Waseda at TRECVID 2015
Semantic Indexing (SIN)

Kazuya UEKI and Tetsunori KOBAYASHI
Waseda University
1. System Description

Our computing environment

Limited CPU and memory resources

Two off-the-shelf computers.
1. System description

Our computing environment

Two off-the-shelf computers.

But these computers each have two GPUs. (Titan Black)
For this year’s submission, we decided to focus on extracting features only from CNNs.

We did not use local features (SIFT or HOG), motion features (dense trajectories), and audio features.
1. System description

Semantic indexing pipelines:

[Step 1] Feature extraction using multiple CNNs

[Step 2] Feature pooling

[Step 3] Classification with SVMs

[Step 4] Fusion of multiple score outputs
1. System description

[Step 1] Feature extraction using multiple CNNs

The network structure: AlexNet

4096-D features
4096-D features
1000-D features
1. System description

[Step 1] Feature extraction using multiple CNNs

Instead of using local features or motion features, 6 different CNNs were used.

SIFT, HOG, and etc

Dense trajectories
1. System description

[Step 1] Feature extraction using multiple CNNs

(1) **ImageNet**
- Trained with the ImageNet dataset
  (1.2 million images and 1,000 categories)
- Provided with the Caffe (CNN) library

(2) **Finetune**
- Created by *finetuning* ImageNet model for TRECVID SIN task
- 1 million keyframe images
- 346 concepts
  (# of units in the output layer: 346)
1. System description

[Step 1] Feature extraction using multiple CNNs

(3) Gradient

- Substitute edge features with CNN features
- Trained with 1 million gradient images
- 346 concepts

Color: Orientation of gradient
Brightness: Magnitude of the orientation gradients

Original image  Gradient image
1. System description

[Step 1] Feature extraction using multiple CNNs

(4) OpticalFlow

- Substitute motion features with CNN features
- Trained with 1 million optical flow images
- 346 concepts

Color: Orientation of the optical flow
Brightness: Magnitude of the optical flow
1. System description

[Step 1] Feature extraction using multiple CNNs

(5) **Places**
- Scene recognition model
- Trained on 205 scene categories
- 2.5 million images
- Provided by MIT (Caffe model zoo)

(6) **Hybrid**
- Scene and object recognition model
- Trained on 1,183 categories
  (205 scene categories + 978 object categories)
- 3.6 million images
- Provided by MIT (Caffe model zoo)
Multiple frames from a shot

Frame

1

2

3

10

[Step 2] Feature pooling

We selected a maximum of 10 frames from a shot at regular intervals.
1. System description

[Step 2] Feature pooling

Frame:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.051</td>
<td>-9.251</td>
<td></td>
<td>-3.482</td>
</tr>
<tr>
<td>2</td>
<td>-1.349</td>
<td>-3.039</td>
<td></td>
<td>-1.498</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
<td>:..</td>
<td>...</td>
</tr>
<tr>
<td>10</td>
<td>2.493</td>
<td>1.455</td>
<td></td>
<td>2.411</td>
</tr>
</tbody>
</table>

The values of the elements in the same dimension were compared across 10 sets, and the maximum value was selected.

Element-wise Max-pooling

\[
\begin{pmatrix}
2.051 \\
-0.148 \\
\vdots \\
5.471
\end{pmatrix}
\]

One fixed-length vector
1. System description

[Step 3] Classification with SVMs

- Create a separate SVM using features from each layer
  To reduce the computational cost and the toll on memory resources

- Use approximately 20,000–30,000 shots for each concept
  With roughly the same number of positive and negative samples

- Utilize flipped images during both the training and the testing
  To enrich the variations of the training and the testing sets

There are far fewer positive samples than negative samples.

=> Use the flipped images exclusively for the positive samples.
1. System description

[Step 3] Classification with SVMs

Training phase

Original images

Flipped images

SVM (normal)

SVM (data augmentation)
[Step 3] Classification with SVMs

Testing phase

Original image

Flipped image

SVM (normal)

SVM (data augmentation)

Score

Score

Score

Fusion
1. System description

[Step 3] Classification with SVMs

Scores from the following 3 scores were combined.

- Original images used for both training and testing

- Both original and flipped images used for training, but only original images used for testing

- Both original and flipped images used for training, and only flipped images used for testing.
[Step 4] Fusion of multiple score outputs

- **Waseda4**: Fusion weight of 2 for ImageNet, Finetune, Places and Hybrid models.
  Fusion weight of 1 for Hybrid and Gradient models.

- **Waseda3**: Fusion weight were optimized to improve the mAP of 30 concepts.

- **Waseda2**: Fusion weight were optimized to improve the mAP of 60 concepts.

- **Waseda1**: Fusion weight were optimized to improve the average precision of each concept.

Fusion weight optimization did not offer significant improvements over averaging of scores.
## 2. Results of Submitted Runs

### Submission results

The mAPs for individual models with the TRECVID 2015 SIN testing set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Layer</th>
<th>Train: original images</th>
<th>Test: original images</th>
<th>Train: original + flipped images</th>
<th>Test: original images</th>
<th>Test: flipped images</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Train: original images</td>
<td>Test: original images</td>
<td></td>
<td>Test: original images</td>
<td>Test: flipped images</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ImageNet</td>
<td>6</td>
<td>24.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>23.61</td>
<td>24.14</td>
<td>23.89</td>
<td>23.53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>18.82 1</td>
<td>19.08 1</td>
<td>18.70 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finetune</td>
<td>6</td>
<td>23.50</td>
<td>23.80</td>
<td>23.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>23.29</td>
<td>23.39</td>
<td>23.44</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>21.53</td>
<td>21.90</td>
<td>21.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gradient</td>
<td>6</td>
<td>20.74</td>
<td>19.41</td>
<td>19.03</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>19.82</td>
<td>18.95</td>
<td>19.17</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>17.71</td>
<td>17.26</td>
<td>17.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>13.22</td>
<td>13.34</td>
<td>13.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>13.12</td>
<td>13.43</td>
<td>13.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Places</td>
<td>6</td>
<td>23.40</td>
<td>23.61</td>
<td>23.74</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>22.29</td>
<td>22.41</td>
<td>22.20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>22 2</td>
<td>22 2</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hybrid</td>
<td>6</td>
<td>25.12</td>
<td>24.75</td>
<td>24.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>25.52</td>
<td>25.17</td>
<td>24.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>23.20</td>
<td>22.93</td>
<td>22.88</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
2. Results of Submitted Runs

- Our 2015 submissions ranked between 5 and 8 in a total of 86 runs.
- Our best run was an mAP of 30.86%, which ranked 2nd among all participants.

The mAP was improved by combining all models. Complementary features.

Comparison of Waseda runs with the runs of other teams on IACC 2 C.
2. Results of Submitted Runs

Average precision of our best run (Waseda1) for each semantic concept.

* One of our runs achieved the best average precision for some concepts.
3. Summary and future works

- Despite the simplicity of our method, it achieved relatively high performance.
- The performance of semantic video indexing was still extremely low.
- In the future, we will investigate the root causes of this poor performance and evaluate the options for improving it.
Thank you for your attention.

Any questions?