MIL-UT Presentation on Abstract Image Challenge

University of Tokyo
*Kuniaki Saito,*Andrew Shin, Yoshitaka Ushiku, Tatsuya Harada
Outline

- Challenge
  - Baseline
  - Our Method
  - Result
- After the Challenge
  - Our Method
  - Result
- Summary
Our Method on Challenge

- Holistic features [Zhang et al.]

- Deep Holistic Features
  - ResNET [He et al., 2015]
  - VGG-19 [Simonyan et al., 2014]

- Region Features
  - 1. Avg. Softmax on Top Regions
  - 2. VLAD on Region Proposals
  - Selective Application Based on Question
1. Average Softmax

- Softmax from Top Regions
  - DeepProposal [Ghodrati et al., ICCV 2015]
  - Fast-RCNN [Girshick ICCV 2015]

Example of bounding boxes

Fast-RCNN and VGG-16 trained on ILSVRC object detection task
2. VLAD Coding

- Local feature coding
  - VLAD [Arandjelovic et al., CVPR 2013]

Regions from selective search

256 dim

Fc7 + Coordinate vec
Fc7 + Coordinate vec
Fc7 + Coordinate vec
Fc7 + Coordinate vec

8 dim

264 dim

VLAD Coding
3. Baseline Method

Elementwise multiplication

softmax

“living room”

Where is this?
We alternate between Avg. Softmax (yes/no, number) and VLAD (others) depending on the type of questions.
We won this challenge

<table>
<thead>
<tr>
<th></th>
<th>Open-Ended</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>method</td>
<td>all</td>
<td>Yes/no</td>
<td>Number</td>
</tr>
<tr>
<td>Baseline</td>
<td>65.02</td>
<td>77.5</td>
<td>52.5</td>
<td>56.4</td>
</tr>
<tr>
<td>Challenge Result</td>
<td>67.39</td>
<td>79.6</td>
<td>57.1</td>
<td>58.2</td>
</tr>
</tbody>
</table>

|                | MultipleChoice              |                  |                  |                  |
|                | method                      | all              | Yes/no           | Number           | other            |
| Baseline       | 69.21                       | 77.5             | 52.9             | 66.7             |
| Challenge Result | 71.18                      | 79.6             | 56.2             | 67.9             |
Further Improvement

Proposal of DualNet for VQA

- Keypoint: structure to fuse various features
  - Elementwise summation and multiplication

Multiplication

\[
\begin{align*}
I' &= \tanh(W_M I) \\
Q' &= \tanh(W_Q Q) \\
F &= I' \odot Q' \\
output &= W_f F
\end{align*}
\]

Summation

\[
\begin{align*}
I' &= \tanh(W_M I) \\
Q' &= \tanh(W_Q Q) \\
F &= I' + Q' \\
output &= W_f F
\end{align*}
\]

How about performing both multiplication and summation?
DualNet on Abstract Image

Projection on common space

\[ I_{M1}' = \tanh(W_{M1}I_1), I_{S1}' = \tanh(W_{S1}I_1) \]
\[ I_{M2}' = \tanh(W_{M2}I_2), I_{S2}' = \tanh(W_{S2}I_2) \]
\[ Q_M' = \tanh(W_{MQ}Q), Q_S' = \tanh(W_{SQ}Q) \]

Elementwise multiplication and summation

\[ F_M = I_{M1}' \odot I_{M2}' \odot Q_M' \]
\[ F_S = I_{S1}' + I_{S2}' + Q_S' \]

Where is this? "living room"

Where is this?

\[ F = \text{Concat}(F_M, F_S) \]
\[ \text{output} = W_{f2} \tanh(W_{f1}F) \]
We expect that the architecture of DualNet is effective for abstract image too.

\[
\begin{align*}
I_{M1}' &= \tanh(W_{M1}I_1), I_{S1}' = \tanh(W_{S1}I_1) \\
I_{M2}' &= \tanh(W_{M2}I_2), I_{S2}' = \tanh(W_{S2}I_2) \\
Q_M' &= \tanh(W_{MQ}Q), Q_S' &= \tanh(W_{SQ}Q) \\
\end{align*}
\]

Elementwise multiplication and summation

\[
\begin{align*}
F_M &= I_{M1}' \odot I_{M2}' \odot Q_M' \\
F_S &= I_{S1}' + I_{S2}' + Q_S' \\
F &= \text{Concat}(F_M, F_S) \\
\text{output} &= W_{f2} \tanh(W_{f1}F)
\end{align*}
\]

Where is this? "living room"
Model Ensemble

- 4 models
  - Different common space dimensions

Where is this? ResNet 152 Holistic feature

Living room
Best results on abstract image including this challenge results

<table>
<thead>
<tr>
<th>method</th>
<th>all</th>
<th>Yes/no</th>
<th>Number</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>65.02</td>
<td>77.5</td>
<td>52.5</td>
<td>56.4</td>
</tr>
<tr>
<td>Challenge Result</td>
<td>67.39</td>
<td>79.6</td>
<td>57.1</td>
<td>58.2</td>
</tr>
<tr>
<td>DualNet (ours)</td>
<td>68.87</td>
<td>80.0</td>
<td>57.9</td>
<td>61.1</td>
</tr>
<tr>
<td>DualNet (ensemble)</td>
<td>69.73</td>
<td>80.7</td>
<td>58.8</td>
<td>62.1</td>
</tr>
</tbody>
</table>
**Result**

- Best results on abstract image including this challenge results

<table>
<thead>
<tr>
<th>method</th>
<th>Open-Ended</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>all</td>
</tr>
<tr>
<td>Baseline</td>
<td>65.02</td>
</tr>
<tr>
<td>Challenge Result</td>
<td>67.39</td>
</tr>
<tr>
<td>DualNet (ours)</td>
<td>68.87</td>
</tr>
<tr>
<td>DualNet (ensemble)</td>
<td>69.73</td>
</tr>
</tbody>
</table>

2.3% improvement from Challenge result!
Best results on abstract image including this challenge results

<table>
<thead>
<tr>
<th>Method</th>
<th>all</th>
<th>Yes/no</th>
<th>Number</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>69.21</td>
<td>77.5</td>
<td>52.9</td>
<td>66.7</td>
</tr>
<tr>
<td>Challenge Result</td>
<td>71.18</td>
<td>79.6</td>
<td>56.2</td>
<td>67.9</td>
</tr>
<tr>
<td>DualNet (ours)</td>
<td>73.29</td>
<td>80.0</td>
<td>58.5</td>
<td>62.0</td>
</tr>
<tr>
<td>DualNet (ensemble)</td>
<td>74.02</td>
<td>80.8</td>
<td>59.2</td>
<td>72.4</td>
</tr>
</tbody>
</table>
Best results on abstract image including this challenge results

<table>
<thead>
<tr>
<th>method</th>
<th>all</th>
<th>Yes/no</th>
<th>Number</th>
<th>other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>69.21</td>
<td>77.5</td>
<td>52.9</td>
<td>66.7</td>
</tr>
<tr>
<td>Challenge Result</td>
<td>71.18</td>
<td>79.6</td>
<td>56.2</td>
<td>67.9</td>
</tr>
<tr>
<td>DualNet (ours)</td>
<td>73.29</td>
<td>80.0</td>
<td>58.5</td>
<td>62.0</td>
</tr>
<tr>
<td>DualNet (ensemble)</td>
<td>74.02</td>
<td>80.8</td>
<td>59.2</td>
<td>72.4</td>
</tr>
</tbody>
</table>

2.9% improvement from Challenge result!
Summary

- Method on the Challenge
  - Deep holistic features
  - + selective region features

- Improved method: DualNet
  - Performing both multiplication and summation
Acknowledgement

- This work was funded by ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan).
Paper

- DualNet: Domain-Invariant Network for Visual Question Answering
Thank you for listening!