Kobe University, NICT and University of Siegen at TRECVID 2016 AVS Task

Yasuyuki Matsumoto, Kuniaki Uehara (Kobe University), Takashi Shinozaki (CiNet, NICT)
Kimiaki Shirahama, Marcin Grzegozek (University of Siegen)

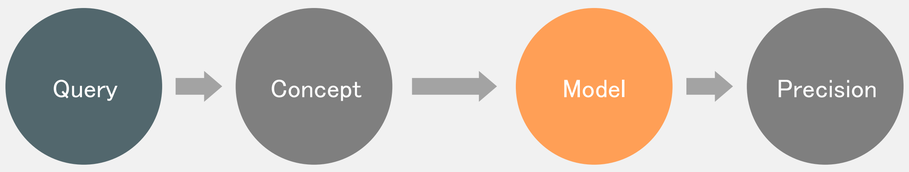
Contribution

A method of using small-scale neural network to greatly accelerate concept classifier training (hours -> minutes).

Transfer learning can acquire temporal characteristics efficiently by combining both small networks and LSTM.

Evaluate the effectiveness of using unbalanced examples at the time of training.

Overview



+ Manual selection
+ Feature extraction
+ MicroNN training
+ LSTM
+ Shot retrieval

Manual Concept Selection

- Begin with manual selection of relevant concepts for each query
- Simple rule makes it easier to automatic concept selection in the future

Query (502)
''Find shots of a man indoors looking at camera where a bookcase is behind him''

Base form "Look"
"man" Pick only noun and verb
"bookcase" "furniture" Synonyms

Concept: Indoor | Speaking_to_camera | Bookshelf | Funiture

Concept Transfer

- Selected concepts are transferred from a dataset to dataset
- A kind of **Curriculum Learning?**

query_id	ImageNet	TRECVID	UCF 101
501		Outdoor	playingGuitar
502	bookshelf →	Indoor, Speaking_to_camera, Furniture	
503	drum →	Indoor	drumming

Shot Retrieval

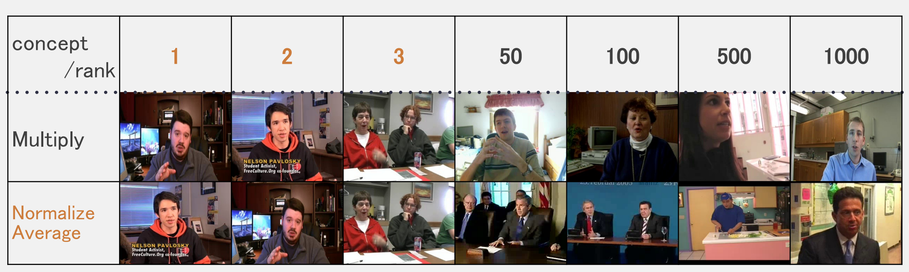
- Firstly, output values of microNNs with target concepts are normalized
- Then, Search Score is calculated as the average among the outputs

Concepts: Indoor | Speaking_to_camera | Bookshelf | Funiture

Outputs of microNNs: 0.7 + 0.2 + 0.5 + 0.6 / 4

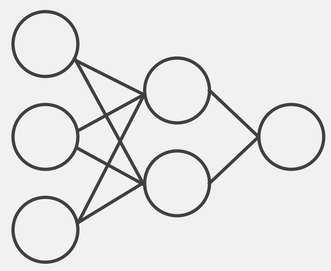
Average Precision (Search Score): 0.5

- Summation is more robust than multiplication of outputs



Micro Neural Networks

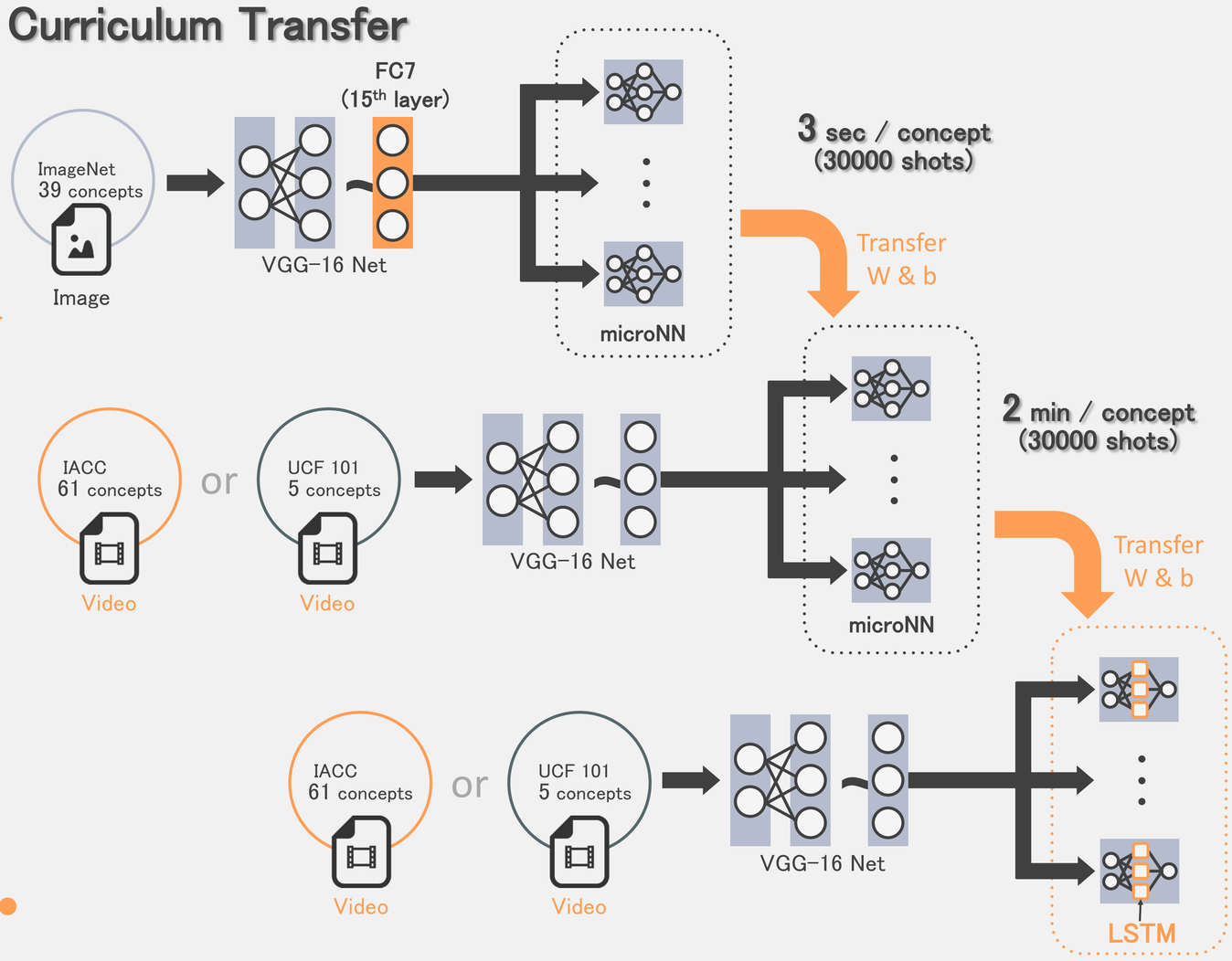
- Binary classifier: just discriminate exist or not
- Very simple network structure
  - 4096 input units, 32 hidden units, 2 output units
  - Fully-connected each other
- Recent DNN technique
  - ReLU and Dropout for hidden layer outputs



Pros
- Much faster
- Learn iteratively
- Easy to extend

Cons
- Less precise a llittle

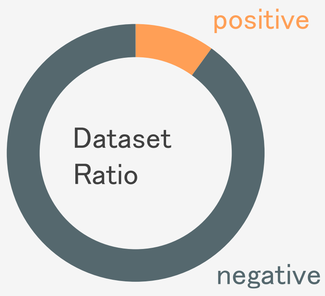
Curriculum Transfer



Our Three Runs

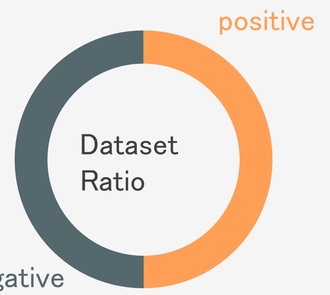
kobe_nict_siegen_D_M_1
Unbalanced

Fine-tuning is carried out using imbalanced numbers of positive and negative examples.



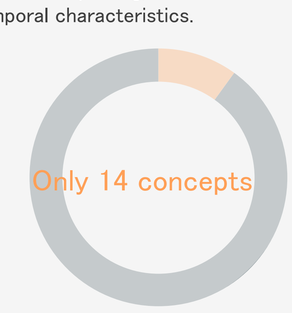
kobe_nict_siegen_D_M_2
Balanced

Fine-tuning is carried out using balanced numbers of positive and negative examples.



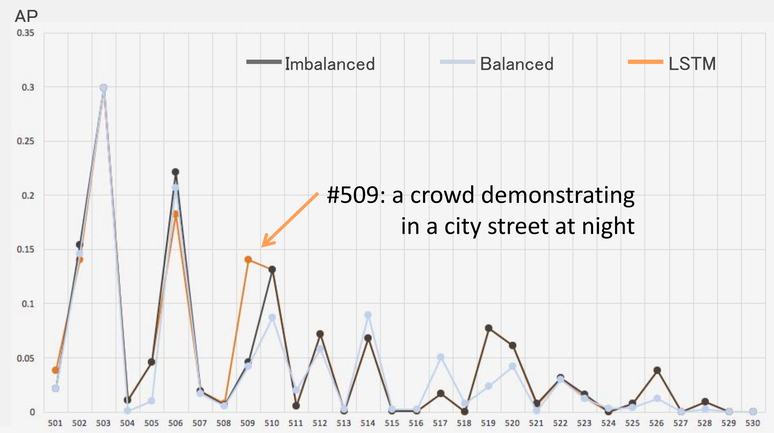
kobe_nict_siegen_D_M_3
(Unbalanced) LSTM

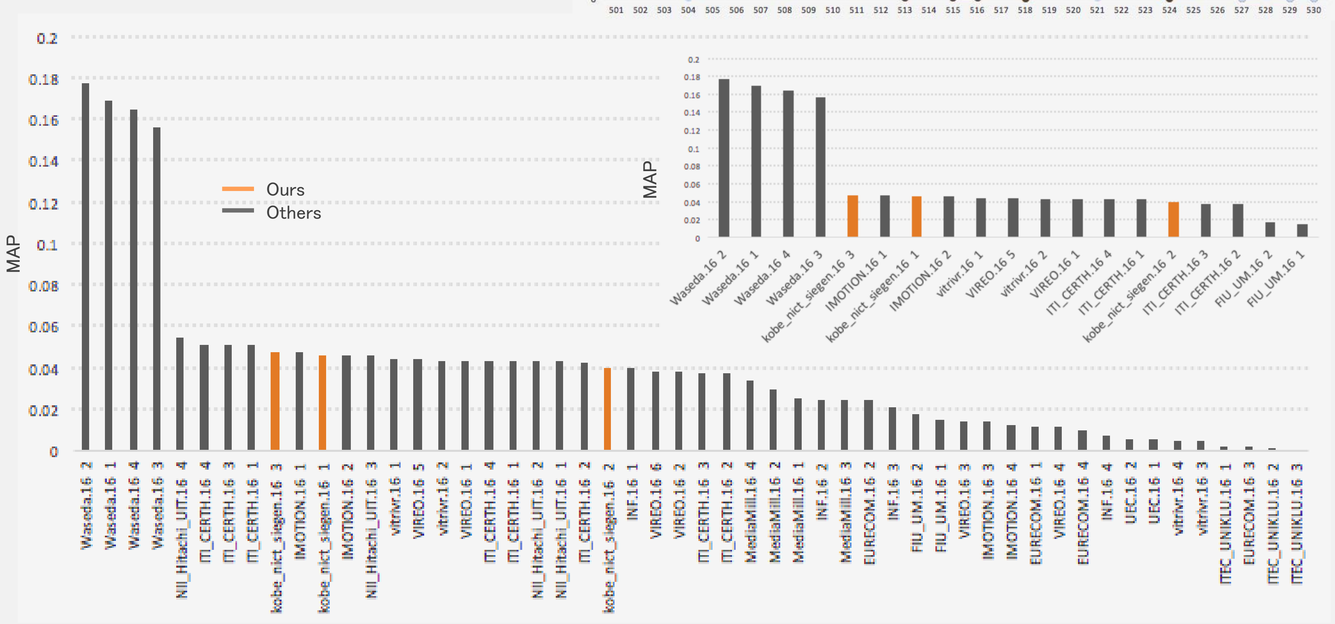
Each LSTM takes as input a vector of extracted by a pre-trained CNN. Unlike max-pooling, obtain temporal characteristics.



Results

- MicroNNs are worked efficiently
  - ➢ Easy to use, and so-so results
  - ➢ Enough speed for plenty of concepts
- Unbalanced condition is better than balanced
  - ➢ Should not persist data balance
- LSTM is also worked correctly
  - ➢ Much better result in some queries
- So much space for improvement…





Future works

- **More temporal resolution for LSTM:** 1 fps ⇒ 5 fps or more
- **Utilize outputs of several types of CNNs:** Scene CNN, Optical flow etc.

Correspond: Takashi Shinozaki (tshino@nict.go.jp) TRECVID 2016, Nov 14-16, NIST, Washington DC, US