Visual Question Answering and Visual Reasoning

Zhe Gan
6/15/2020
Overview

• Goal of this part of the tutorial:

  • Use VQA and visual reasoning as example tasks to understand Vision-and-Language representation learning

  • After the talk, everyone can confidently say: “yeah, I know VQA and visual reasoning pretty well now”

  • Focus on high-level intuitions, not technical details

  • Focus on static images, instead of videos

  • Focus on a selective set of papers, not a comprehensive literature review
Agenda

• Task Overview
  • What are the main tasks that are driving progress in VQA and visual reasoning?

• Method Overview
  • What are the state-of-the-art approaches and the key model design principles underlying these methods?

• Summary
  • What are the core challenges and future directions?
Agenda

• Task Overview
  • *What are the main tasks that are driving progress in VQA and visual reasoning?*

• Method Overview
  • *What are the state-of-the-art approaches and the key model design principles underlying these methods?*

• Summary
  • *What are the core challenges and future directions?*
What is V+L about?

• V+L research is about how to train a smart AI system that can see and talk
What is V+L about?

• V+L research is about how to train a smart AI system that can see and talk.

Prof. Yann LeCun’s cake theory

- Reinforcement Learning
- Supervised Learning
- Unsupervised/Self-supervised Learning

In our V+L context

- Multimodel Intelligence
- BERT
- ResNet
- Language Understanding
- Visual Understanding
Task Overview: VQA and Visual Reasoning

- Large-scale annotated datasets have driven tremendous progress in this field
Image credit: https://visualqa.org/, https://visualdialog.org/
The left image contains twice the number of dogs as the right image, and at least two dogs in total are standing.

**VizWiz**


**NLVR2**
Visual Entailment

Premise: 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.

Hypothesis: Entailment

Answer: Neutral

Contradiction

VQA-Rephrasings

2. Cycle-Consistency for Robust Visual Question Answering, CVPR 2019
GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering, CVPR 2019
TextVQA

A dataset to benchmark visual reasoning based on text in images.

Q: Which American president is associated with the stuffed animal seen here?
A: Teddy Roosevelt

Q: What is the price of the bananas per kg?
A: $11.98

Q: What does the red sign say?
A: Stop

More datasets...

SQuINTing at VQA Models:
Interrogating VQA Models with Sub-Questions

Figure 1: A potential reasoning failure: Current models answer “Yes” correctly to the Reasoning question “Is the banana ripe enough to eat?”. We might assume that correctly answering the Reasoning question stems from perceiving relevant concepts correctly – perceiving yellow bananas in this example. But when asked “Are the bananas mostly green or yellow?”, it answers “Green” incorrectly – indicating that the model possibly answered the original for the wrong reasons even if the answer was right. We quantify the extent to which this phenomenon occurs in VQA and introduce a new dataset aimed at stimulating research on well grounded reasoning.

VQA-LOL: Visual Question Answering under the Lens of Logic

Annotations from COCO
OBJECTS (B):
person, bottle, bowl, microwave, fridge, clock
CAPTIONS (C):
“a man bending over to look inside of a fridge.”
“A person standing in front of an open refrigerator.”

<table>
<thead>
<tr>
<th>Question</th>
<th>Pred. Answer</th>
<th>LXMER accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1: Is there beer?</td>
<td>YES (96.26%) NO (3.74%)</td>
<td>86.65</td>
</tr>
<tr>
<td>Q2: Is the man wearing shoes?</td>
<td>NO (90.03%) YES (9.97%)</td>
<td>50.79</td>
</tr>
<tr>
<td>Q2 \land Q1: Is the man not wearing shoes and is there beer?</td>
<td>NO (62.00%) YES (37.99%)</td>
<td>50.51</td>
</tr>
<tr>
<td>Q2 \lor Q1: Is there beer and does this seem like a man bending over to look inside of a fridge?</td>
<td>NO (100%) YES (0.00%)</td>
<td>50.51</td>
</tr>
<tr>
<td>Q1 \land antonym(B): Is there beer and is there a wine glass?</td>
<td>YES (84.37%) NO (15.60%)</td>
<td>50.51</td>
</tr>
</tbody>
</table>
Diagnostic Datasets

- CLEVR (Compositional Language and Elementary Visual Reasoning)
  - Has been extended to visual dialog (CLEVR-Dialog), referring expressions (CLEVR-Ref+), and video reasoning (CLEVRER)

Beyond VQA: Visual Grounding

- Referring Expression Comprehension: RefCOCO(+/g)
  - ReferIt Game: Referring to Objects in Photographs of Natural Scenes
- Flickr30k Entities

Beyond VQA: Visual Grounding

- PhraseCut: Language-based image segmentation

Visual Question Answering

Challenge 2016 Winner

Challenge 2017 Winner

Challenge 2018 Winner

Challenge 2019 Winner

Revisiting Grid Features for VQA

2020 VQA Challenge Winner

<table>
<thead>
<tr>
<th>Rank</th>
<th>Participant team</th>
<th>yes/no</th>
<th>number</th>
<th>other</th>
<th>overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Renaissance (StructBERT-base Ensemble)</td>
<td>90.71</td>
<td>59.80</td>
<td>66.92</td>
<td>76.01</td>
</tr>
<tr>
<td>2</td>
<td>DL-61 (BGN, ensemble)</td>
<td>90.89</td>
<td>61.13</td>
<td>66.28</td>
<td>75.92</td>
</tr>
<tr>
<td>3</td>
<td>MS D365 AI (VILLA Ensemble)</td>
<td>91.30</td>
<td>59.23</td>
<td>66.20</td>
<td>75.85</td>
</tr>
</tbody>
</table>

Image Credit: CVPR 2019 Visual Question Answering and Dialog Workshop
Agenda

• Task Overview
  • What are the main tasks that are driving progress in VQA and visual reasoning?

• Method Overview
  • What are the state-of-the-art approaches and the key model design principles underlying these methods?

• Summary
  • What are the core challenges and future directions?
Overview

• How a typical system looks like

What is she eating?

Image Feature Extraction

Multi-Modal Fusion

Answer Prediction

Hamburger
Overview

• Better image feature preparation
• Enhanced multimodal fusion
  • Bilinear pooling: how to fuse two vectors into one
  • Multimodal alignment: cross-modal attention
  • Incorporation of object relations: intra-modal self-attention, graph attention
  • Multi-step reasoning
• Neural module networks for compositional reasoning
• Robust VQA (briefly mention)
• Multimodal pre-training (briefly mention)
Better Image Feature Preparation

• From *grid* features to *region* features, and to *grid* features again
2017 VQA Challenge Winner

In Defense of Grid Features for VQA

Table 2. Evaluation of Pixel-BERT with other methods on VQA.

<table>
<thead>
<tr>
<th>Model</th>
<th>test-dev</th>
<th>test-std</th>
</tr>
</thead>
<tbody>
<tr>
<td>MUTAN[5]</td>
<td>60.17</td>
<td>-</td>
</tr>
<tr>
<td>BUTD[2]</td>
<td>65.32</td>
<td>65.67</td>
</tr>
<tr>
<td>ViLBERT[21]</td>
<td>70.55</td>
<td>70.92</td>
</tr>
<tr>
<td>VisualBERT[19]</td>
<td>70.80</td>
<td>71.00</td>
</tr>
<tr>
<td>VLBERT[29]</td>
<td>71.79</td>
<td>72.22</td>
</tr>
<tr>
<td>LXMERT[33]</td>
<td>72.42</td>
<td>72.54</td>
</tr>
<tr>
<td>UNITER[6]</td>
<td>72.27</td>
<td>72.46</td>
</tr>
<tr>
<td>Pixel-BERT (r50)</td>
<td>71.35</td>
<td>71.42</td>
</tr>
<tr>
<td>Pixel-BERT (x152)</td>
<td>74.45</td>
<td>74.55</td>
</tr>
</tbody>
</table>

Bilinear Pooling

• Instead of simple concatenation and element-wise product for fusion, bilinear pooling methods have been studied.

• Bilinear pooling and attention mechanism can be enhanced with each other.
Multimodal Compact Bilinear Pooling

2016 VQA Challenge Winner

However, the feature after FFT is very high dimensional.
Multimodal Tucker Fusion

Bilinear Super-diagonal Fusion

FiLM: Feature-wise Linear Modulation

$$
\gamma_{i,c} = f_c(x_i) \quad \beta_{i,c} = h_c(x_i), \\
FiLM(F_{i,c} | \gamma_{i,c}, \beta_{i,c}) = \gamma_{i,c} F_{i,c} + \beta_{i,c}.
$$

Something similar to conditional batch normalization

Multimodal Alignment

• Cross-modal attention:
  • Tons of work in this area
  • Early work: questions attend to image grids/regions
  • Current focus: image-text co-attention
Parallel Co-attention and Alternative Co-attention

SAN: Stacked Attention Networks for Image Question Answering, CVPR 2016
HierCoAttn: Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering, CVPR 2018

DAN: Dual Attention Network
DCN: Dense Co-attention Network

2018 VQA Challenge Runner-Up
- Multiple Glimpses
- Counter Module
- Residual Learning
- Glove Embeddings

Relational Reasoning

• Intra-modal attention
  • Recently becoming popular
  • Representing image as a graph
  • Graph Convolutional Network & Graph Attention Network
  • Self-attention used in Transformer

2016/9
2017/6
2018/6
2019/2
2019/3
2019/5

Graph-Structured
Relation Network
Graph Learner
MuRel
ReGAT
LCGN

THE UNIVERSITY of ADELAIDE
Google DeepMind
aimbrain
SORBONNE UNIVERSITÉ
Microsoft
Berkeley UNIVERSITY of CALIFORNIA
Graph-Structured Representations for Visual Question Answering

Relational Network: A fully-connected graph is constructed

$e_n = F([v_n \| q])$, \hspace{1em} n = 1, 2, ..., N

$E \in \mathbb{R}^{N \times d_e}$

$A = EE^T$ so that $A_{i,j} = e_i^T e_j$.

$N(i) = topm(a_i)$

[1] Learning Conditioned Graph Structures for Interpretable Visual Question Answering, NeurIPS 2018

Language-Conditioned Graph Networks for Relational Reasoning, ICCV 2019
• **Explicit** Relation: Semantic & Spatial relation
• **Implicit** Relation: Learned dynamically during training

[1] Relation-Aware Graph Attention Network for Visual Question Answering, ICCV 2019
[1] Relation-Aware Graph Attention Network for Visual Question Answering, ICCV 2019
MCAN: Deep Modular Co-Attention Network

- Winning entry to VQA Challenge 2019
- Similar idea also explored in DFAF, close to **V+L pre-training** models

MCAN: Deep Modular Co-Attention Network

• Winning entry to VQA Challenge 2019
• Similar idea also explored in DFAF, close to *V+L pre-training* models

MAC: Memory, Attention and Composition

- Multi-step reasoning via recurrent MAC cells, while retaining end-to-end differentiability

[1] Compositional Attention Networks for Machine Reasoning, ICLR, 2018
MAC: Memory, Attention and Composition

• Each cell maintains recurrent dual states:
  • *Control $c_i$*: the reasoning operation that should be accomplished at this step.
  • *Memory $m_i$*: the retrieved information relevant to the query, accumulated over previous iterations.
• Implementation-wise:
  • *Attention-based average* of a given query (question)
  • *Attention-based average* of a given Knowledge Base (image)

[1] Compositional Attention Networks for Machine Reasoning, ICLR, 2018
Neural State Machine

• We see and reason with concepts, not visual details, 99% of the time
• We build semantic world models to represent our environment

Neural Module Network

• All the previously mentioned work can be considered as Monolithic Network
• Design Neural Modules for compositional visual reasoning

[5] Explainable Neural Computation via Stack Neural Module Networks, ECCV 2018
Q: How many spheres are the left of the big sphere and the same color as the small rubber cylinder?

A: 1
Consider a compositional model

**Q:** How many spheres are the left of the big sphere and the same color as the small rubber cylinder?

**Q:** How many spheres are the right of the big sphere and the same color as the small rubber cylinder?

**Q:** Is the big sphere the same material as the thing on the right of the cube?

Overview of the NMN approach

Overview of the NMN approach

Uses some pre-trained parser

Trained separately

Inferring and Executing Programs

What do the modules learn?

How many other things are of the same size as the green matte ball?

Question encoder (RNN)

Question features

Layout policy (RNN)

find() 
relocate(_) 
count()

Layout prediction (reverse Polish notation)

How many other things are of the same size as the green matte ball?

Network builder

Module network

4

count

relocate

the same size as
d
find

green matte ball

Answer

Image encoder (CNN)

Image features

Robust VQA: two examples

• Overcoming language prior with adversarial regularization

Robust VQA: two examples

- Self-critical reasoning

See the right image region, but still predicts wrong

[1 Self-Critical Reasoning for Robust Visual Question Answering, NeurIPS 2019]
Agenda

• Task Overview
  • What are the main tasks that are driving progress in V+L representation learning?

• Method Overview
  • What are the state-of-the-art approaches and the key model design principles underlying these methods?

• Summary
  • What are the core challenges and future directions?
Take-away Messages

• Popular tasks:
  • VQA, GQA, VCR, RefCOCO, NLVR2, etc.

• Methods:
  • Grid vs. region features
  • Bilinear pooling and FiLM
  • Multimodal alignment with cross-modal attention
  • Relational reasoning with intra-modal attention (self-attention, graph attention)
  • Transformer model becomes popular in the field
  • Multi-step reasoning
  • Neural state machine
  • Neural module network
Challenges & Future Directions

• Can we have something like GLUE and SuperGLUE?
• Can we use a Visual Transformer to encode images to train a large V+L Transformer model end-to-end?
• Instead of Transformer, can we perform FiLM-like fusion for multi-modal pre-training?
• Since all the reasoning is performed in the embedding/neural space, it is not clear whether the model “truly” learns how to reason
• Adversarial robustness of V+L models is less explored in the current literature
Thank you!
Any Questions?