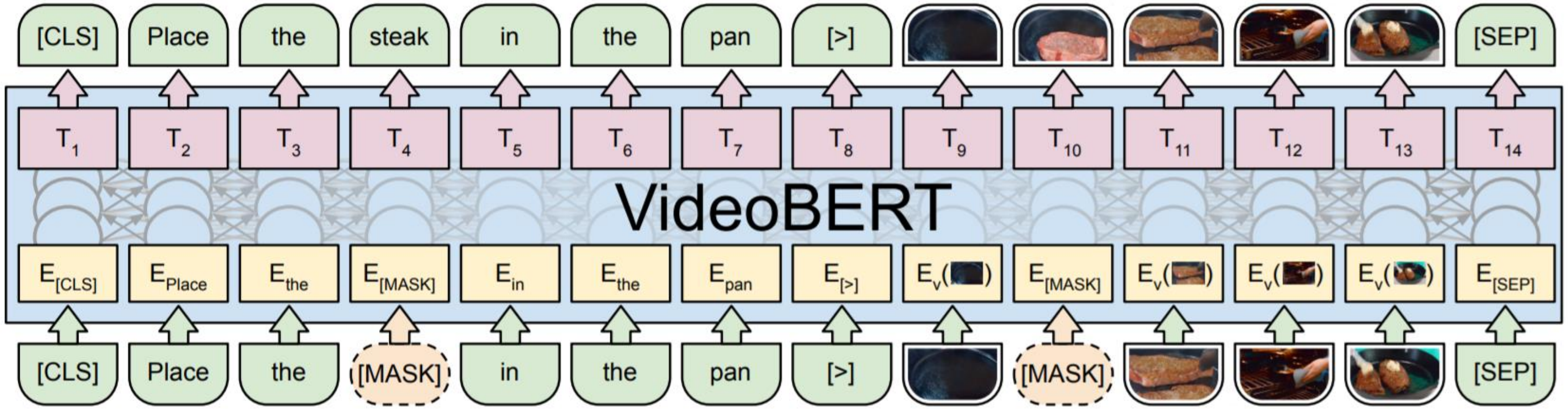


❖ Adding more data gives better results across all downstream tasks

[1] Zhukov, Dimitri, et al. "Cross-task weakly supervised learning from instructional videos." CVPR 2019

VideoBERT: A Joint Model for Video and Language Representation Learning

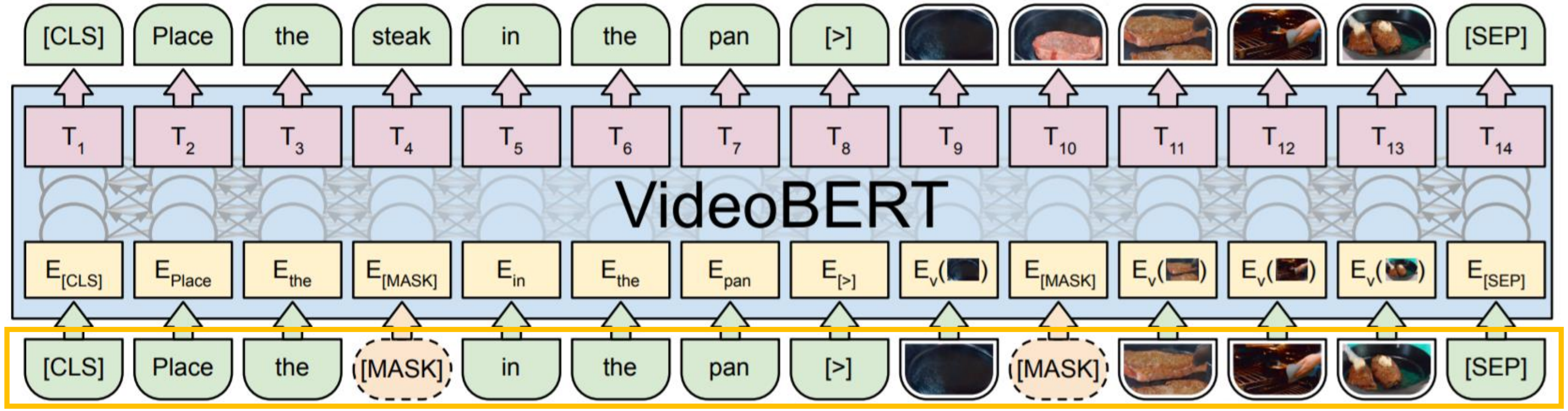
Pre-training



[Sun et al, ICCV 2019]

VideoBERT: A Joint Model for Video and Language Representation Learning

Pre-training



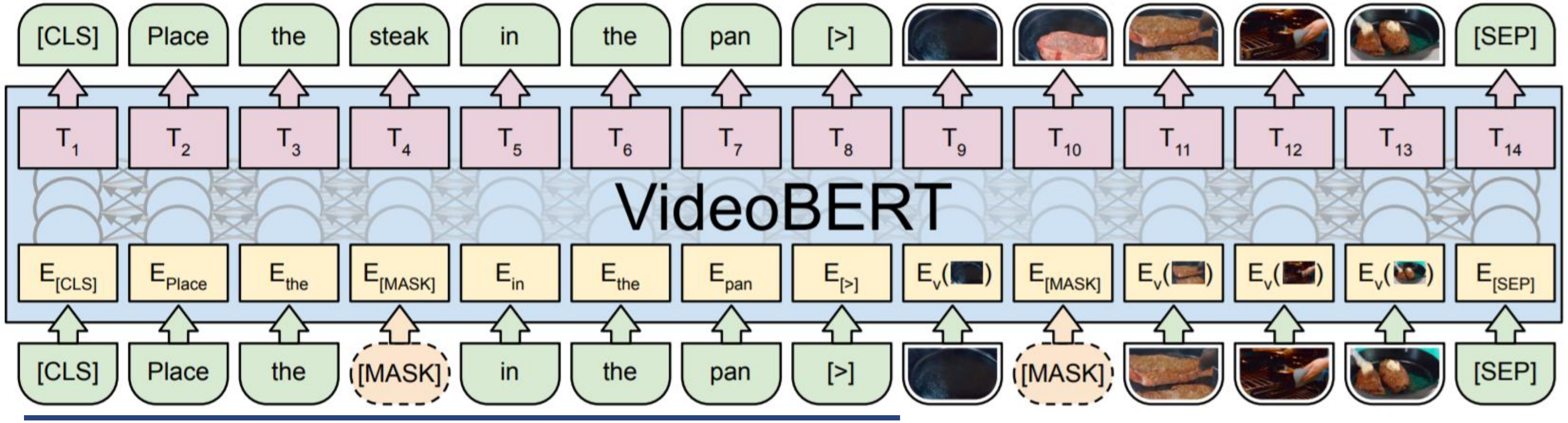
<u>Large-scale Pre-training Dataset</u>

- 312K cooking/recipe videos from YouTube

[Sun et al, ICCV 2019]

VideoBERT: A Joint Model for Video and Language Representation Learning

Pre-training



Large-scale Pre-training Dataset

- 312K cooking/recipe videos from YouTube

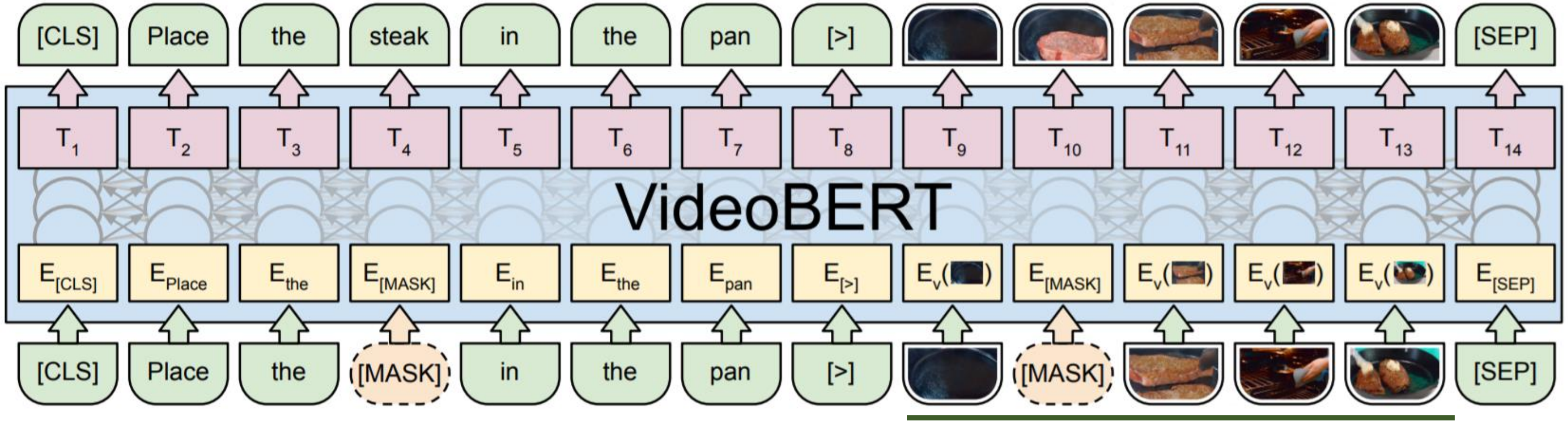
Text Representations

- Tokenized into WordPieces, following BERT

[Sun et al, ICCV 2019]

VideoBERT: A Joint Model for Video and Language Representation Learning

Pre-training



Video Representations
- 3D features from Kinetics pretrained S3D
- Tokenized into 21K clusters using hierarchical k-means

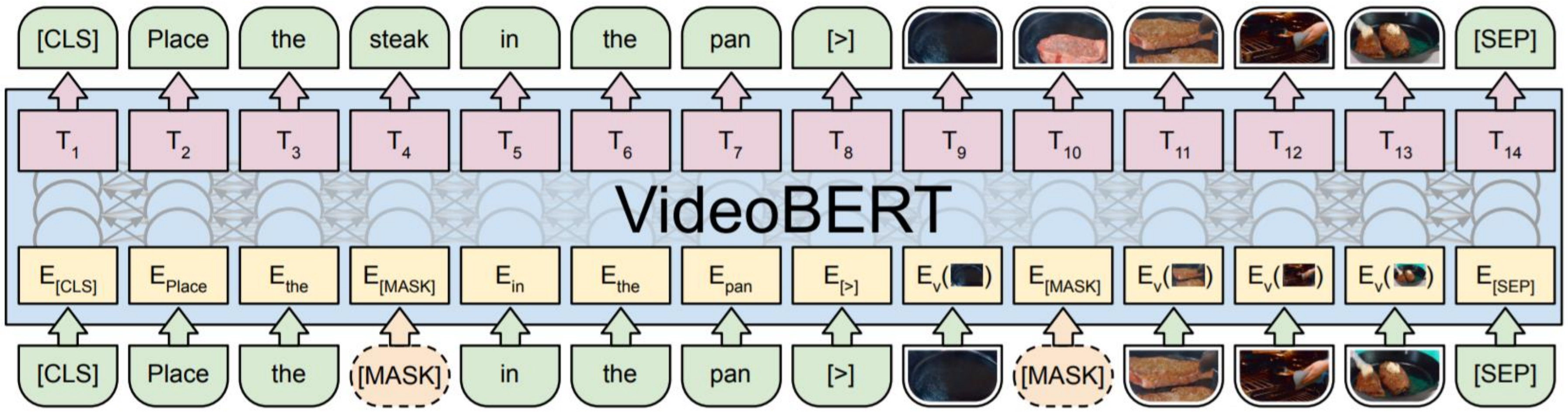
Large-scale Pre-training Dataset
- 312K cooking/recipe videos from YouTube

Text Representations
- Tokenized into WordPieces, following BERT

[Sun et al, ICCV 2019]

VideoBERT: A Joint Model for Video and Language Representation Learning

Pre-training



<u>Large-scale Pre-training Dataset</u>

- 312K cooking/recipe videos from YouTube

<u>Text Representations</u>

- Tokenized into WordPieces, following BERT

<u>Video Representations</u>

- 3D features from Kinetics pretrained S3D
- Tokenized into 21K clusters using hierarchical k-means

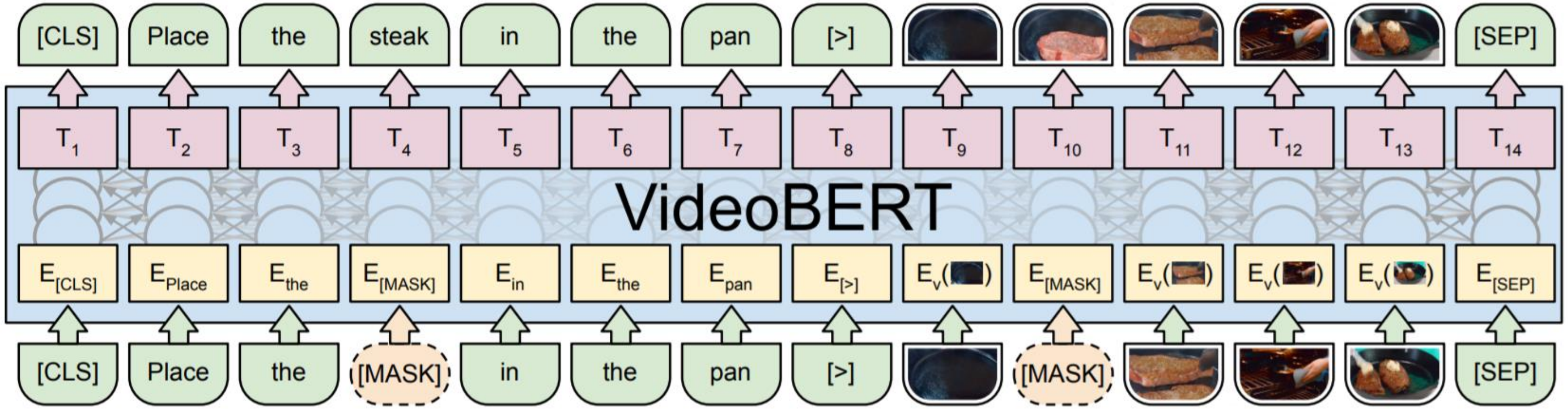
<u>Pre-training Joint Embedding</u>

- Transformer-based Video-Text encoder
- Pre-training tasks: Masked Language Modeling (MLM) + Masked Frame Modeling (MFM)

[Sun et al, ICCV 2019]

VideoBERT: A Joint Model for Video and Language Representation Learning

Pre-training



Downstream Tasks

Captioning

Now, let's [MASK] the [MASK] to the [MASK] and [MASK] the [MASK].

Now, let's place the tomatoes to the cutting board and slice the tomatoes.

Zero-shot Action classification

Now, let's show you how to [MASK] the [MASK].

Top Verbs: make, assemble, prepare
Top Nouns: pizza, sauce, pasta

[Sun et al, ICCV 2019]

VideoBERT: A Joint Model for Video and Language Representation Learning

Model	Verb top-5	Object top-5
Fully-supervised Method [1]	46.9	30.9
VideoBERT (Zero-Shot)	43.3	33.7

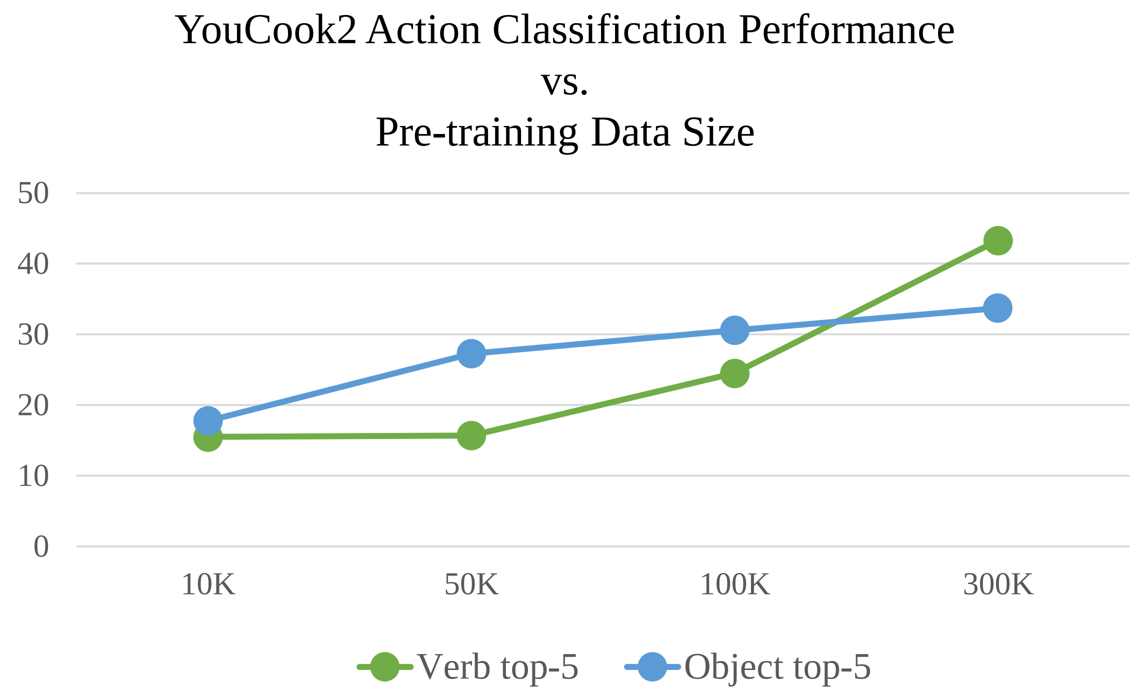
YouCook2 Action Classification

- ❖ VideoBERT (Zore-Shot) performs competitively to supervised method

Model	BLEU-4	METEOR	ROUGE-L	CIDEr
SOTA w/o PT [2]	3.84	11.55	27.44	0.38
VideoBERT	4.04	11.01	27.50	0.49
VideoBERT + S3D	4.33	11.94	28.80	0.55

YouCook2 Captioning

- ❖ VideoBERT outperforms SOTA
- ❖ Adding S3D features to visual tokens further boosts performance



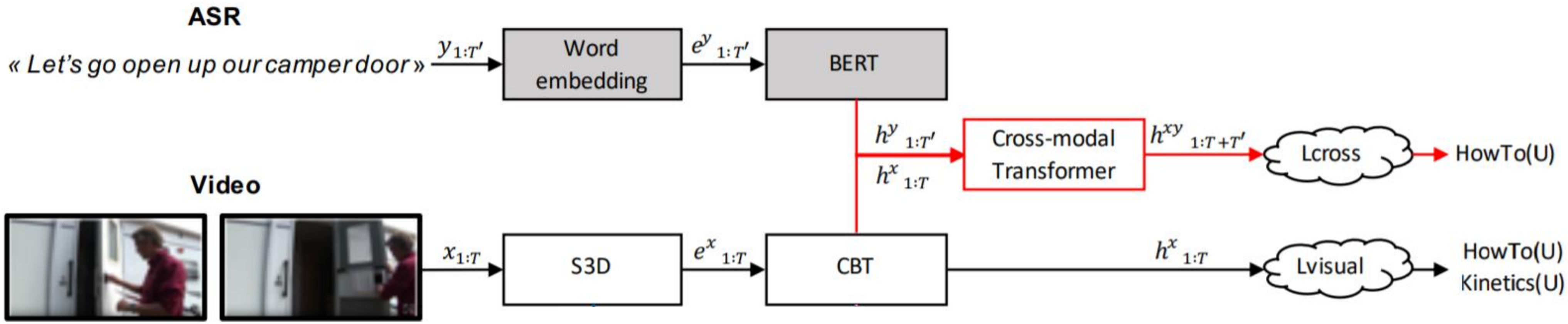
- ❖ Adding more data generally gives better results

[1] Xie, Saining, et al. "Rethinking spatiotemporal feature learning for video understanding." ECCV 2018

[2] Zhou, Luowei, et al. "End-to-end dense video captioning with masked transformer." CVPR 2018

CBT: Learning Video Representations using Contrastive Bidirectional Transformer

Pre-training



Large-scale Pre-training Dataset
- HowTo100M

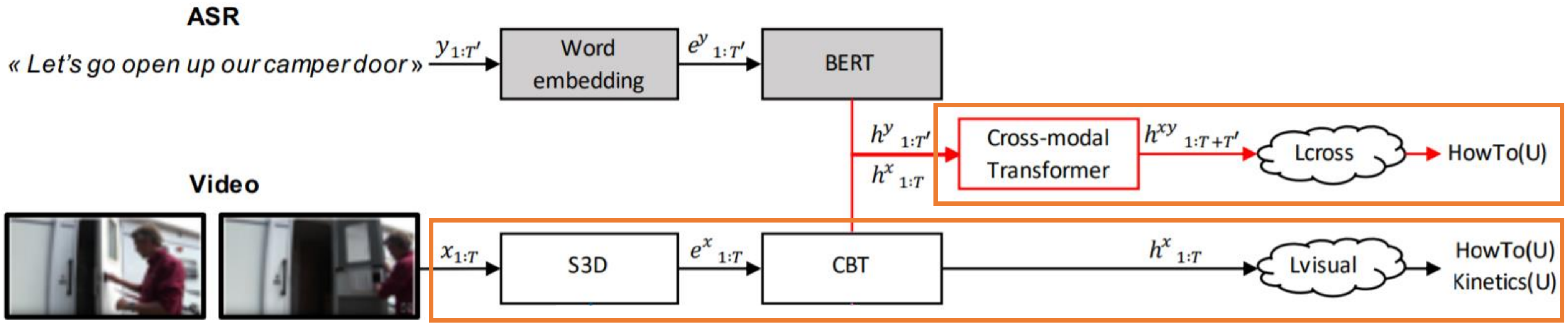
Text Representations
- Tokenized into WordPieces, following BERT

Video Representations
- 3D features from Kinetics pretrained S3D

[Sun et al, 2019]

CBT: Learning Video Representations using Contrastive Bidirectional Transformer

Pre-training



Large-scale Pre-training Dataset
- HowTo100M

Text Representations
- Extract contextualized word embeddings from BERT

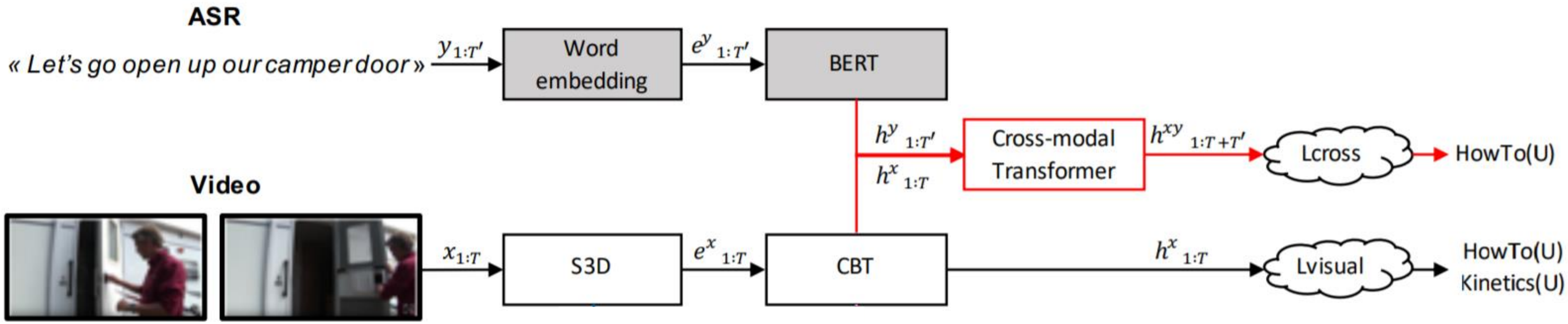
Video Representations
- 3D features from Kinetics pretrained S3D

Pre-training for Better Video Representations
- 3 Transformers: BERT, CBT and Cross-modal Transformer
- Pre-train through Noise Contrastive Estimation (NCE)
  - Video-only Pre-training (end-to-end)
  - Video-Text Alignment (fixed S3D and BERT)

[Sun et al, 2019]

CBT: Learning Video Representations using Contrastive Bidirectional Transformer

Pre-training



Downstream Tasks

Captioning

Now, let's place the tomatoes to the cutting board and slice the tomatoes.

Action/Video classification

Preparing Pizza

Video Segmentation

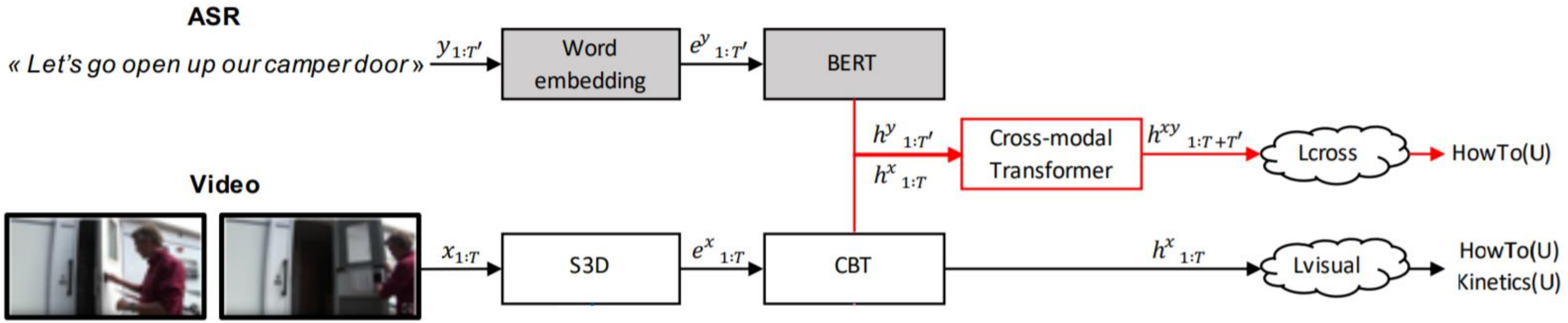
Segment #1

Segment #2

[Sun et al, 2019]

CBT: Learning Video Representations using Contrastive Bidirectional Transformer

Pre-training



Downstream Tasks

Captioning

Now, let's place the tomatoes to the cutting board and slice the tomatoes.

Action/Video classification

Preparing Pizza

Video Segmentation

Segment #1

Segment #2

[Sun et al, 2019]

CBT: Learning Video Representations using Contrastive Bidirectional Transformer

Model	BLEU-4	METEOR	ROUGE-L	CIDEr
SOTA w/o PT [1]	4.38	11.55	27.44	0.38
S3D	3.24	9.52	26.09	0.31
VideoBERT + S3D	4.33	11.94	28.80	0.55
CBT	<u>5.12</u>	<u>12.97</u>	<u>30.44</u>	<u>0.64</u>

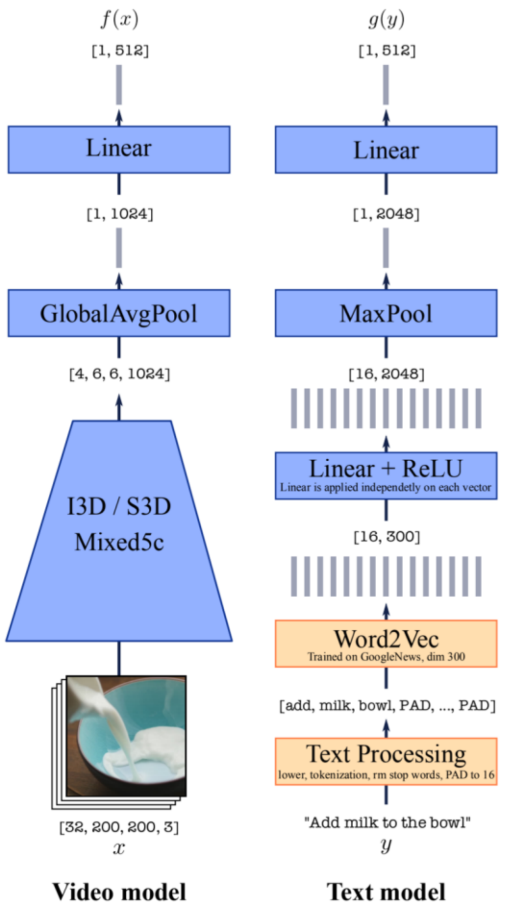
<u>YouCook2 Captioning</u>

- CBT achieves the new state of the art, as contrastive learning encourages better video representations

[1] Zhou, Luowei, et al. "End-to-end dense video captioning with masked transformer." CVPR 2018

MIL-NCE: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Pre-training



[Miech et al, CVPR 2020]

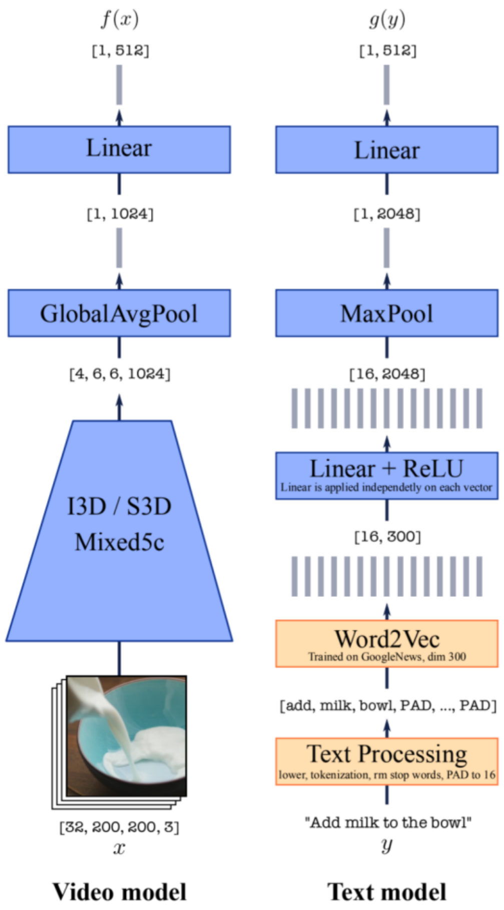
Large-scale Pre-training Dataset
- HowTo100M

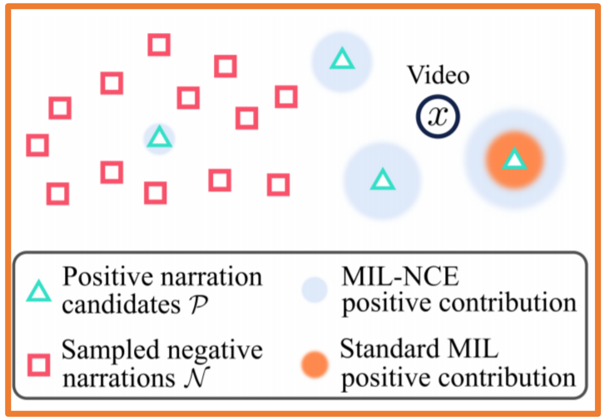
Video Representations
- 3D features from I3D/S3D

Text Representations
- GoogleNews pre-trained word2vec embeddings

MIL-NCE: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Pre-training





[Miech et al, CVPR 2020]

Large-scale Pre-training Dataset
- HowTo100M

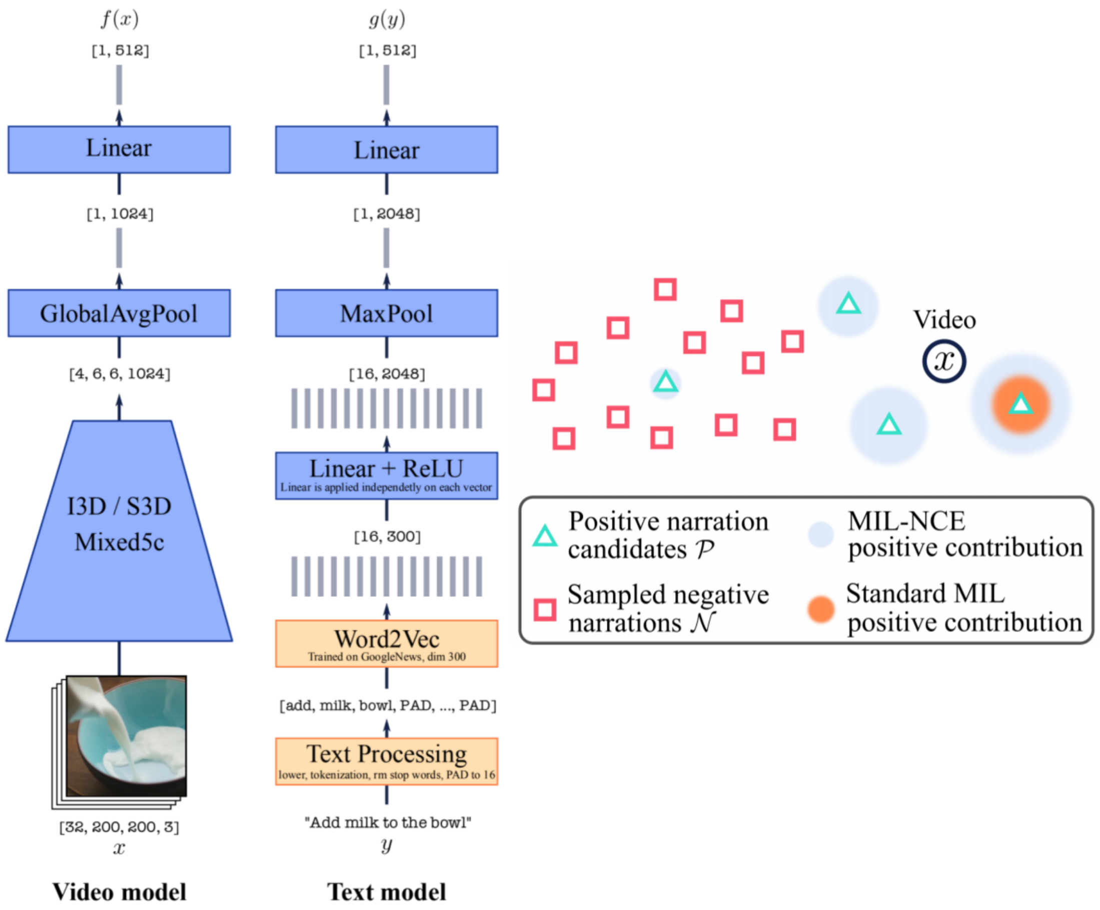
Video Representations
- 3D features from I3D/S3D

Text Representations
- GoogleNews pre-trained word2vec embeddings

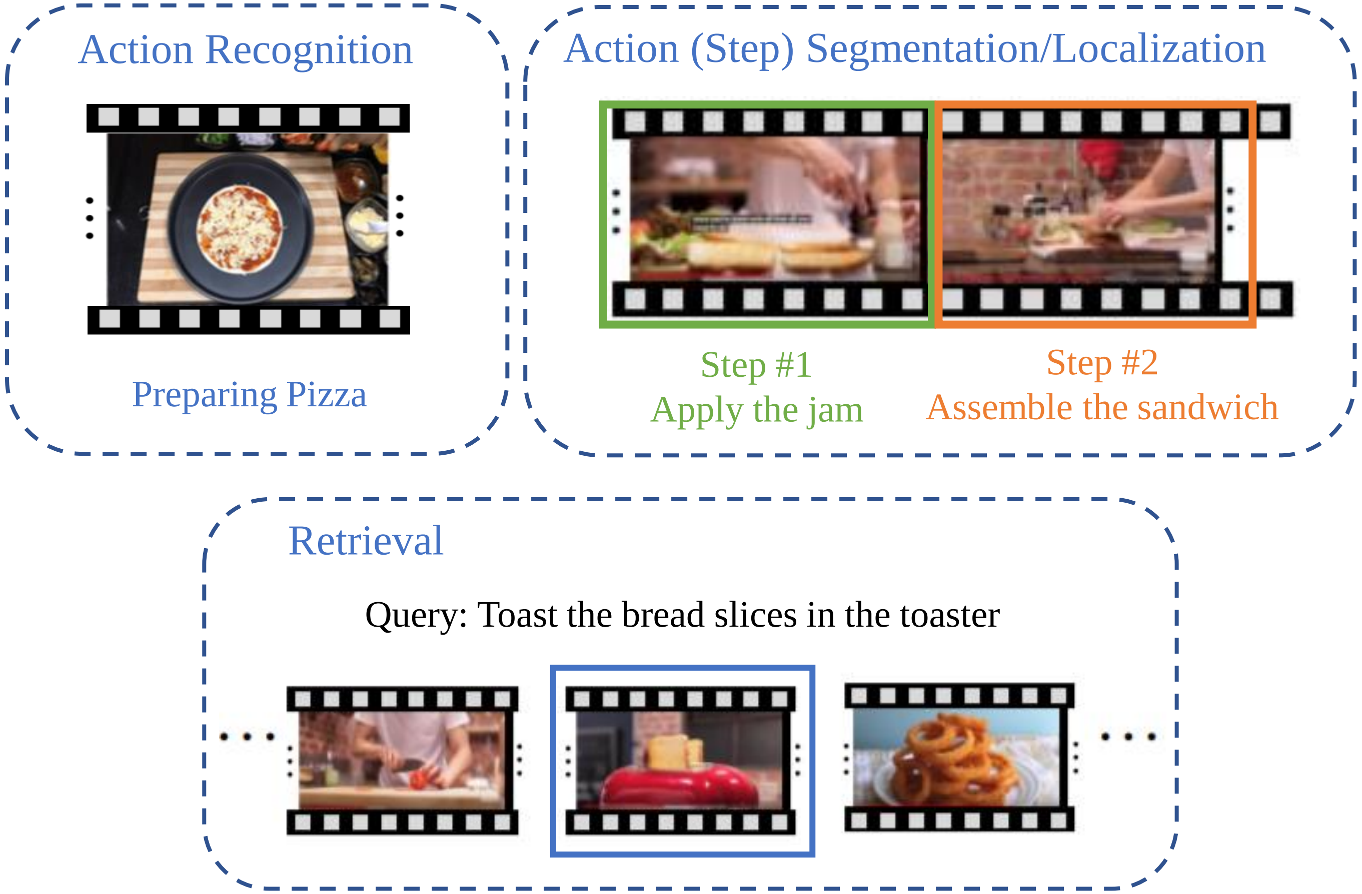
Pre-training Joint Embedding
- MIL-NCE pre-training
  - Multiple Instance Learning (MIL)
  - Noise Contrastive Estimation (NCE)

MIL-NCE: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Pre-training



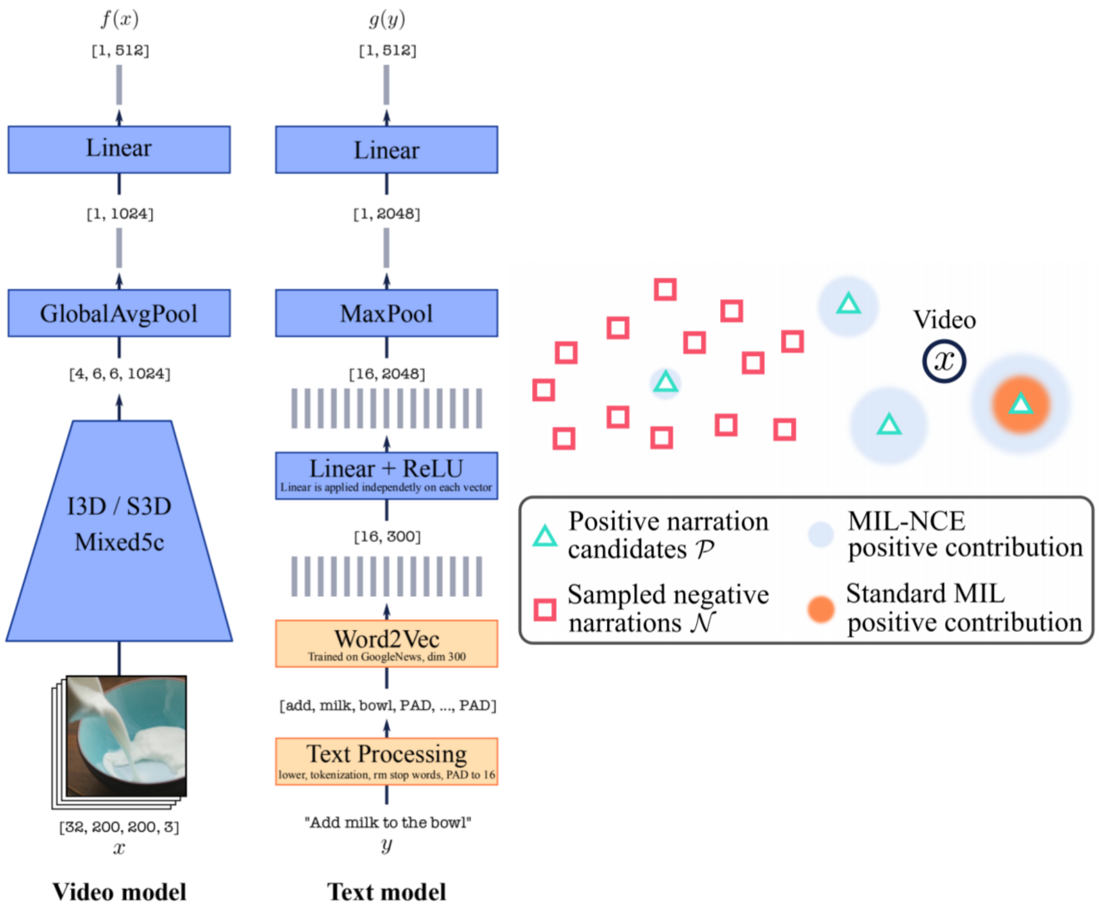
Downstream Tasks



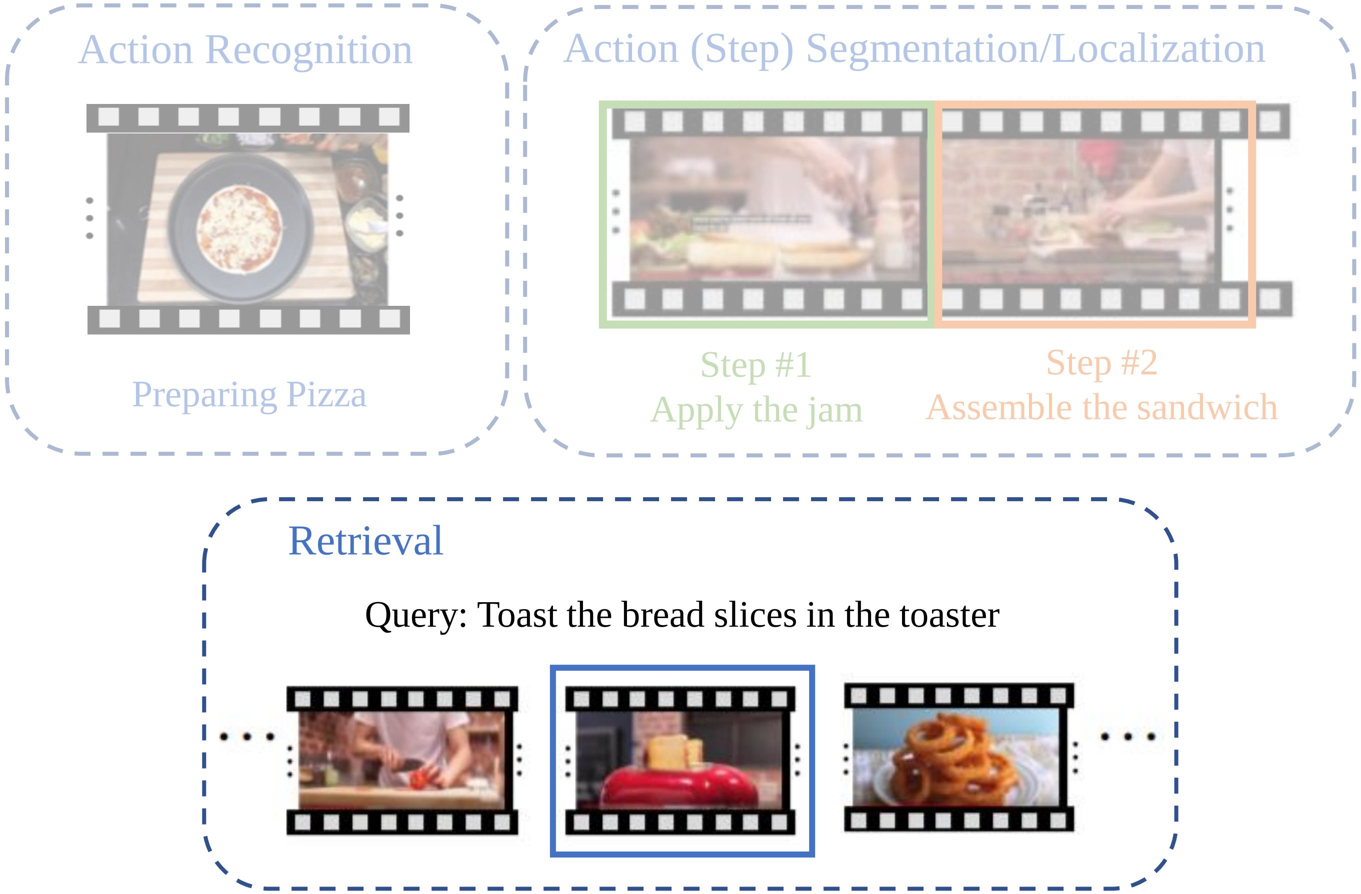
[Miech et al, CVPR 2020]

MIL-NCE: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Pre-training



Downstream Tasks



[Miech et al, CVPR 2020]

MIL-NCE: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Model	Labeled Dataset Used	YouCook2 (Median R)	MSRVTT (Median R)
HowTo100M	ImageNet + Kinetics400	46	38
	ImageNet + Kinetics400 + YouCook2	24	-
MIL-NCE	None	16	35

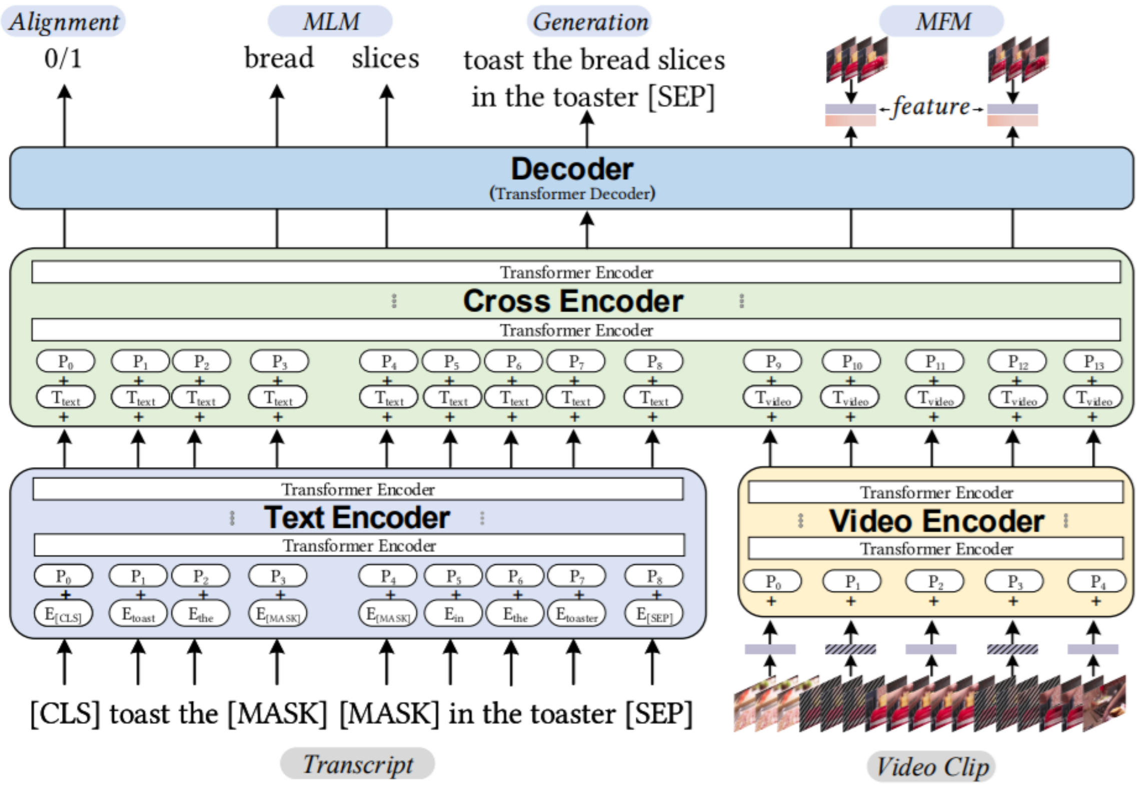
Zero-shot Clip Retrieval

❖ *On both datasets*, MIL-NCE improves over HowTo100M without using any labeled data

❖ *On YouCook2*, MIL-NCE even surpasses supervised HowTo100M model

UniViLM: a Unified Video and Language pre-training Model for multimodal understanding and generation

Pre-training



Large-scale Pre-training Dataset
- 380K videos from HowTo100M
- All food domain related videos

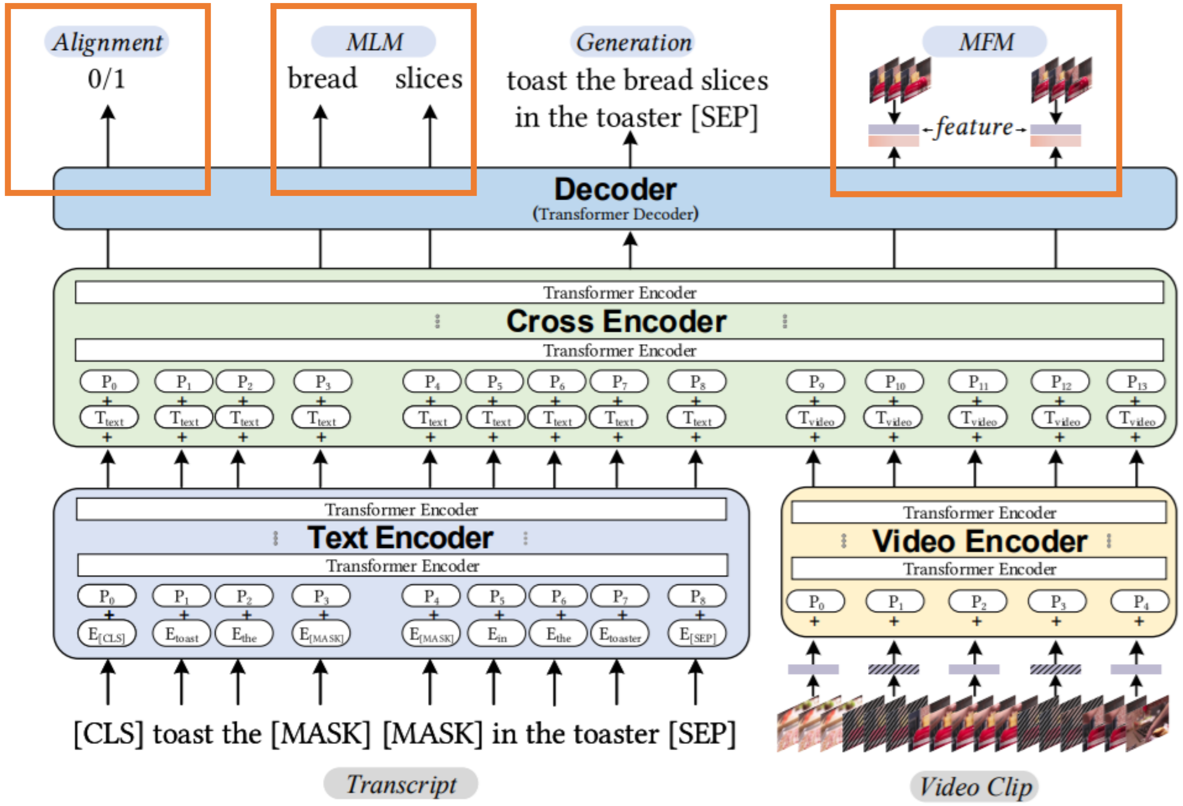
Video Representations
- 2D features from ImageNet pre-trained ResNet-152
- 3D features from Kinetics pre-trained ResNeXt-101

Text Representations
- Tokenized into WordPieces, following BERT

[Luo et al, 2020]

UniViLM: a Unified Video and Language pre-training Model for multimodal understanding and generation

Pre-training



Large-scale Pre-training Dataset
- 380K videos from HowTo100M
- All food domain related videos

Video Representations
- 2D features from ImageNet pre-trained ResNet-152
- 3D features from Kinetics pre-trained ResNeXt-101

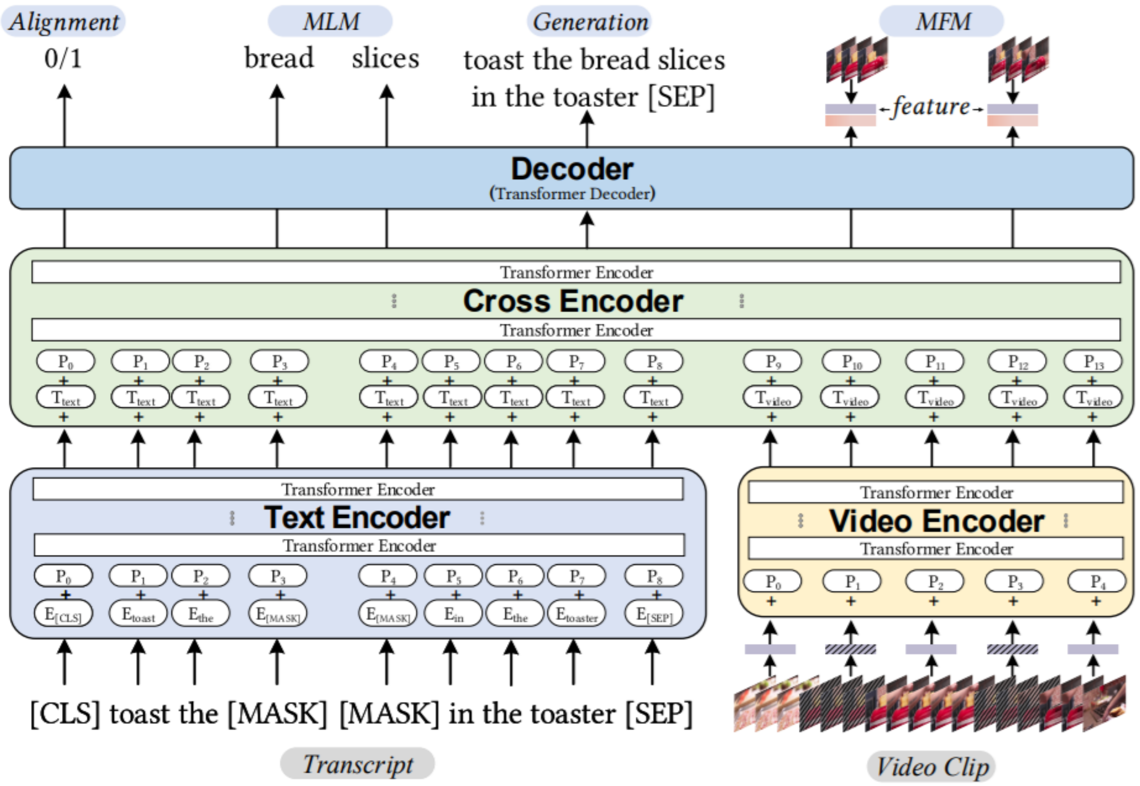
Text Representations
- Tokenized into WordPieces, following BERT

Pre-training Joint Embedding
- Pre-training tasks: MLM + MFM + Video-Text Alignment

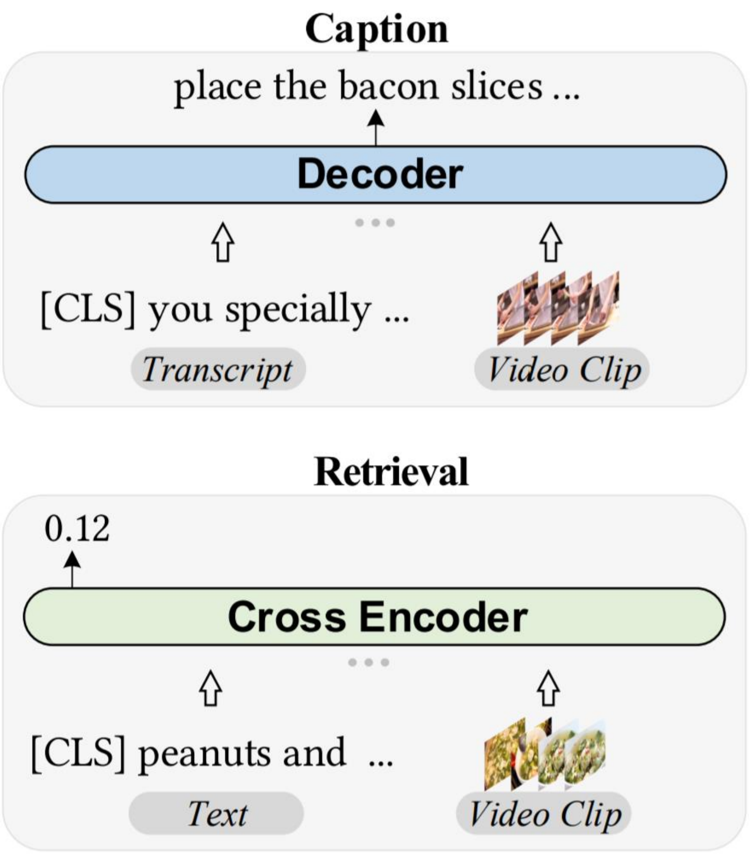
[Luo et al, 2020]

UniViLM: a Unified Video and Language pre-training Model for multimodal understanding and generation

Pre-training



Downstream Tasks



[Luo et al, 2020]

UniViLM: a **Uni**fied **Vi**deo and **L**anguage pre-training **M**odel for multimodal understanding and generation

Model	Pre-training Data Size	YouCook2 (Median R)	MSRVTT (Median R)
HowTo100M	1.2M	24	9
	380K	25	16
UniViLM	380K	20	9

Clip Retrieval

- *On YouCook2 (in-domain)*, UniViLM improves over HowTo100M with less pre-training data
- *On MSRVTT (out-of-domain)*, UniViLM surpasses HowTo100M with the same amount of pre-training data

YouCook2 Captioning

- UniViLM w/o pre-training achieves worse performance
- UniViLM w/ pre-training slightly outperforms SOTA

Model	Pre-training Data Size	BLEU-4	METEOR	ROUGE-L	CIDEr
SOTA [1]	0	9.01	17.77	36.65	1.12
UniViLM	0	8.67	15.38	35.18	1.00
	380K	10.42	16.93	38.04	1.20

[1] Shi, Botian, et al. "Dense procedure captioning in narrated instructional videos." ACL 2019

Conclusion

- Video + Language Pre-training is still at its early stage
  - Video + Language inputs are directly concatenated, losing the temporal alignment
  - Pre-training tasks directly borrowed from Image + Text Pre-training
  - Pre-training datasets limited to narrated instructional videos from YouTube
- Video + Language downstream tasks are relatively “simple”
  - Mostly focus on visual clues only
  - Subtitles/Narrations contain a lot of information, but usually discarded

Thank you!
Any questions?