Waseda_Meisei at TRECVID 2017
Ad-hoc Video Search (AVS)

Kazuya UEKI  Koji HIRAKAWA  Kotaro KIKUCHI
Tetsuji OGAWA  Tetsunori KOBAYASHI

Waseda University
Meisei University
- AVS’s task objective:
  To return a list of at most 1000 shot IDs ranked according to their likelihood for each query.

- Our system:
  Based on a large semantic concept bank.
  *(More than 50,000 concepts)*

- This is our first submission to full automatic run:
  Problem: Word ambiguity in concept selection step.
  WordNet/Word2Vec-based methods were proposed.
  WordNet-based one outperformed Word2Vec-based one.
1. System outline
1. System outline

New

- Query
  - Keyword 1
  - Keyword 2
  - …
  - Keyword N

Concept bank > 50K

- Concept 1
  - Concept 1M₁
  - Concept 2₁
  - Concept 2M₂
  - Concept N₁
  - Concept NMₙ

CNN/SVM of each concept

Score calculation

- Score 1₁
  - Score 1M₁
  - Score 2₁
  - Score 2M₂
  - Score N₁
  - Score NMₙ

Score fusion

Score for Video & Query

Same as 2016 system
<table>
<thead>
<tr>
<th>Training Dataset</th>
<th>Type</th>
<th>#Concepts, Data</th>
<th>Network</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRECVID346</td>
<td>Object, Scene, Action</td>
<td>346 concepts</td>
<td>GoogLeNet</td>
<td>CNN/SVM tandem</td>
</tr>
<tr>
<td>PLACES205</td>
<td>Scene</td>
<td>205 concepts 2500K pictures</td>
<td>AlexNet</td>
<td>CNN</td>
</tr>
<tr>
<td>PLACES365</td>
<td>Scene</td>
<td>365 concepts 1800K pictures</td>
<td>GoogLeNet</td>
<td>CNN</td>
</tr>
<tr>
<td>Hybrid1183</td>
<td>Object, Scene</td>
<td>1183 concepts 3600K pictures</td>
<td>AlexNet</td>
<td>CNN</td>
</tr>
<tr>
<td>ImageNet1000</td>
<td>Object</td>
<td>1000 concepts 1200K pictures</td>
<td>AlexNet</td>
<td>CNN</td>
</tr>
<tr>
<td>ImageNet4000,4437,8201,12988</td>
<td>Object</td>
<td>4000,4437,8201,12988 concepts</td>
<td>GoogLeNet</td>
<td>CNN</td>
</tr>
<tr>
<td>ImageNet21841</td>
<td>Object</td>
<td>21841 concepts 14200K pictures</td>
<td>GoogLeNet</td>
<td>CNN</td>
</tr>
<tr>
<td>FCVID239</td>
<td>Object, Scene, Action</td>
<td>239 concepts 91223 movies</td>
<td>GoogLeNet</td>
<td>CNN/SVM tandem</td>
</tr>
<tr>
<td>UCF101</td>
<td>Action</td>
<td>101 concepts 13320 movies</td>
<td>GoogLeNet</td>
<td>CNN/SVM tandem</td>
</tr>
</tbody>
</table>

*ImageNet* refers to the ImageNet dataset.
2. Detail of concept selection
2. Detail: Step 1 Extract keyword

Search keyword from query.

Query: “One or more people at train station platform”

N/A  N/A  “people”  “train”  “station”  “platform”  “train_station_platform”  (Collocation)
Problem:
Representation of the keyword is not the same as that of the index word. Which concept should be used for the keyword.
2. Detail: Step 2 Choose concepts for each keyword

- **Manual runs**
  - The concept for the keyword is manually selected.

- **Automatic runs**
  - WordNet based method.
    - Exact match of *synset*.
  - Word2Vec based method.
    - Similarity of skipgram.
  - Hybrid of WordNet & Word2Vec.
Each “Word” has a set of “Lexeme”s. Lexemes which have the same meaning make synset.
2. Detail: Step 2 Choose concepts for each keyword

Automatic approach #1: WordNet *synset* matching

Query

```
One or more people at train station ...
```

Keyword $i$

```
Synset of Keyword $i$
```

Exact matched

---

Concept bank

```
Index 1: Model of Concept 1
```

```
Index 2: Model of Concept 2
```

```
Index 3: Model of Concept 3
```

```
Index N: Model of Concept N
```

---

```
Synset of Index 1
```

```
Synset of Index 2
```

```
Synset of Index 3
```

```
Synset of Index N
```

---

: WordNet
Automatic approach #2: Word2Vec similarity

2. Detail: Step 2 Choose concepts for each keyword
Automatic approach #2: Word2Vec similarity

2. Detail: Step 2 Choose concepts for each keyword

Query:
One or more people at train station

Concept bank:
Index 1: Model of Concept 1
Index 2: Model of Concept 2
Index 3: Model of Concept 3
Index N: Model of Concept N

Keyword i:
Vector rep. of Keyword i

Vector rep. of Index 1
Vector rep. of Index 2
Vector rep. of Index 3
Vector rep. of Index N

w_i
w_i-1
w_i+1
w_i+2

: Word2Vec
Hybrid method:

Apply WordNet-based method, first.

If failed /* WordNet-based method find no concepts */
then Apply Word2Vec-based one.
Word2Vec-based approach tends to select too many concepts.

WordNet-based approach tends to lack some concepts.

Desired (ideal) Concept Set
2. Detail: Step 2 Calculate score

- TRECVID346
- FCVID239
- UCF101

CNN/SVM tandem
connectionist architecture

1\textsuperscript{st} frame  2\textsuperscript{nd} frame  10\textsuperscript{th} frame

\[
\begin{pmatrix}
2.051 \\
-1.349 \\
\vdots \\
2.493
\end{pmatrix}
\begin{pmatrix}
-9.251 \\
-3.039 \\
\vdots \\
1.455
\end{pmatrix}
\begin{pmatrix}
-3.482 \\
-1.498 \\
\vdots \\
2.411
\end{pmatrix}
\begin{pmatrix}
2.051 \\
-0.148 \\
\vdots \\
5.471
\end{pmatrix}
\]

max pooling

hidden layer

at most 10 images

\[\ldots\]

CNN

SVM

Detail:
Step 2 Calculate score
The shot scores were obtained directly from the output layer (before softmax was applied)

at most 10 images

CNN

\[
\begin{pmatrix}
2.051 & -1.349 & -3.482 \\
-9.251 & -3.039 & -1.498 \\
2.493 & 1.455 & 2.411 \\
\end{pmatrix}
\]
3. Results
## 3. Results (Manual runs)

Comparison of Waseda_Meisei manual runs

<table>
<thead>
<tr>
<th>Name</th>
<th>Fusion method</th>
<th>Fusion weight</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual-1</td>
<td>Multiply(log)</td>
<td>✓</td>
<td>21.6</td>
</tr>
<tr>
<td>Manual-2</td>
<td>Multiply(log)</td>
<td></td>
<td>20.4</td>
</tr>
<tr>
<td>Manual-3</td>
<td>Sum(linear)</td>
<td>✓</td>
<td>20.7</td>
</tr>
<tr>
<td>Manual-4</td>
<td>Sum(linear)</td>
<td></td>
<td>18.9</td>
</tr>
</tbody>
</table>

**Fusion method:** Multiply(log) > Sum(linear)

**Fusion weight:** w/ weight > w/o weight
3. Results (Manual runs)

Comparison of Waseda Meisei runs with the runs of other teams for all submitted manually assisted runs.
### 3. Results (Automatic runs)

#### Comparison of Waseda_Meisei automatic runs

<table>
<thead>
<tr>
<th>Name</th>
<th>WordNet synset</th>
<th>Word2Vec</th>
<th>FCVID239 + UCF101</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-1</td>
<td>✓</td>
<td></td>
<td></td>
<td>15.9</td>
</tr>
<tr>
<td>Auto-2</td>
<td></td>
<td>✓</td>
<td></td>
<td>14.3</td>
</tr>
<tr>
<td>Auto-3</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>14.1</td>
</tr>
<tr>
<td>Auto-4</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>12.5</td>
</tr>
</tbody>
</table>

WordNet vs. Word2Vec:  **WordNet > Word2Vec**
### Results for 2016 TRECVID dataset

<table>
<thead>
<tr>
<th>Name</th>
<th>WordNet synset</th>
<th>Word2Vec</th>
<th>FCVID239 + UCF101</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auto-1</td>
<td>✔</td>
<td></td>
<td></td>
<td>17.8</td>
</tr>
<tr>
<td>Auto-2</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>17.4</td>
</tr>
<tr>
<td>Auto-3</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>17.4</td>
</tr>
<tr>
<td>Auto-4</td>
<td>✔</td>
<td>✔</td>
<td></td>
<td>17.8</td>
</tr>
</tbody>
</table>
3. Results (Automatic runs)

Comparison of Waseda Meisei runs with the runs of other teams for all the fully automatic runs.

- Auto 1: WordNet synset
- Auto 2: Word2Vec
- Auto 3: Word2Vec (rich DB incl. FCVID239 + UCF101)
- Auto 4: WordNet+Wrd2Vec Hybrid (Bug)
3. Results: Difference btw. our Auto & our Manu.

- **534** Find shots of *a person talking behind a podium* wearing a suit outdoors during daytime → “Speaker_At_Podium” is used in manual.
- **542** Find shots of *at least two* planes both visible → Object counting module is installed in manual condition.
- **559** Find shots of a man and woman *inside a car* → “car_interior” is used and “car” is not used in manual.
  (All, parsing (linguistic) problem)
543 Find shots of a person communicating using \textit{sign language} → No concept for “sign language”. (Short of concepts)

554 Find shots of a person holding or operating \textit{a TV or movie camera} → “TV” contaminated. (Parsing problem)

558 Find shots of a person wearing a \textit{scarf} → “\texttt{scarf\_joint}” contaminated. (Word-concept matching problem)

Scarf itself is difficult to recognize. (Scoring problem)
4. Summary & future works
Summary

• We joined in “ad-hoc video search” task.

• This is our first attempt to “automatic run”.
  In step2 (selection of concepts from keyword),
  WordNet-based/Word2Vec-based methods proposed

• WordNet-based concept selection outperformed
  Word2Vec-based one.
4. Summary and future works

Future works

• To improve the concept selection methods.
  e.g. Other use of WordNet / Word2Vec

• To improve linguistic part.
  e.g. a person talking behind xxxx,
       inside car,
       at least two xxxx
       TV or movie camera

• To handle action type concepts.
Thank you for your attention.

Any questions?