Multimodal Learning for Image Captioning and Visual Question Answering

Xiaodong He
Deep Learning Technology Center
Microsoft Research
UC Berkeley, April 7\textsuperscript{th}, 2016
Collaborators:

Hao Fang
Saurabh Gupta
Forrest Iandola
Rupesh Srivastava
Li Deng
Piotr Dollár
Jianfeng Gao
Xiaodong He
Margaret Mitchell
John Platt
Lawrence Zitnick
Geoffrey Zweig
Jacob Devlin
Kenneth Tran
Lei Zhang
Jian Sun
Chris Buehler
Chris Thrasher
Chris Sienkiewicz
Cornelia Carapcea
Yuxiao Hu
Yandong Guo
Zichao Yang
Alex Smola
...

...
Outline

Motivation
Image captioning
Visual question answering
Summary
Barack Obama is an American politician serving as the 44th President of the United States. Born in Honolulu, Hawaii, … in 2008, he defeated Republican nominee and was inaugurated as president on January 20, 2009. (Wikipedia.org)
Image Captioning (one step from perception to cognition)

describe objects, attributes, and relationship in an image, in a natural language form

-- Let’s do a Turing Test!
Two major paradigms

Two entries tied at the 1st place at COCO 2015 Caption Challenge

End-to-end using LSTM (e.g., Google)
Adopted encoder-decoder framework from machine translation, Popular: Google, Montreal, Stanford, Berkeley


Compositional framework (e.g., MSR)
Visual concept detection => caption candidates generation =>
Deep semantic ranking

Compositional framework can potentially exploit non paired image-caption data more effectively

MSR, Stage 1: Multiple Instance Learning (MIL)

- Treat training caption as bag of image labels
- Train one binary classifier per label on all images
- "Noisy-Or" classifier
  - Image divided into 12x12 overlapping regions
  - fc7 vector used for image features

\[
p(w \text{ in } r_j \text{ of image } i) = 1 - \prod_{j \in r_i} (1 - \sigma(f_{ij} \cdot v_w))
\]

- \(i\) = image id
- \(f_{ij}\) = fc7 vector
- \(\sigma(x)\) = sigmoid
- \(r_i\) = regions
- \(v_w\) = learned classifier weights

E.g., the visual "attention" of word **sitting**.

\[
h(x, y) = \sum_{r_i \text{ s.t. } (x, y) \in r_i} \sigma(f_{ij} \cdot v_{sitting})
\]
Multiple Instance Learning illustration

A man sitting on a chair with a dog in his lap

\[ \hat{P}(w \text{ in region}) = \frac{1}{1 + e^{W_{MIL}\times v_{fc7}}} \]

Raw pixels from input box

Tuned image features from AlexNet (Krizhevsky et al., 2012) or VGG (Simonyan and Zisserman, 2014).
MaxEnt LM (MELM) for modeling language

Table 1. Features used in the maximum entropy language model.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>Definition</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute</td>
<td>0/1</td>
<td>$\bar{w}<em>l \in \tilde{V}</em>{l-1}$</td>
<td>Predicted word is in the attribute set, i.e. has been visually detected and not yet used.</td>
</tr>
<tr>
<td>N-gram+</td>
<td>0/1</td>
<td>$\bar{w}_{l-N+1}, \ldots, \bar{w}_l = \kappa$ and $\bar{w}<em>l \in \tilde{V}</em>{l-1}$</td>
<td>N-gram ending in predicted word is $\kappa$ and the predicted word is in the attribute set.</td>
</tr>
<tr>
<td>N-gram-</td>
<td>0/1</td>
<td>$\bar{w}_{l-N+1}, \ldots, \bar{w}_l = \kappa$ and $\bar{w}<em>l \notin \tilde{V}</em>{l-1}$</td>
<td>N-gram ending in predicted word is $\kappa$ and the predicted word is not in the attribute set.</td>
</tr>
<tr>
<td>End</td>
<td>0/1</td>
<td>$\bar{w}<em>l = \kappa$ and $\tilde{V}</em>{l-1} = \emptyset$</td>
<td>The predicted word is $\kappa$ and all attributes have been mentioned.</td>
</tr>
<tr>
<td>Score</td>
<td>$\mathbb{R}$</td>
<td>score($\bar{w}_l$) when $\bar{w}<em>l \in \tilde{V}</em>{l-1}$</td>
<td>The log-probability of the predicted word when it is in the attribute set.</td>
</tr>
</tbody>
</table>

$$
\begin{align*}
Pr(w_l = \bar{w}_l | \bar{w}_{l-1}, \ldots, \bar{w}_1, <s>, \tilde{V}_{l-1}) = \\
\frac{\exp \left[ \sum_{k=1}^{K} \lambda_k f_k(\bar{w}_l, \bar{w}_{l-1}, \ldots, \bar{w}_1, <s>, \tilde{V}_{l-1}) \right]}{\sum_{v \in \mathcal{V} \cup <s>} \exp \left[ \sum_{k=1}^{K} \lambda_k f_k(v, \bar{w}_{l-1}, \ldots, \bar{w}_1, <s>, \tilde{V}_{l-1}) \right]}
\end{align*}
$$

where $<s>$ denotes the start-of-sentence token, $\bar{w}_j \in \mathcal{V} \cup <s>$, and $f_k(w_l, \cdots, w_1, \tilde{V}_{l-1})$ and $\lambda_k$ respectively denote the $k$-th max-entropy feature and its weight. The basic discrete ME features we use are summarized in Table 1.

$$
L(\Lambda) = \sum_{s=1}^{S} \sum_{l=1}^{#(s)} \log Pr(\bar{w}_l^{(s)} | \bar{w}_{l-1}^{(s)}, \ldots, \bar{w}_1^{(s)}, <s>, \tilde{V}_{l-1}^{(s)})
$$
MELM for candidate generation

Language model

MaxEnt LM

\[ p(\text{cabinets} \mid \text{with wooden}) \]

cabinets

a kitchen with wooden cabinets

Repeat to generate 500 candidates

1. wooden cabinets in a kitchen
2. a sink and cabinets
...
500. a room with stove on the floor

[Fang, et al., CVPR 2015]
Deep Multimodal Similarity Model

- Project sentence and image into a comparable semantic vector space
- Whole sentence language model
- DMSM + basic features → re-ranked caption list

### Mathematical Formulation

**Relevance:**
\[ R(Q, D) = \text{cosine}(y_Q, y_D) = \frac{y_Q^T y_D}{\|y_Q\| \|y_D\|} \]

**Caption probability:**
\[ P(D|Q) = \frac{\exp(\gamma R(Q, D))}{\sum_{D' \in B} \exp(\gamma R(Q, D'))} \]

**Objective:**
\[ L(\Lambda) = -\log \prod_{(Q, D^+)} P(D^+|Q) \]

---


---

Serves as a semantic matching checker.

**Cosine similarity**

**Global vector**

**Image (2D) Convolutional neural network**

**Text (1D) Convolutional neural network**

A man sitting on a bench
The convolutional network at the image side

Feed the pre-trained image feature vector into the image side of the DMSM

Tuned image features from AlexNet (Krizhevsky et al., 2012) or VGG (Simonyan and Zisserman, 2014).
The convolutional network at the caption side

Models fine-grained structural language information in the caption

Using a convolutional neural network for the text caption side

Semantic layer: $y$
Semantic projection matrix: $W_s$
Max pooling layer: $ν$
Max pooling operation
Convolutional layer: $h_t$
Convolution matrix: $W_c$
Word hashing layer: $f_t$
Word hashing matrix: $W_f$
Word sequence: $x_t$

Figure Credit: [Shen, He, Gao, Deng, Mesnil, WWW, April 2014]
What does the model learn at the convolutional layer?

Capture the local context dependent word sense
- Learn one embedding vector for each local context-dependent word

<table>
<thead>
<tr>
<th>car body shop</th>
<th>cosine similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>car body kits</td>
<td>0.698</td>
</tr>
<tr>
<td>auto body repair</td>
<td>0.578</td>
</tr>
<tr>
<td>auto body parts</td>
<td>0.555</td>
</tr>
<tr>
<td>wave body language</td>
<td>0.301</td>
</tr>
<tr>
<td>calculate body fat</td>
<td>0.220</td>
</tr>
<tr>
<td>forcefield body armour</td>
<td>0.165</td>
</tr>
</tbody>
</table>

The similarity between different “body” within contexts

$h_t = W_c \times [f_{t-1}, f_t, f_{t+1}]$
What happens at the max-pooling layer?

- Aggregate *local topics* to form the *global intent*
- Identify salient words/phrase at the max-pooling layer

Words that win the most active neurons at the max-pooling layers:

```
[auto, body, repair, cost, calculator, software]
```

Usually, those are salient words containing clear intents/topics
DMSM learning

Evaluation: on a 5K val set, for each image, rank the 5K captions and check the rank of the true caption

Mean Reciprocal Rank % (ranking among 5000 candidates on the 5K validation set)

- CDSSM d=300
- CDSSM d=1000
- DSSM d=300

Harmonic Mean Rank (ranking among 5000 candidates on the 5K val set)

- CDSSM d=300
- CDSSM d=1000
- DSSM d=300
Evaluation

Human judgment is the ultimate metric

*Turing Test Results*
at the MS COCO Captioning Challenge 2015

<table>
<thead>
<tr>
<th>% of captions that pass the Turing Test</th>
<th>Official Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSR</td>
<td>32.2%</td>
</tr>
<tr>
<td>Google</td>
<td>31.7%</td>
</tr>
<tr>
<td>MSR Captivator</td>
<td>30.1%</td>
</tr>
<tr>
<td>Montreal/Toronto</td>
<td>27.2%</td>
</tr>
<tr>
<td>Berkeley LRCN</td>
<td>26.8%</td>
</tr>
</tbody>
</table>

Other groups: Baidu/UCLA, Stanford, Tsinghua, etc.

Human: 67.5%

Still a big gap!
A brief comparison:

**DMSM’s objective:**
the score of the reference to be higher than other generic captions.

**MRNN’s objective:**
the score of the reference to be higher than arbitrary word sequences

DMSM focuses on semantics rather than syntax. E.g., ensures the reference (semantically interesting) scores higher than generic ones (grammatically correct but semantically incorrect or boring), while MRNN focus on syntax more.
Auto metric & Human Judge

- MELM+DMSM and MRNN obtain same BLEU score
- But humans prefer MELM+DMSM’s output more

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU %</th>
<th>Better or Equal to Human</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1: MELM + DMSM</td>
<td>25.7</td>
<td>34.0%</td>
</tr>
<tr>
<td>Model 2: MRNN</td>
<td>25.7</td>
<td>29.0%</td>
</tr>
</tbody>
</table>

Human judgers shown generated caption and human caption, choose which is “better”, or equal.

Image Diversity

- Test images bucketed based on visual overlap with training
  - MELM+DMSM does well on images with low overlap
  - MRNN does well on images with high overlap

<table>
<thead>
<tr>
<th>Condition</th>
<th>Train/Test Visual Overlap BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Whole Set</td>
</tr>
<tr>
<td>D-ME+DMSM</td>
<td>25.7</td>
</tr>
<tr>
<td>MRNN</td>
<td>25.7</td>
</tr>
</tbody>
</table>

BLEU scores
- Rare images w.r.t. training set
- Common images w.r.t. training set
Language Analysis

- MRNN weakness: Repeated captions
  - Table 1: MRNN repeat captions seen in training data verbatim more often
  - Table 2: Systems produce same captions multiple times; MRNN does it the most

**Example:**

MELM+DMSM: “A plate with a sandwich and a cup of coffee”

MRNN: “A close up of a plate of food” *(more generic)*
From COCO domain to open-domain

- Fast runtime
- Better accuracy per human judgment
- Broader coverage
- Richer information (e.g. people names, locations)
- Output with uncertainty information
Rich Image Captioning in the Wild

ConvNets

Visual concepts
- Celebrity
- Landmark
- Features vector

Language Model

DMSM

Confidence Model

A small boat in Ha Long Bay

This image contains: water, boat, lake, mountain, etc.

[Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, Chris Sienkiewicz submitted to CVPR Deep Vision 2016]
Deep ResNet for visual concepts detection

ResNet

- ImageNet winning solution
- Treat as multiclass problem
- Sigmoid output
- No softmax normalization

Trained on multiple GPUs

man, tennis, court, holding, shirt, yellow, racquet, ...
The deep multimodal semantic model projects images and captions to a abstract semantic space. The overall semantics of a image will be represented by a vector in this space. The overall semantics of a caption will also be represented by a vector in this space. If these two vectors are close to each other, then the caption is a good match for the image. Otherwise, not a matching caption.

DMSM: Bridge the gap between image and language!

Text: *a man holding a tennis racquet on a tennis court*

Deep Residual Network

[Fang, et al., CVPR 2015]

[Fang, et al., CVPR 2015]

[Huang, He, Gao, Deng et al., 2013]

[He, Zhang, Ren, Sun, 2015]
Entity Recognition

- Extreme classification with a **big** set of celebrities

  “Sasha Obama, Malia Obama, Michelle Obama, Peng Liyuan et al. posing for a picture with Forbidden City in the background.”

- Integrating entities (celebrities, landmarks, etc.) makes captions much richer.

[Guo, Zhang, Hu, He, Gao, MS-Celeb-1M: Challenge of Recognizing One Million Celebrities in the Real World, 2016]
Describe with uncertainty

Confidence score \([0, 1]\) 

\[
s = \frac{e^{W \cdot f}}{1 + e^{W \cdot f}}
\]

caption: \(a\) \(man\) holding \(a\) \(tennis\) \(racquet\) \(on\) \(a\) \(tennis\) \(court\)

Image
## Test results - COCO

**Beat previous SOTA on in-domain data (MS COCO)**

<table>
<thead>
<tr>
<th>System</th>
<th>Excellent</th>
<th>Good</th>
<th>Bad</th>
<th>Embarrassing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fang et al., 2015</td>
<td>40.6%</td>
<td>26.8%</td>
<td>28.8%</td>
<td>3.8%</td>
</tr>
<tr>
<td>New system</td>
<td>51.8%</td>
<td>23.4%</td>
<td>22.5%</td>
<td>2.4%</td>
</tr>
</tbody>
</table>

Human evaluation on 1000 random samples of the COCO test set.
Test results - Instagram

Significantly beat previous SOTA on data in the wild

<table>
<thead>
<tr>
<th>System</th>
<th>Excellent</th>
<th>Good</th>
<th>Bad</th>
<th>Embarrassing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fang et al., 2015</td>
<td>12.0%</td>
<td>13.4%</td>
<td>63.0%</td>
<td>11.6%</td>
</tr>
<tr>
<td>New system</td>
<td>25.4%</td>
<td>24.1%</td>
<td>45.3%</td>
<td>5.2%</td>
</tr>
</tbody>
</table>

Human evaluation on Instagram test set, which contains 1380 random images that we scraped from Instagram.
Confidence score distribution - Instagram

Confidence score aligns with human judgement well

<table>
<thead>
<tr>
<th>Conf. score</th>
<th>Excellent</th>
<th>Good</th>
<th>Bad</th>
<th>Embarrassing</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>0.59</td>
<td>0.51</td>
<td>0.26</td>
<td>0.20</td>
</tr>
<tr>
<td>Std dev</td>
<td>0.21</td>
<td>0.23</td>
<td>0.21</td>
<td>0.19</td>
</tr>
</tbody>
</table>
Image Captioning as a Cloud Service

API Accessible via MS Cognitive Services

http://CaptionBot.ai

I think it's a view of a plane flying over a snow covered mountain.
I am not really confident, but I think it's Leonardo da Vinci sitting in front of a mirror and she seems 😊.
when Jen-Hsun Huang was giving a keynote showing off a GPU-powered VR visiting of mt. Everest -- here is what our CaptionBot has to say.
From Captioning to Question Answering

- Answer natural language questions according to the content of a reference image.

**Question:** What are sitting in the basket on a bicycle?

**Answer:** dogs

---

Image Question Answering (IQA)
Caption vs. QA: need reasoning

Image QA: reasoning is a key.

Question: What are sitting in the basket on a bicycle?

Multiple-steps of reasoning over the image to infer the answer

Answer: dogs
Stacked Attention Network for Reasoning

IQA: Need perform multiple steps of reasoning over the image to infer the answer.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Smola, "Stacked Attention Networks for Image Question Answering," CVPR 2016 (oral)
Components of the SAN

- Image Model

Figure 2: CNN based image model

\[ f_I = \text{CNN}_{vgg}(I). \]
Components of the SAN

- **Question Model**
  Code the question into a vector using LSTM

![LSTM diagram]

Figure 3: LSTM based question model
Components of the SAN

- **Question Model**
  
  Code the question into a vector using CNN

```
What are sitting in the basket on a bicycle?

v_{question}
```
Illustration of computing the attention layer

\[ \tilde{v}_I = \sum_i p_i v_i, \]

\[ u = \tilde{v}_I + v_Q. \]
Components of the SAN

- Stacked Attention Mechanism

1st attention layer:

\[ h_A^1 = \tanh(W_{I,A} v_I \oplus (W_{Q,A} v_Q + b_A)) \]
\[ p_I^1 = \text{softmax}(W_P h_A^1 + b_P) \]

Deeper attention layers \((k = 2, \ldots)\):

\[ \tilde{v}_I^k = \sum_i p_i^k v_i, \]
\[ u^k = \tilde{v}_I^k + v_Q. \]
\[ h_A^k = \tanh(W_{I,A}^k v_I \oplus (W_{Q,A}^k u^{k-1} + b_A^k)) \]
\[ p_I^k = \text{softmax}(W_P^k h_A^k + b_P^k) \]
\[ \tilde{v}_I^k = \sum_i p_i^k v_i \]
\[ u^k = \tilde{v}_I^k + u^{k-1} \]
\[ p_{\text{ans}} = \text{softmax}(W_u u^K + b_u) \]
## Results

<table>
<thead>
<tr>
<th>Methods</th>
<th>test-dev</th>
<th>test-std</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>All</td>
<td>Yes/No</td>
</tr>
<tr>
<td>VQA: [1]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Question</td>
<td>48.1</td>
<td>75.7</td>
</tr>
<tr>
<td>Image</td>
<td>28.1</td>
<td>64.0</td>
</tr>
<tr>
<td>Q+I</td>
<td>52.6</td>
<td>75.6</td>
</tr>
<tr>
<td>LSTM Q</td>
<td>48.8</td>
<td>78.2</td>
</tr>
<tr>
<td>LSTM Q+I</td>
<td>53.7</td>
<td>78.9</td>
</tr>
<tr>
<td>SAN(2, CNN)</td>
<td>58.7</td>
<td>79.3</td>
</tr>
</tbody>
</table>

Table 5: VQA results on the official server, in percentage

**Big improvement** on the VQA benchmark and other benchmarks

Other: Object
Color
Location ...
Q: what stands between two blue lounge chairs on an empty beach?

Ans: umbrella
More examples:

(a) What are pulling a man on a wagon down on dirt road?
   Answer: horses Prediction: horses

(b) What is the color of the box?
   Answer: red Prediction: red

(c) What next to the large umbrella attached to a table?
   Answer: trees Prediction: tree

(d) How many people are going up the mountain with walking sticks?
   Answer: four Prediction: four

(e) What is sitting on the handle bar of a bicycle?
   Answer: bird Prediction: bird

(f) What is the color of the horns?
   Answer: red Prediction: red

Original Image | First Attention Layer | Second Attention Layer | Original Image | First Attention Layer | Second Attention Layer
Error analysis: 22% wrong attention, 42% wrong prediction, 31% ambiguous answer, 5% label error

(a) What swim in the ocean near two large ferries? 
Answer: ducks Prediction: boats

(b) What is the color of the shirt? 
Answer: purple Prediction: green

(c) What is the young woman eating? 
Answer: banana Prediction: donut

(d) How many umbrellas with various patterns? 
Answer: three Prediction: two

(e) The very old looking what is on display? 
Answer: pot Prediction: vase

(f) What are passing underneath the walkway bridge? 
Answer: cars Prediction: trains
Go deeper?

before: a herd of elephants standing next to a man

Now, + Entity: a herd of elephants standing next to Obama

Next, + knowledge & reasoning:

Obama is the president from the Democratic party, whose competitor is the Republic party, whose mascot is Elephant.

Obama is chased by his republic competitors 😊

Who is that person?
What are behind that man?
Why these elephants are chasing him?
Other relevant work

• Question answering/Inference
  • David Golub, Xiaodong He, “Character-Level Question Answering with Attention,” arXiv 1604.00727
  • Ji He, Jianshu Chen, Xiaodong He, Jianfeng Gao, Lihong Li, Li Deng, Mari Ostendorf, "Deep Reinforcement Learning with an Action Space Defined by Natural Language," arXiv:1511.04636
  • Wen-tau Yih, Ming-Wei Chang, Xiaodong He, and Jianfeng Gao, Semantic Parsing via Staged Query Graph Generation: Question Answering with Knowledge Base, ACL 2015
  • Wen-tau Yih, Xiaodong He, and Christopher Meek, Semantic Parsing for Single-Relation Question Answering, ACL 2014

• Knowledge/Language representation
  • Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng, Embedding Entities and Relations for Learning and Inference in Knowledge Bases, ICLR 2015
  • Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang, Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval, in NAACL 2015
Summary

- *Language* is a valuable supervision for teaching machines to understand complex scenes *as humans do*.

- Deep learning models can perform certain level of *reasoning* in the image-language joint space and answer questions

- Need to add *knowledge* to give machines the common sense beyond in an isolated image

Please try out [http://CaptionBot.ai](http://CaptionBot.ai) 😊