THE UNIVERSITY OF TOKYO

MIL-UT Presentation on Abstract Image Challenge

University of Tokyo

*Kuniaki Saito ,*Andrew Shin,
Yoshitaka Ushiku , Tatsuya Harada

Outline

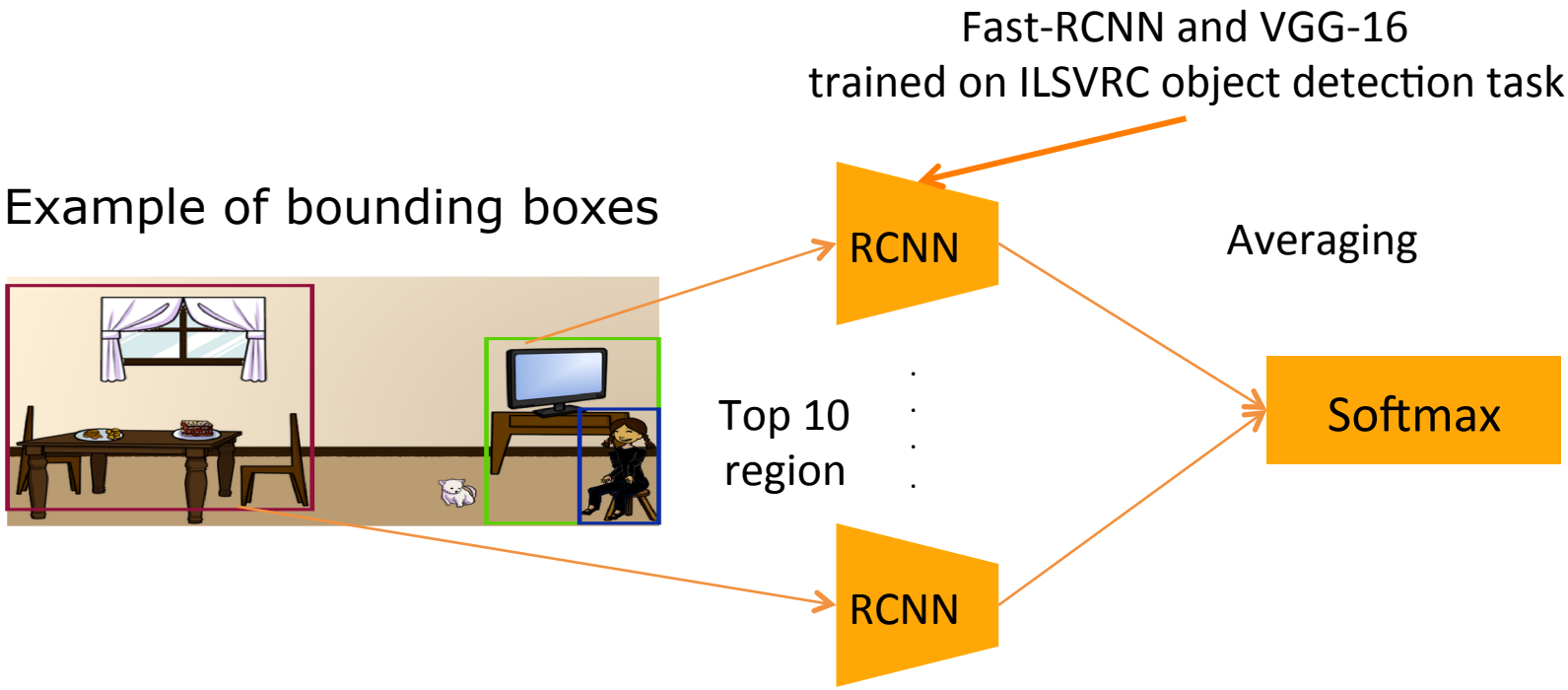
- Challenge
  - Baseline
  - Our Method
  - Result
- After the Challenge
  - Our Method
  - Result
- Summary

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

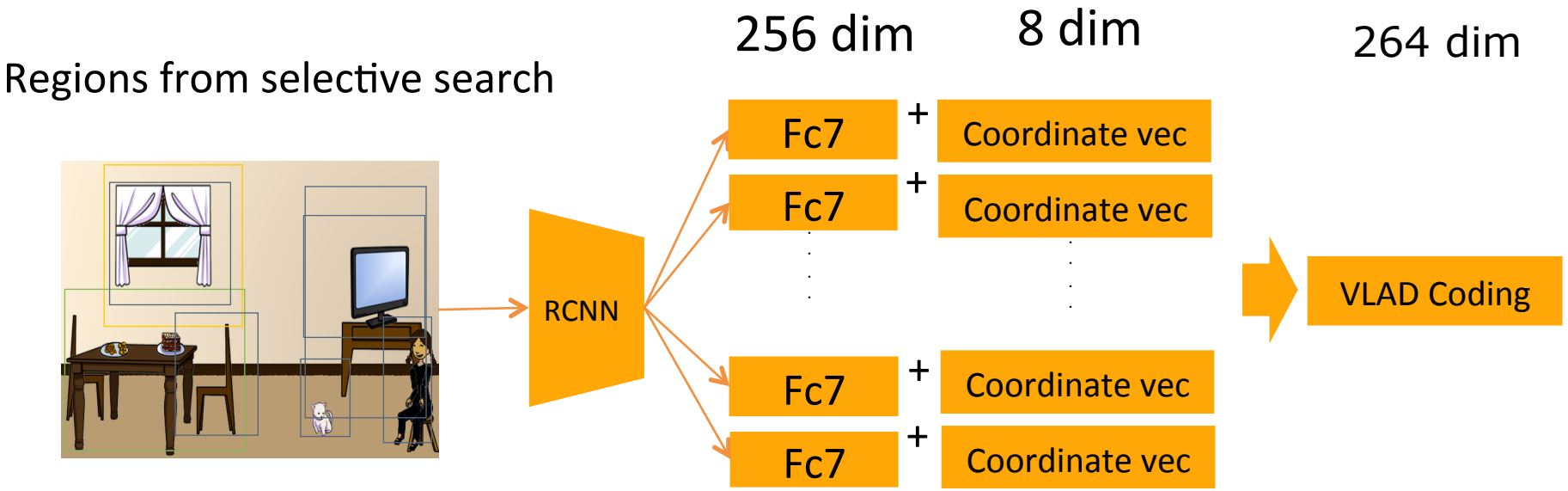
1. Average Softmax

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

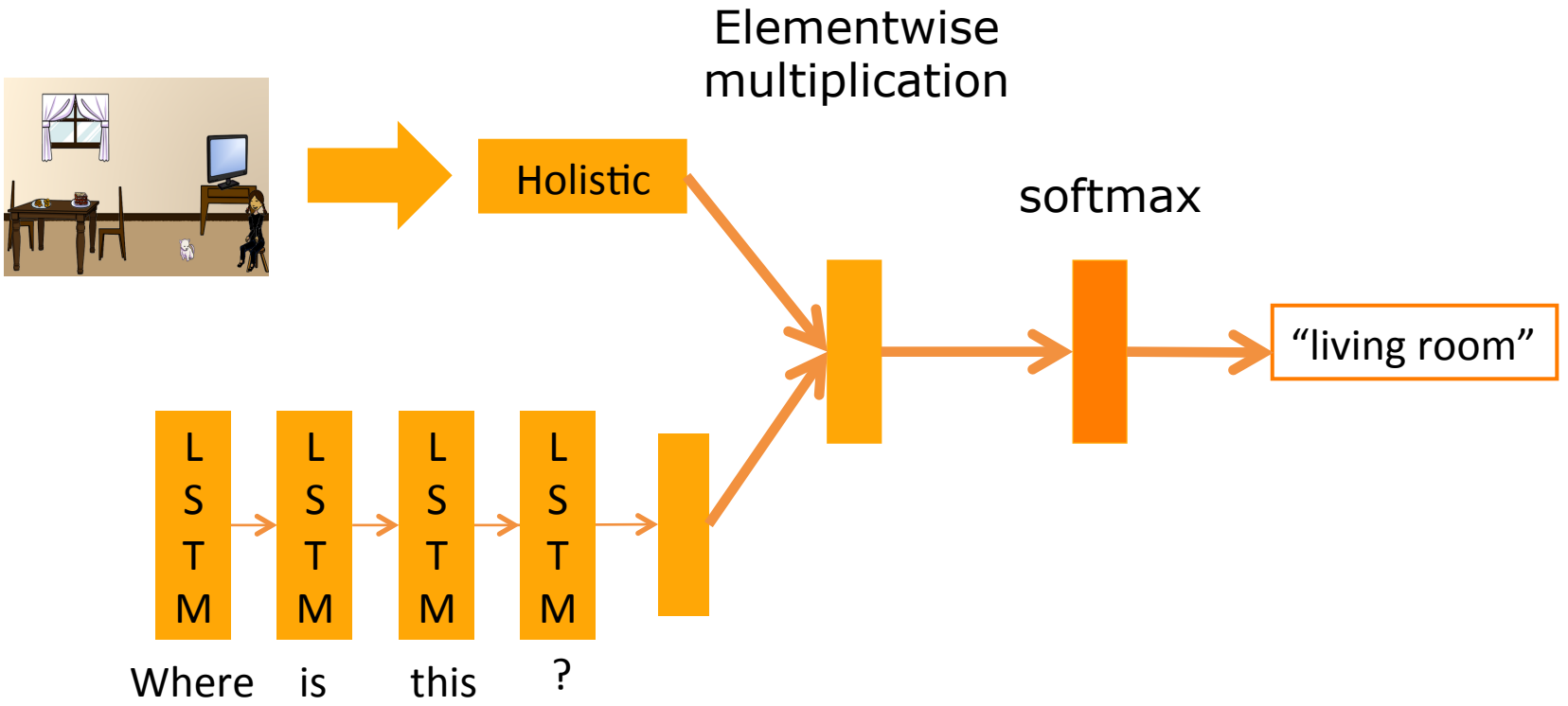


2. VLAD Coding

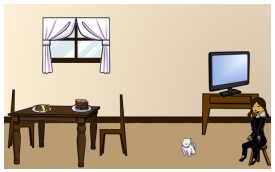
- Local feature coding
  - VLAD [Arandjelovic et al., CVPR 2013]



3. Baseline Method



Network Architecture



Holistic

VGG

ResNet

VLAD or Softmax

L2normalize

①

②

L S T M

L S T M

L S T M

L S T M

Where is this ?

“living room”

① $I' = tanh(W_M(Concat(I_1, I_2, I_3, I_4)))$

② $Q'_M = tanh(W_q Q)$

$F_M = I' \odot Q'_M$

$output = W_{f_1} F_M$

We alternate between Avg. Softmax (yes/no, number) and VLAD (others) depending on the type of questions

Result

- We won this challenge

Open-Ended				
method	**all**	**Yes/no**	**Number**	**other**
Baseline	65.02	77.5	52.5	56.4
Challenge Result	67.39	79.6	57.1	58.2

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

$$
\begin{aligned}
I' &= tanh(W_M I) \\
Q' &= tanh(W_q Q) \\
F &= I' \odot Q' \\
output &= W_f F
\end{aligned}
$$

Summation

$$
\begin{aligned}
I' &= tanh(W_M I) \\
Q' &= tanh(W_q Q) \\
F &= I' + Q' \\
output &= W_f F
\end{aligned}
$$

How about performing both muliplication and summation?

DualNet on Abstract Image

Projection on common space

① $I_{M_1}' = tanh(W_{M1} I_1), I_{S_1}' = tanh(W_{S1} I_1)$

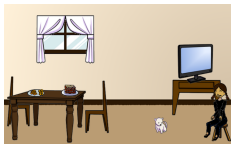
$I_{M_2}' = tanh(W_{M2} I_2), I_{S_2}' = tanh(W_{S2} I_2)$

$Q_M' = tanh(W_{Mq} Q), Q_S' = tanh(W_{Sq} Q)$

Elementwise multiplication and summation

② $F_M = I_{M_1}' \odot I_{M_2}' \odot Q_M'$

$F_S = I_{S_1}' + I_{S_2}' + Q_S'$



Holistic feature

ResNet 152

L S T M

L S T M

L S T M

L S T M

Where is this ?

① ②

③

"living room"

③ $F = Concat(F_M, F_S)$

$output = W_{f_2} tanh(W_{f_1} F)$

DualNet on Abstract Image

Projection on common space

① $I_{M_1}' = tanh(W_{M1} I_1), I_{S_1}' = tanh(W_{S1} I_1)$

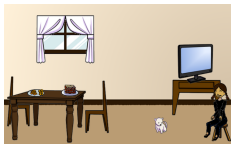
$I_{M_2}' = tanh(W_{M2} I_2), I_{S_2}' = tanh(W_{S2} I_2)$

$Q_M' = tanh(W_{Mq} Q), Q_S' = tanh(W_{Sq} Q)$

Elementwise multiplication and summation

② $F_M = I_{M_1}' \odot I_{M_2}' \odot Q_M'$

$F_S = I_{S_1}' + I_{S_2}' + Q_S'$



Holistic feature

ResNet 152

LSTM LSTM LSTM LSTM

Where is this ?

① ②

③

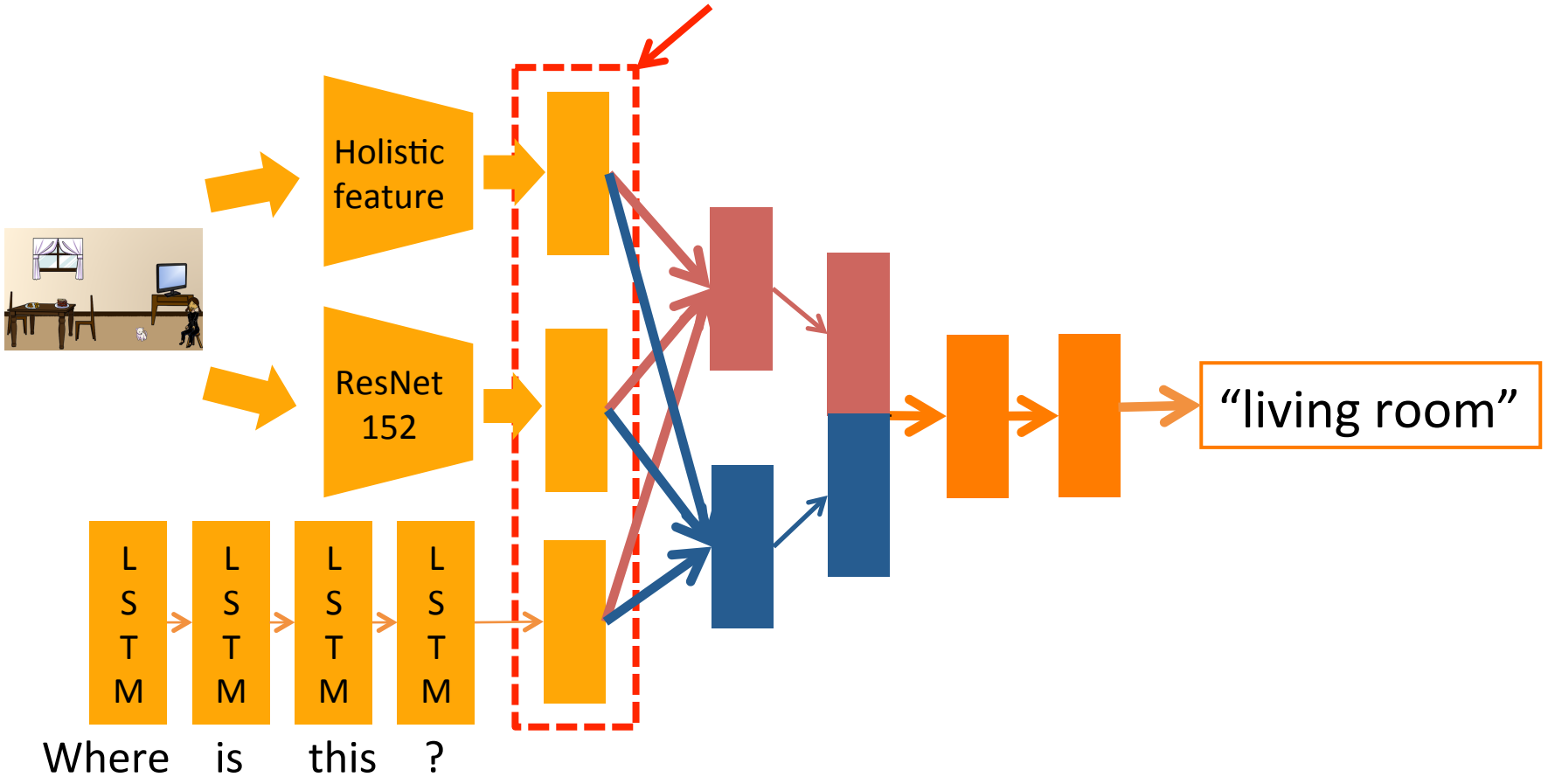
"living room"

③ $F = Concat(F_M, F_S)$

$output = W_{f_2} tanh(W_{f_1} F)$

Model Ensemble

- 4 models
  - Different common space dimensions



Result

- Best results on abstract image including this challenge results

Open-Ended				
method	**all**	**Yes/no**	**Number**	**other**
Baseline	65.02	77.5	52.5	56.4
Challenge Result	67.39	79.6	57.1	58.2
DualNet (ours)	68.87	80.0	57.9	61.1
DualNet (ensemble)	69.73	80.7	58.8	62.1

Result

- Best results on abstract image including this challenge results

Open-Ended				
method	**all**	**Yes/no**	**Number**	**other**
Baseline	65.02	77.5	52.5	56.4
Challenge Result	67.39	79.6	57.1	58.2
DualNet (ours)	68.87	80.0	57.9	61.1
DualNet (ensemble)	69.73	80.7	58.8	62.1

2.3% improvement from Challenge result!

Result

- Best results on abstract image including this challenge results

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
DualNet (ours)	73.29	80.0	58.5	62.0
DualNet (ensemble)	74.02	80.8	59.2	72.4

Result

- Best results on abstract image including this challenge results

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
DualNet (ours)	73.29	80.0	58.5	62.0
DualNet (ensemble)	74.02	80.8	59.2	72.4

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
  https://arxiv.org/pdf/1606.06108.pdf

Thank you for listening!