Images and Languages

CS 6384 Computer Vision
Professor Yu Xiang
The University of Texas at Dallas
Image Classification

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

Understand Images with Natural Languages

• Image captioning

• Object grounding

• Visual question answering

• Representation learning with images and languages
Image Captioning

• Automatically generate texture descriptions of images

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

the person is riding a surfboard in the ocean
A Traditional Method for Image Captioning

Image Captioning with RNNs

Image Captioning with RNNs

Image Captioning with Attentions

1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation

Image Captioning with Attentions

## Image Captioning with Attentions

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Google NIC (Vinyals et al., 2014)\textsuperscript{1\Sigma}</td>
<td>63</td>
<td>41</td>
<td>27</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear (Kiros et al., 2014a)\textsuperscript{0}</td>
<td>65.6</td>
<td>42.4</td>
<td>27.7</td>
<td>17.7</td>
<td>17.31</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>67</td>
<td>44.8</td>
<td>29.9</td>
<td>19.5</td>
<td>18.93</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>67</td>
<td>45.7</td>
<td>31.4</td>
<td>21.3</td>
<td>20.30</td>
</tr>
<tr>
<td>Flickr8k</td>
<td>Google NIC\textsuperscript{0\Sigma}</td>
<td>66.3</td>
<td>42.3</td>
<td>27.7</td>
<td>18.3</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear</td>
<td>60.0</td>
<td>38</td>
<td>25.4</td>
<td>17.1</td>
<td>16.88</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>66.7</td>
<td>43.4</td>
<td>28.8</td>
<td>19.1</td>
<td>18.49</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>66.9</td>
<td>43.9</td>
<td>29.6</td>
<td>19.9</td>
<td>18.46</td>
</tr>
<tr>
<td>Flickr30k</td>
<td>CMU/MS Research (Chen &amp; Zitnick, 2014)\textsuperscript{a}</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>20.41</td>
</tr>
<tr>
<td></td>
<td>MS Research (Fang et al., 2014)\textsuperscript{t\alpha}</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>20.71</td>
</tr>
<tr>
<td></td>
<td>BRNN (Karpathy &amp; Li, 2014)\textsuperscript{0}</td>
<td>64.2</td>
<td>45.1</td>
<td>30.4</td>
<td>20.3</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Google NIC\textsuperscript{t\alpha\Sigma}</td>
<td>66.6</td>
<td>46.1</td>
<td>32.9</td>
<td>24.6</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Log Bilinear\textsuperscript{0}</td>
<td>70.8</td>
<td>48.9</td>
<td>34.4</td>
<td>24.3</td>
<td>20.03</td>
</tr>
<tr>
<td></td>
<td>Soft-Attention</td>
<td>70.7</td>
<td>49.2</td>
<td>34.4</td>
<td>24.3</td>
<td>23.90</td>
</tr>
<tr>
<td></td>
<td>Hard-Attention</td>
<td>71.8</td>
<td>50.4</td>
<td>35.7</td>
<td>25.0</td>
<td>23.04</td>
</tr>
</tbody>
</table>

**BLEU (BiLingual Evaluation Understudy)**

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

Image Captioning with Object Detection

Object detection features \( \{v_1, \ldots, v_k\} \)

RoI pooling from Faster R-CNN

LSTM-based model

\[
\tilde{v} = \frac{1}{k} \sum_i v_i
\]

Attention

\[
a_{i,t} = w_a^T \tanh (W_{va} v_i + W_{ha} h_{t}^1) \\
\alpha_t = \text{softmax} (a_t)
\]

\[
\hat{v}_t = \sum_{i=1}^{K} \alpha_{i,t} v_i
\]

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

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

Object Grounding

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

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

```
A cat with white paws jumps over a fence in front of a yellow tree

∅ no object
```

maximum number of tokens: 256

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

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

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

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

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


- Input
  - An image
  - A free-form, open-ended, natural language question

- Output
  - Case 1: open-ended answer
  - Case 2: multiple-choice task

\[
\text{accuracy} = \min\left(\frac{\# \text{humans that provided that answer}}{3}, 1\right)
\]
Visual Question Answering

“How many horses are in this image?”

Visual Question Answering

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

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

Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. Li et al., ECCV, 2020
Classify “polluted” triplets with wrong tags

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

• Fine-tuning for image captioning

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

• Fine-tuning for question answering

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

Replace with a Question

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

- Contrastive pre-training

400 million (image, text) pairs from Internet

CLIP: Contrastive Language-Image Pre-Training

• Contrastive pre-training

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

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

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

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

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

Multi-class N-pair Loss

\[
\mathcal{L}_{N\text{-pair}}(x, x^+, \{x_i^-\}_{i=1}^{N-1}) = \log \left(1 + \sum_{i=1}^{N-1} \exp(f(x)^T f(x_i^-) - f(x)^T f(x^+))\right)
\]

Softmax for multi-class classification

\[
= -\log \frac{\exp(f(x)^T f(x^+))}{\exp(f(x)^T f(x^+)) + \sum_{i=1}^{N-1} \exp(f(x)^T f(x_i^-))}
\]

CLIP: Contrastive Language-Image Pre-Training

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

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

Summary

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

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

• Baby Talk: Understanding and Generating Image Descriptions, 2011

• Deep Visual-Semantic Alignments for Generating Image Descriptions, 2015
  https://arxiv.org/abs/1412.2306

• Show, Attend and Tell: Neural Image Caption Generation with Visual Attention, 2015

• Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering, 2018

• MDETR - Modulated Detection for End-to-End Multi-Modal Understanding, 2021
  https://arxiv.org/abs/2104.12763

• VQA: Visual Question Answering, 2015
  https://arxiv.org/abs/1505.00468

• Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks, 2020

• Learning Transferable Visual Models From Natural Language Supervision, 2021