Recent Advances in Transferable Representation Learning

Multi-modal Contextualized Language Representation

Muhao Chen, Kai-Wei Chang, Dan Roth

AAAI 2020 Tutorial
How to answer diverse visual reasoning questions?

To answer the question on the right, the model needs to:

- Identify objects (umbrella) in the image
- Implicitly ground natural language to the image (raining → umbrella)
- Infer the correct answer

Hard to learn only from one dataset!

An example from the VCR dataset
Transferable Representations

Main training objective is to predict missing words.

VisualBERT

The model projects words and image regions into the same vector space and uses multiple Transformer layers to build joint representations.

Several people [MASK] on a [MASK] in the [MASK] with [MASK].

Input consists of an image and a caption with some masked words. Such data is easy to obtain from the internet.

Several people walking on a sidewalk in the rain with umbrellas.

Unsupervised pre-training on vision and language

Is it raining outside?

a) Yes, it is snowing.

b) Yes, [person8] and [person10] are outside.

c) No, it looks to be fall.

d) Yes, it is raining heavily.

*An example from the VCR dataset*

Transfer to answering commonsense questions
man wearing white shirt is walking on sidewalk alongside other pedestrians
Challenge 1: Grounding Language in X

man wearing white shirt is walking on sidewalk alongside other pedestrians
Challenge 2: (Contextualized) Commonsense Embedding

Based on BERT (NAACL 19)

12 layers of self-attention captures association between text-text, text-image
A (potentially non-exhaustive) list of BERT with Vision Architecture

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Pre-training resource*</th>
<th>Pre-training Tasks**</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViLBERT (Lu et al., 2019)</td>
<td>Two-stream</td>
<td>CC</td>
</tr>
<tr>
<td>B2T2 (Alberti et al., 2019)</td>
<td>Single-stream</td>
<td>CC</td>
</tr>
<tr>
<td>LXMERT (Tan &amp; Bansal, 2019)</td>
<td>Two-stream</td>
<td>COCO, VG, VQA, GQA</td>
</tr>
<tr>
<td>VisualBERT (Li et al., 2019a)</td>
<td>Single-stream</td>
<td>COCO</td>
</tr>
<tr>
<td>Unicoder-VL (Li et al., 2019b)</td>
<td>Single-stream</td>
<td>CC</td>
</tr>
<tr>
<td>VL-BERT (Su et al., 2019)</td>
<td>Single-stream</td>
<td>CC</td>
</tr>
<tr>
<td>UNITER (Chen et al., 2019)</td>
<td>Single-stream</td>
<td>COCO, VG, CC, SBU</td>
</tr>
</tbody>
</table>

* CC stands for Conceptual Captions, VG stands for Visual Genome

** 1 means cross modality alignment; 2 means grounded masked LM; 3 means masked visual classification; 4 means visual regression; 5 means cross modality QA
Architectural Difference

An example of single-stream architecture: VisualBERT [Li+19]

A joint Transformer for fusing textual and visual input
Architectural Difference

Separate Transformers for text and vision at first and then a joint Transformer

An example of two-stream architecture: LXMERT [Tan+19]
Pre-training Objectives

Pre-training objectives of LXMERT

[Tan+19]
Downstream Tasks

VCR: Visual Commonsense [Zellers+18]
(e.g., actions, goals, and mental states)

Q: What is hanging above the toilet?
A: Towel

VQA: Comprehensive Visual QA
(e.g., shape, size, color, object) [Goyal+16]
Downstream Tasks

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

NLVR2: Binary sentence classification, focus on semantic diversity, compositionality, and visual reasoning [Suhr+19]

Flickr30K: locating objects given the sentence

A man with pierced ears is wearing glasses and an orange hat.
Performance Improvement

Image caption data (MSCOCO): 
~300,000 images, 5 captions per image

**VCR** 71.6 (best single model 72.6)

**VQA** 70.8 (baseline: 68.5, best ~75)

**NLVR** 67.4 (best on leaderboard: 54.1)

**Flickr30k** R@10: 86.61 (Best: 86.35)
Learning High-level Concepts

are rescuing someone that has gotten in a car accident

they are wearing yellow jackets

[SEP] obj-1 obj0 obj1 obj2 obj3 obj4 obj5 obj6 obj7 obj8 obj9

are rescuing someone that has gotten in a car accident

they are wearing yellow jackets

[SEP] obj-1 obj0 obj1 obj2 obj3 obj4 obj5 obj6 obj7 obj8 obj9
Learning High-level Concepts

[CLS]
they
are
celebrating
someone
"
"
s
birthday
"

[SEP]
obj-1
obj0
obj1
obj2
obj3
obj4
obj5

[CLS]
they
are
celebrating
someone
"
"
s
birthday
"
What does BERT with Vision Look At?

**Entity Grounding**

- Man
- Shirt
- Sidewalk
- Pedestrians
- Sidewalk*
What does BERT with Vision Look At?

- Entity grounding

1) Certain heads are accurate
2) Accuracy peaks at higher layers
What does BERT with Vision Look At?

- **Syntactic grounding**
  1. Certain heads are accurate
  2. Accuracy peaks at higher layers