Self-supervised Learning for Vision-and-Language

Licheng Yu, Yen-Chun Chen, Linjie Li
Nowadays Machine Learning
Nowadays Machine Learning
Datasets + Labels

- **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]

**Jigsaw puzzles**

[Noroozi et al. ECCV 2016]

**Image Inpainting**

[Pathak et al. CVPR 2016]

**Relative Location Prediction**

[Doersch et al. ICCV 2015]
Datasets + Labels: Self-Supervised Learning for Vision

MOCO; He et al, 2019

CPC; Ord et al, 2019

CMC; Tian 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

Large, Noisy, Cheap Data

Model

Pre-training Task I
Pre-training Task II
Pre-training Task III

Fine-tune on Downstream Task

Small, Clean, Labeled Data
Two-Stage Training Pipeline

Large, Noisy, Cheap Data

Pre-training Task I
Pre-training Task II
Pre-training Task III

Fine-tune on Downstream Task

Small, Clean, Labeled Data
Generalization

Large, Noisy, Cheap Data

Pre-training Task I
Pre-training Task II
Pre-training Task III

Model I
Model II
Model III
Model IV
Model V
Model VI
Model VII
Model VIII
Model IX
Downstream Tasks
- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expressions
- Image-Text Retrieval
- Image Captioning

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

Bluebird Farm Alpacas

The alpaca was actually walking me, and I'm okay with that.

#neverstopexploring #newyork
#alpaca #positivevibes #teamcozy
#shecozy #citylimitless
#portraitphotography #portrait
#vacationmode

144 likes
5 HOURS AGO
Common Pre-training Data for Vision + Language

<table>
<thead>
<tr>
<th>Split</th>
<th>COCO Captions</th>
<th>VG Dense Captions</th>
<th>Conceptual Captions</th>
<th>SBU Captions</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>533K (106K)</td>
<td>5.06M (101K)</td>
<td>3.0M (3.0M)</td>
<td>990K (990K)</td>
</tr>
<tr>
<td>val</td>
<td>25K (5K)</td>
<td>106K (2.1K)</td>
<td>14K (14K)</td>
<td>10K (10K)</td>
</tr>
</tbody>
</table>

**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

[Anderson et al, CVPR 2018]  
[Ren et al, NeurIPS 2015]  
[Jiang et al, CVPR 2020]
Model Architecture
Model Architecture:

(a) Single-stream Model.

(b) Two-stream Model.

[Behand the Scene; Cao et al 2020]
Model Architecture:

[Behand the Scene; Cao et al 2020]
Model Architecture:

(a) Single-stream Model.

(b) Two-stream Model.

[Behand the Scene; Cao et al 2020]
Single-Stream Architecture

Transformer

man with his dog on a couch

[UNITER; Chen et al 2019]
Single-Stream Architecture

[UNITER; Chen et al 2019]
Single-Stream Architecture

[UNITER; Chen et al 2019]
Pre-training Tasks
Pretraining Tasks

Masked Language Modeling (MLM)

[uniter; Chen et al 2019]
Pretraining Tasks

Masked Language Modeling (MLM)

Masked Region Modeling (MRM)
Pretraining Tasks

UNITER Model

Transformer

Image Embedder

Image Feature

LN

FC

FC

R-CNN

Location

Text Embedder

Text Feature

LN

Emb

Emb

Token

Position

Masked Language Modeling (MLM)

UNITER

man with his [MASK]...

dog

Masked Region Modeling (MRM)

UNITER

man with his dog on a couch...

Image-Text Matching (ITM)

UNITER

[CLS] the bus is...

0
Pretraining Tasks

Masked Language Modeling (MLM)

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 Masked Language Modeling (MLM):

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

Masked Region Modeling (MRM)

- **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 Masked Region Modeling:**

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

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

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

UNITER

$g_\theta(v_m^{(i)}) \in \mathbb{R}^K$
$c(v_m^{(i)}) \in \mathbb{R}^K$

Masked Region Modeling (MRM)

Image Regions: $v = \{v_1, \ldots, v_K\}$
Sentence Tokens: $w = \{w_1, \ldots, w_T\}$
Masking Indices: $m \in \mathbb{N}^M$

Loss Function of Masked Region Modeling:
$$\mathcal{L}_{MRM}(\theta) = E_{(w,v) \sim D} f_\theta(v_m|v_{\backslash m}, w).$$

2) Objective of Masked Region Classification (MRC)
$$f_\theta(v_m|v_{\backslash m}, w) = \sum_{i=1}^M \text{CE}(c(v_m^{(i)}), g_\theta(v_m^{(i)}))$$
Pretraining Tasks

**UNITER**

Masked Region Modeling (MRM)

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 **Masked Region Modeling**:

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

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

\[
f_\theta(\mathbf{v}_m | \mathbf{v}_{\setminus m}, \mathbf{w}) = \sum_{i=1}^{M} D_{KL}(\tilde{c}(\mathbf{v}_m^{(i)}) \| g_\theta(\mathbf{v}_m^{(i)}))
\]
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)

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

What color are her eyes?

black
Downstream Task 2: Visual Entailment

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

Premise + Hypothesis = Answer

- Entailment
- Neutral
- Contradiction

[Xie et al., 2019]
Downstream Task 2: Visual Entailment

Two women 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

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

true
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.

- He is telling ... [CLS]
- He just told ... [CLS]
- He is feeling ... [CLS]
- He is giving ... [CLS]
Downstream Task 5: Referring Expression Comprehension

woman washing dishes

[Kazemzadeh et al., EMNLP 2014]
Downstream Task 5: Referring Expression Comprehension

UNITER

woman washing dishes

× × √ ×
Downstream Task 6: Image-Text Retrieval

“a girl with a cat on grass”
Downstream Task 6: Image-Text Retrieval

“a girl with a cat on grass”

Image 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”

Text DB
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

Conventional Batching

Dynamic Batching

Saved computation
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

<table>
<thead>
<tr>
<th></th>
<th>Fp-16</th>
<th>Fp-32</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>Fast</td>
<td>Slow</td>
</tr>
<tr>
<td>Memory</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Numerical Stability</td>
<td>Bad</td>
<td>Good</td>
</tr>
</tbody>
</table>

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
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:** Vision-And-Language Understanding Evaluation
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, …)
Modality Probing

- Visual Probing
- Linguistic Probing
- Cross-Modality Probing
  - Multimodal fusion degree
  - Modality importance
  - Visual coreference
VALUE:
Vision-And-Language Understanding Evaluation

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: Object-Semantics Aligned Pre-training

$$x \triangleq [ w, q, v ] = [ w, q', v ] \triangleq x'$$

[OSCAR; Li et al., 2020]
VILLA: Vision-and-Language Large-scale Adversarial training

[Image of a dog lying on the grass next to a frisbee with boxes indicating word embedding, regional feature, and adversarial perturbation.]

Adversarial Pre-training:
- Masked Language Modeling (MLM)
- Image-Text Matching (ITM)
- ...

Adversarial Finetuning:
- VQA
- VCR
- NLVR2
- Visual Entailment
- Referring Expression Comprehension
- Image-Text Retrieval
- ...

[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.

<table>
<thead>
<tr>
<th>Method</th>
<th>VQA test-dev</th>
<th>VQA test-std</th>
<th>VCR Q→A</th>
<th>VCR QA→R</th>
<th>VCR Q→AR</th>
<th>NLVR2 dev</th>
<th>NLVR2 test-P</th>
<th>NLVR2 val</th>
<th>NLVR2 test</th>
<th>SNLI-VE val</th>
<th>SNLI-VE test</th>
</tr>
</thead>
<tbody>
<tr>
<td>VL-BERT_LARGE</td>
<td>71.79</td>
<td>72.22</td>
<td>75.5 (75.8)</td>
<td>77.9 (78.4)</td>
<td>58.9 (59.7)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Oscar_LARGE</td>
<td>73.61</td>
<td>73.82</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>79.12</td>
<td>80.37</td>
<td>-</td>
<td>-</td>
<td>79.39</td>
<td>79.38</td>
</tr>
<tr>
<td>UNITER_LARGE</td>
<td>73.82</td>
<td>74.02</td>
<td>77.22 (77.3)</td>
<td>80.49 (80.8)</td>
<td>62.59 (62.8)</td>
<td>79.12</td>
<td>79.98</td>
<td>79.39</td>
<td>79.38</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VILLA_LARGE</td>
<td>74.69</td>
<td>74.87</td>
<td>78.45 (78.9)</td>
<td>82.57 (82.8)</td>
<td>65.18 (65.7)</td>
<td>79.76</td>
<td>81.47</td>
<td>80.18</td>
<td>80.02</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

(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

*: 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
Video + Language Pre-training

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

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.

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

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

[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

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

Text Representations
• GoogleNews pre-trained word2vec embedding models

[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

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

[Miech et al, ICCV 2019]
**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

<table>
<thead>
<tr>
<th>Model</th>
<th>CrossTask (Averaged Recall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully-supervised Upper-bound [1]</td>
<td>31.6</td>
</tr>
<tr>
<td>HowTo100M PT only (weakly supervised)</td>
<td>33.6</td>
</tr>
</tbody>
</table>

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>CrossTask (Averaged Recall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully-supervised Upper-bound [1]</td>
<td>31.6</td>
</tr>
<tr>
<td>HowTo100M PT only (weakly supervised)</td>
<td>33.6</td>
</tr>
</tbody>
</table>

**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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>CrossTask (Averaged Recall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully-supervised Upper-bound [1]</td>
<td>31.6</td>
</tr>
<tr>
<td>HowTo100M PT only (weakly supervised)</td>
<td>33.6</td>
</tr>
</tbody>
</table>

**Step Localization**
- HowTo100M PT is better than training a fully supervised model on a small training set

<table>
<thead>
<tr>
<th>Model</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSMDC</td>
<td>25</td>
</tr>
<tr>
<td>YouCook2</td>
<td>21.5</td>
</tr>
<tr>
<td>MSRVTT</td>
<td>48</td>
</tr>
<tr>
<td>HowTo100M PT</td>
<td>52.8</td>
</tr>
</tbody>
</table>

**Clip Retrival**
- 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

**VideoBERT**: A Joint Model for Video and Language Representation Learning

Pre-training
**VideoBERT**: A Joint Model for Video and Language Representation Learning

Pre-training

**Large-scale Pre-training Dataset**
- 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

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

Text Representations
- Tokenized into WordPieces, following BERT

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

[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

- **Text Representations**
  - Tokenized into WordPieces, following BERT

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

---

[Sun et al, ICCV 2019]
**VideoBERT**: A Joint Model for Video and Language Representation Learning

Pre-training

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Verb top-5</th>
<th>Object top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fully-supervised Method [1]</td>
<td>46.9</td>
<td>30.9</td>
</tr>
<tr>
<td>VideoBERT (Zero-Shot)</td>
<td>43.3</td>
<td>33.7</td>
</tr>
</tbody>
</table>

**YouCook2 Action Classification**
- VideoBERT (Zero-Shot) performs competitively to supervised method

<table>
<thead>
<tr>
<th>Model</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA w/o PT [2]</td>
<td>3.84</td>
<td>11.55</td>
<td>27.44</td>
<td>0.38</td>
</tr>
<tr>
<td>VideoBERT</td>
<td>4.04</td>
<td>11.01</td>
<td>27.50</td>
<td>0.49</td>
</tr>
<tr>
<td>VideoBERT + S3D</td>
<td>4.33</td>
<td>11.94</td>
<td>28.80</td>
<td>0.55</td>
</tr>
</tbody>
</table>

**YouCook2 Captioning**
- VideoBERT outperforms SOTA
- Adding S3D features to visual tokens further boosts performance

---

CBT: Learning Video Representations using Contrastive Bidirectional Transformer

Pre-training

Large-scale Pre-training Dataset
- HowTo100M

Video Representations
- 3D features from Kinetics pretrained S3D

Text Representations
- Tokenized into WordPieces, following BERT

[Sun et al, 2019]
CBT: Learning Video Representations using Contrastive Bidirectional Transformer

Pre-training

Large-scale Pre-training Dataset
- HowTo100M

Video Representations
- 3D features from Kinetics pretrained S3D

Text Representations
- Extract contextualized word embeddings from BERT

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**

**Video**

```
<table>
<thead>
<tr>
<th>ASR</th>
</tr>
</thead>
<tbody>
<tr>
<td>« Let’s go open up our camper door »</td>
</tr>
</tbody>
</table>
```

**Word embedding**

```
\[ y_{1:T} \]
```

**BERT**

**Cross-modal Transformer**

```
\[ h_{1:T} \]
```

**Lvisual**

**HowTo(U)**

**HowTo(U) Kinetics(U)**

---

**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**

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

*YouCook2 Captioning*

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

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

### Pre-training

**Video Representations**
- 3D features from I3D/S3D

**Text Representations**
- GoogleNews pre-trained word2vec embeddings

### Large-scale Pre-training Dataset
- HowTo100M
**MIL-NCE**: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

**Pre-training**

**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)

[Miech et al, CVPR 2020]
**MIL-NCE**: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

**Pre-training**

- Linear
- GlobalAvgPool
- I3D / Si3D Mixed5c
- \( f(x) \) [1, 1024]
- \( s(y) \) [1, 1024]
- MaxPool
- \([4, 0, 0, 1, 128]\)
- \([16, 1024]\)
- Linear + RelU
- \([18, 1020]\)
- Word2Vec
- \([36, 800, 800, 5]\; x\)

**Downstream Tasks**

- **Action Recognition**
  - Preparing Pizza
  - Step #1: Apply the jam
  - Step #2: Assemble the sandwich

- **Action (Step) Segmentation/Localization**
  - Pre-training: [Miech et al, CVPR 2020]

**Retrieval**

Query: Toast the bread slices in the toaster
MIL-NCE: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

Pre-training

Downstream Tasks

Action Recognition

Preparing Pizza

Action (Step) Segmentation/Localization

Step #1
Apply the jam

Step #2
Assemble the sandwich

Retrieval

Query: Toast the bread slices in the toaster

Miech et al., CVPR 2020
**MIL-NCE**: End-to-End Learning of Visual Representations from Uncurated Instructional Videos

<table>
<thead>
<tr>
<th>Model</th>
<th>Labeled Dataset Used</th>
<th>YouCook2 (Median R)</th>
<th>MSRVTT (Median R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HowTo100M</td>
<td>ImageNet + Kinetics400</td>
<td>46</td>
<td>38</td>
</tr>
<tr>
<td>MIL-NCE</td>
<td>ImageNet + Kinetics400 + YouCook2</td>
<td>24</td>
<td>-</td>
</tr>
<tr>
<td>MIL-NCE</td>
<td>None</td>
<td>16</td>
<td>35</td>
</tr>
</tbody>
</table>

*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

- **Alignment**: 0/1
- **MLM**: bread slices
- **Generation**: toast the bread slices in the toaster [SEP]

**Text Representations**
- Tokenized into WordPieces, following BERT

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

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

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

[Lu et al, 2020]
**UniViLM**: a Unified Video and Language pre-training Model for multimodal understanding and generation

**Pre-training**

**Downstream Tasks**

- **Caption**
  - place the bacon slices ...
  - [CLS] you specially ...

- **Retrieval**
  - [CLS] peanuts and ...

[Luo et al, 2020]
**UniViLM**: a Unified Video and Language pre-training Model for multimodal understanding and generation

<table>
<thead>
<tr>
<th>Model</th>
<th>Pre-training Data Size</th>
<th>YouCook2 (Median R)</th>
<th>MSRVTT (Median R)</th>
</tr>
</thead>
<tbody>
<tr>
<td>HowTo100M</td>
<td>1.2M</td>
<td>24</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>380K</td>
<td>25</td>
<td>16</td>
</tr>
<tr>
<td>UniViLM</td>
<td>380K</td>
<td>20</td>
<td>9</td>
</tr>
</tbody>
</table>

**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

<table>
<thead>
<tr>
<th>Model</th>
<th>Pre-training Data Size</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOTA [1]</td>
<td>0</td>
<td>9.01</td>
<td>17.77</td>
<td>36.65</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>8.67</td>
<td>15.38</td>
<td>35.18</td>
<td>1.00</td>
</tr>
<tr>
<td>UniViLM</td>
<td>380K</td>
<td>10.42</td>
<td>16.93</td>
<td>38.04</td>
<td>1.20</td>
</tr>
</tbody>
</table>

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?