Visual Question Answering with Deep Learning
What is the mustache made of?
New Features

• Model
  • Naïve Method => Neural Reasoner Based Model

• Language
  • English => French

• Content
  • Image => Video
Previous Model

Image → Image Feature → Combine → LSTM → Final Answer

Question → Question Vector → Combine
Neural Reasoner Based Model

NEURAL REASONER

• Neural Reasoner is a framework for neural network-based reasoning over natural language sentences
• Reasoning is widely used in natural language processing tasks
• Architecture
  • Encoder Layer
  • Reasoning layer
  • Answering Layer
Neural Reasoner Based Model

**NATURAL LANGUAGE PROCESSING**

- $Q \xrightarrow{\text{encode}} q^{(0)}, F_k \xrightarrow{\text{encode}} f_k^{(0)}, k = 1, 2, ..., K.$
- $\{q^{(l)}, f_1^{(l)} ..., f_K^{(l)}\} \xrightarrow{\text{reasoning}} \{q^{(l+1)}, f_1^{(l+1)} ..., f_K^{(l+1)}\}$
Neural Reasoner Based Model

**REASONING**

- Question-Fact Iteration
  \[ [q^{(l)}_k, f^{(l)}_k] \overset{\text{def}}{=} g_{DNN_l} \left( \left[ (q^{(l-1)})^T, f^{(l-1)}_k T \right]^T ; \Theta_l \right) \]

- Pooling
  - Max/Average Pooling
  - Gating
  - Model-Based (CNN / RNN)
Neural Reasoner Based Model

**IMAGE PROCESSING**

- **Input Image** → $F_k$, $k = 1, 2, ..., K$.
- $Q \xrightarrow{encoder} q^{(0)}, F_k \xrightarrow{encoder} f_k^{(0)}, k = 1, 2, ..., K$.
- $\left\{ q^{(l)}, f_1^{(l)}, ..., f_K^{(l)} \right\} \xrightarrow{reasoning} \left\{ q^{(l+1)}, f_1^{(l+1)}, ..., f_K^{(l+1)} \right\}$
Neural Reasoner Based Model

OBJECT LOCALIZATION

• BING ++
• Edge Boxes
• Objectness
• YOLO Darknet ✓
Neural Reasoner Based Model

**YOLO**

- Divide image into $S \times S$ grid
- Within each grid cell predict:
  - B Boxes: 4 coordinates + confidence
  - Class scores: C numbers
- Regression from image to $7 \times 7 \times (5 \times B + C)$ tensor
- Direct prediction using a CNN
Neural Reasoner Based Model

OVERALL STRUCTURE

Question

Object Localization

Object 1
Object 2
Object ...
Object N - 1
Object N

CNN

Fact 1
Fact 2
Fact ...
Fact N - 1
Fact N

Query

Word Embedding

Iterative Reasoning

Answer
French Support

FEATURES:

• Support French Q&A (French Question and French Answer)
• Not need French VQA dataset for training.
French Support

**BASIC IDEA:**

- Key component: Bilingual Model
- Described in paper: *Bilingual word representations with monolingual quality in mind.*
- By using Bilingual Model, we can achieve the equivalence between English and French.
These two monolingual models embed words in two languages with same meaning into 2 vectors that are close.

For example:
“Cat” (in English)
“Chat” (in French)
French Support

**BILINGUAL MODEL TRAINING**

- Step 1: Training two monolingual models separately using Skip-Gram.
- Keep on
  - Moving words in context closer and closer.
  - Moving words outside context further and further.
French Support

BILINGUAL MODEL TRAINING

• Step 2: Training two monolingual model together using Biskip-Gram.
• Biskip-Gram is based on the idea of Skip-Gram.
• Language1: A1 B1 C1 D1 E1 F1 G1 H1 I1 J1 K1 L1 M1 N1
• Language2: A2 B2 C2 D2 E2 F2 G2 H2 I2 J2 K2 L2 M2 N2
French Support

TRAINING & PREDICTING

• Training: Based on the properties of bilingual model, we can consider French words and English words as the similar vectors. Therefore, we only need English-version dataset for training.
Video Question Answering

**METHODOLOGY**

- **Scene Detector**
  - Key Frame 1
  - Key Frame 2
  - Key Frame ...
  - Key Frame N - 1
  - Key Frame N

- **Question**

- **Word Embedding**

- **Query**

- **AI Model**
  - Answer Set 1
  - Answer Set 2
  - Answer Set ...
  - Answer Set N - 1
  - Answer Set N

- **Optimal Answer**
Video Question Answering

VIDEO SCENE DETECTOR

- Content-Aware Detector
- Threshold Detector
Video Question Answering

**CONTENT-AWARE DETECTOR**

- The content-aware scene detector finds areas where the difference between two subsequent frames exceeds the threshold value that is set.

**THRESHOLD DETECTOR**

- The threshold-based scene detector compares the intensity/brightness of the current frame with a set threshold, and triggering a scene cut/break when this value crosses the threshold.
Training

NEURAL REASONER BASED MODEL

Loss

Accuracy
**Result**

**NEURAL REASONER BASED MODEL**

- **Accuracy**

<table>
<thead>
<tr>
<th></th>
<th>First Semester</th>
<th>Second Semester</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes/No</td>
<td>74.62</td>
<td>80.62</td>
</tr>
<tr>
<td>Number</td>
<td>31.76</td>
<td>31.78</td>
</tr>
<tr>
<td>Other</td>
<td>31.32</td>
<td>40.02</td>
</tr>
<tr>
<td>Overall</td>
<td>45.87</td>
<td>48.53</td>
</tr>
</tbody>
</table>
Result

• First Semester
  • What is this animal?
    • zebra, giraffe, horse, cow, zebras
  • How many animals are there?
    • 2, 3, 4, 1, 5

• Second Semester
  • What is this animal?
    • zebra, zebras, giraffe, cow, horse
  • How many animals are there?
    • 4, 3, 2, 1, 6
Result

• First Semester
  • What are flying through the sky?
    • kites, plane, kite, clouds, airplane
  • How many objects in the sky?
    • 13, 10, 4, 5, 1

• Second Semester
  • What are flying through the sky?
    • plane, airplane, kites, kite, clouds
  • How many objects in the sky?
    • 4, 5, 2, 3, 1
Result

FRENCH Q&A

- *English:*  
  - What is this boy doing?  
  - Eating.

- *French:*  
  - Qu’est-ce qu’il fait?  
  - Manger
Result

FRENCH Q&A

• *English*: What animal is this?
  - Elephant

• *French*: Qu’est-ce que c’est animal?
  - l’éléphant
Result

VIDEO QUESTION ANSWERING
Result

VIDEO QUESTION ANSWERING

• Which animal is this?
  • Dog

• What are they doing?
  • Posing
Discussion

NEURAL REASONER BASED MODEL

• Positive
  • Neural Reasoner

• Negative
  • Object localization algorithms
  • Training set
Discussion

VIDEO QUESTION ANSWERING

• Regards each frame as individual (Ignore the relation between frames and frames)
  • Motions
  • Actions
• Key frames to represent the whole video (Lose information)
Discussion

FRENCH SUPPORT

• Although we can find some correct answers in our French Q&A, the most of answers are not correct. After analyzing the model, we think there are two reasons causing this problem.
Discussion

FRENCH SUPPORT

• Reason 1: Semantic structures of French and English are not the same.
  • What animal is this
  • Qu’est-ce que c’est animal
Discussion

FRENCH SUPPORT

• Reason 2: The bilingual model is not accurate enough.
• $v_1 = \text{English\_word\_to\_vec(“cat”)}$
• $v_2 = \text{French\_word\_to\_vec(“chat”)}$
• $|v_1 - v_2| / |v_1| = 1.113$
• $|v_1 - v_2| / |v_2| = 0.818$
Thank You