# Kobe University,<br>NICT,<br>and University of Siegen<br>at TRECVID 2016 AVS Task

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

# Our Contribution

A method of using **small-scale neural network** to greatly accelerate concept classifier training.

**Transfer learning** can be used to acquire temporal characteristics effiently by combining both small networks and **LSTM**.

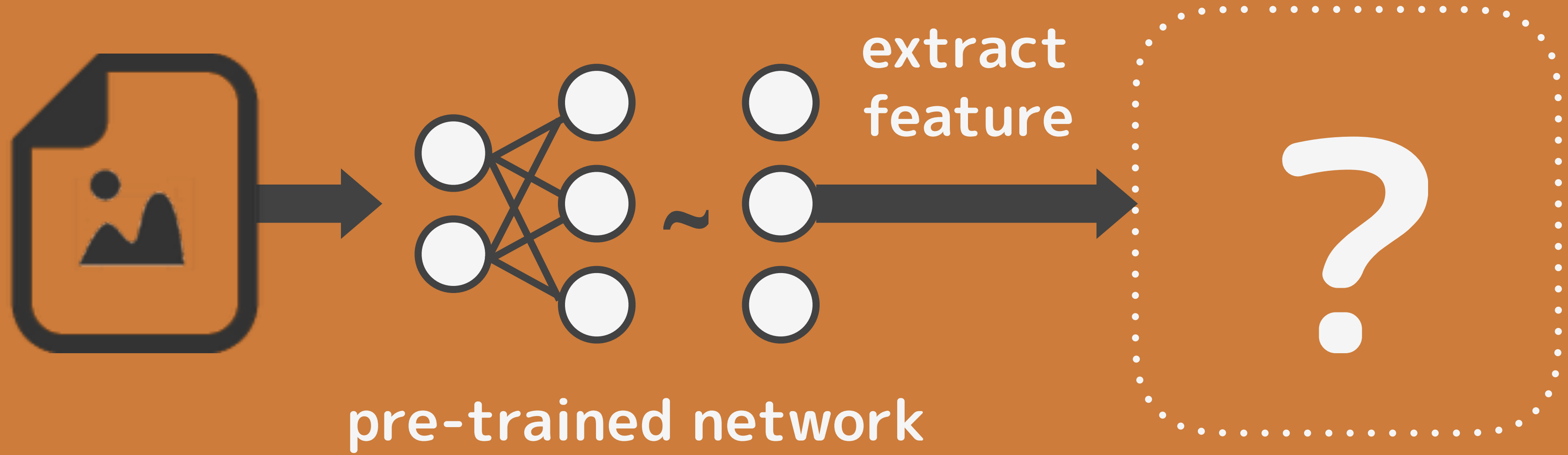
Evaluate the effectiveness of using **balanced examples** at the time of training.

## The Problem

Using pre-trained neural networks to extract features is a very popular approach.

However, training of classifiers takes long time.

This training gets even worse if classifiers required are many.



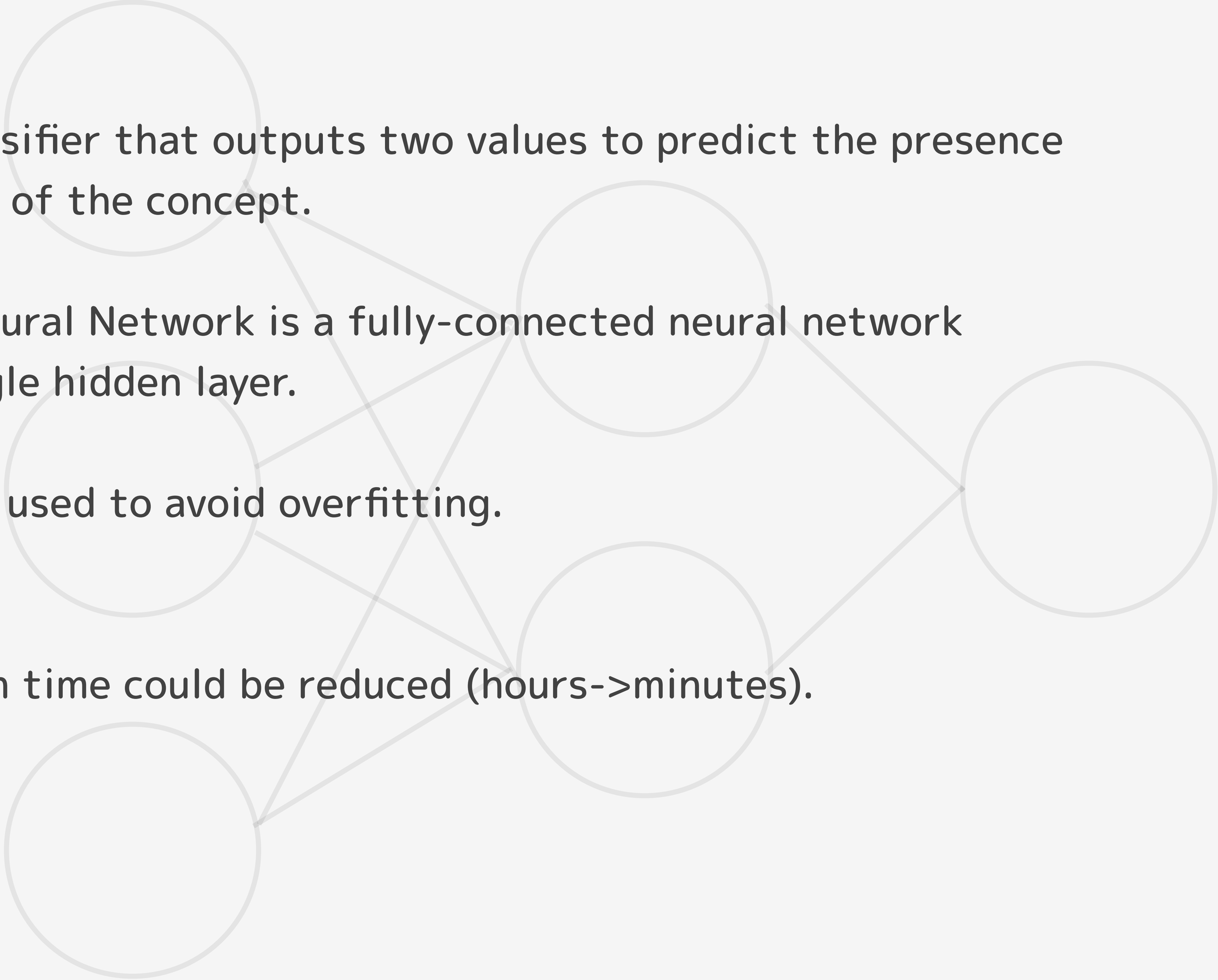
# Micro Neural Networks

Binary classifier that outputs two values to predict the presence or absence of the concept.

A micro Neural Network is a fully-connected neural network with a single hidden layer.

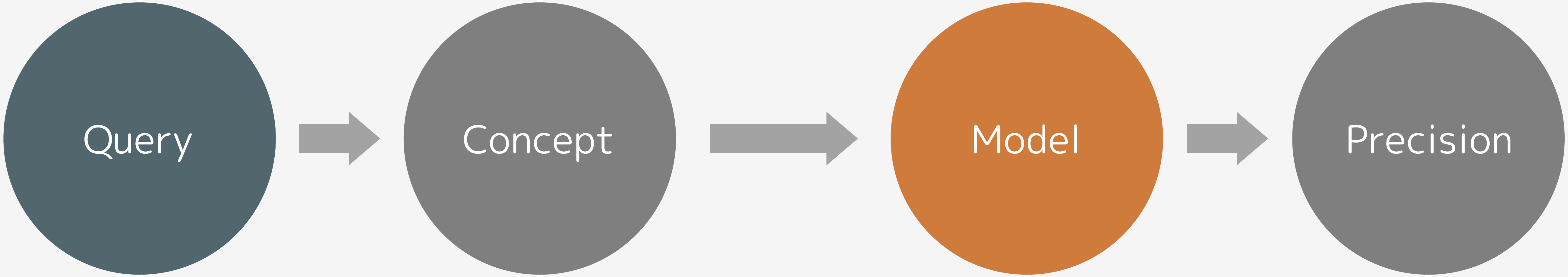
Dropout is used to avoid overfitting.

Calculation time could be reduced (hours->minutes).



# Our Approach - Overview

Overview of our method for TRECVID 2016 AVS task



Query → Concept → Model → Precision

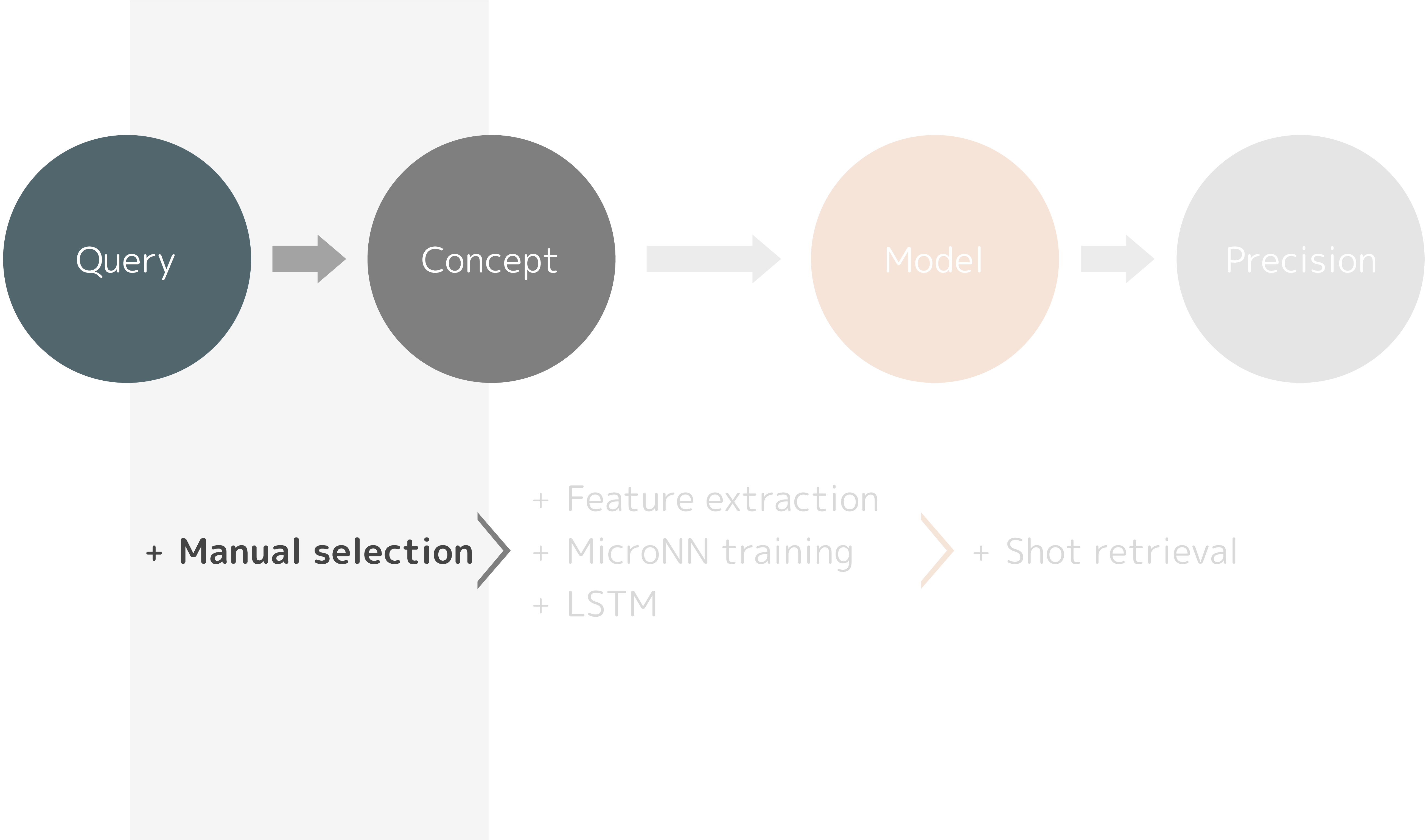
- Manual selection

- Feature extraction
- MicroNN training
- LSTM

- Shot retrieval

# Our Approach - Overview

## How we extracted concepts from the queries



Query → Concept → Model → Precision

- **Manual selection**

- Feature extraction
- MicroNN training
- LSTM

- Shot retrieval

# Our Approach - Manual Selection

Begin with manually selecting relevant concepts for each query

Simple rule is used to make it easier to automate the concept selection in the future.

“look’’

**Base form**

Query (502)

’’Find shots of a man indoors looking at camera where a bookcase is behind him’’

“man’’

**Pick only noun and verb**

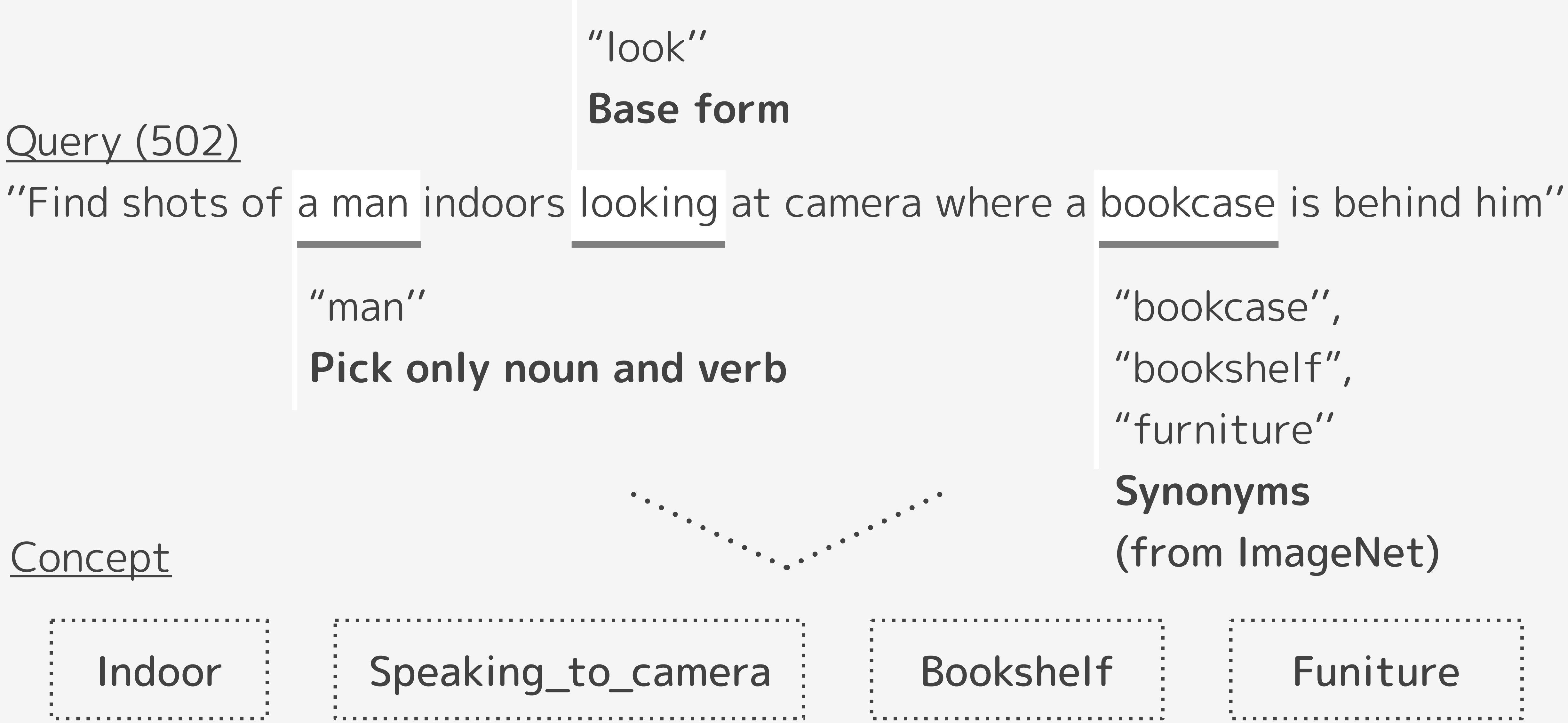
“bookcase’’,
“bookshelf”,
“furniture’’

**Synonyms**
**(from ImageNet)**

# Our Approach - Manual Selection

Begin with manually selecting relevant concepts for each query

Simple rule is used to make it easier to automate the concept selection in the future.



“look”

**Base form**

Query (502)

”Find shots of a man indoors looking at camera where a bookcase is behind him”

“man”

**Pick only noun and verb**

“bookcase”,

“bookshelf”,

“furniture”

**Synonyms**

**(from ImageNet)**

Concept

Indoor

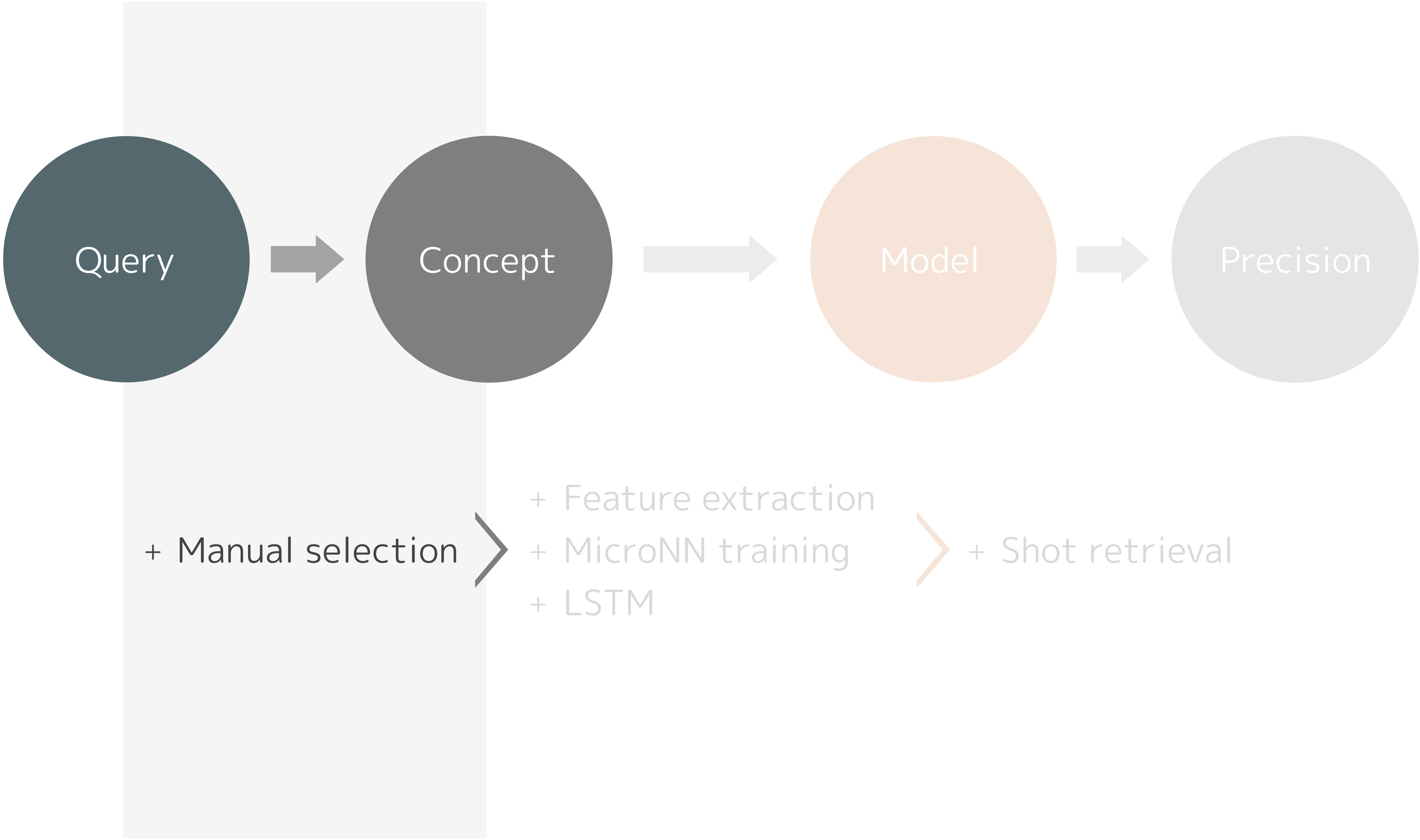
Speaking_to_camera

Bookshelf

Funiture

# Our Approach - Overview

Overview of our method for TRECVID 2016 AVS task



Query → Concept → Model → Precision

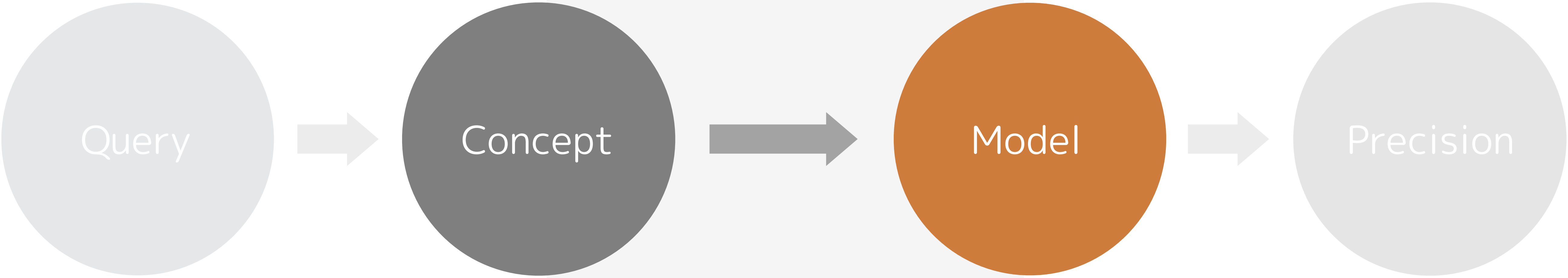
- Manual selection

- Feature extraction
- MicroNN training
- LSTM

- Shot retrieval

# Our Approach - Overview

Combine the concepts from each query.



Query

Concept

Model

Precision

- Manual selection

- Feature extraction
- MicroNN training
- LSTM

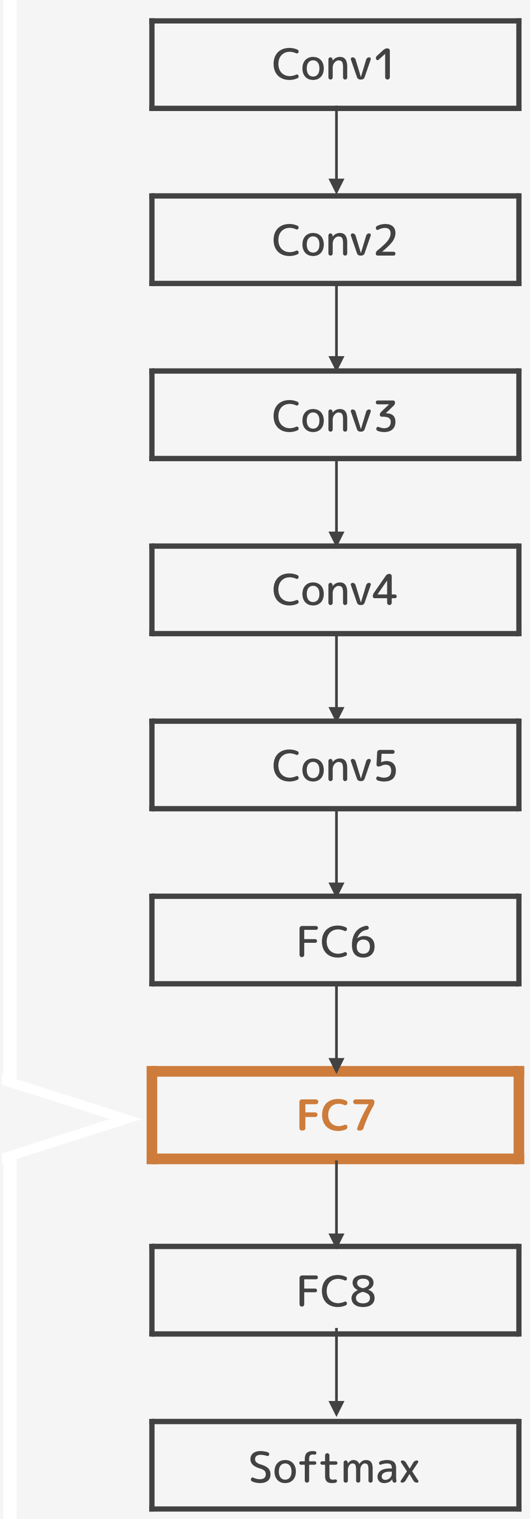
- Shot retrieval

# Our Approach - Feature Extraction

Pre-trained network is usually transferred into classifiers suitable for the target problem

Use pre-trained VGGNet

- ILSVRC 2014
- CNN with very deep architecture
- The 16 layer version is used
- FC7 : Use output at the second fully connected layer

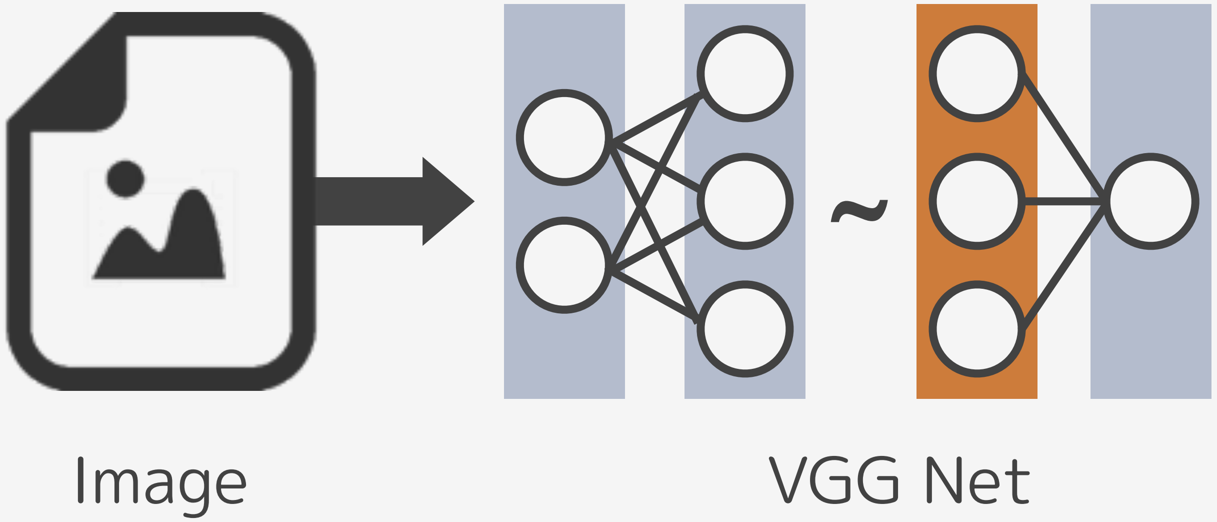


K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition”

# Our Approach - MicroNN Training

Perform gradual transfer learning for each concept in the following step

① Start with training microNN using **images**

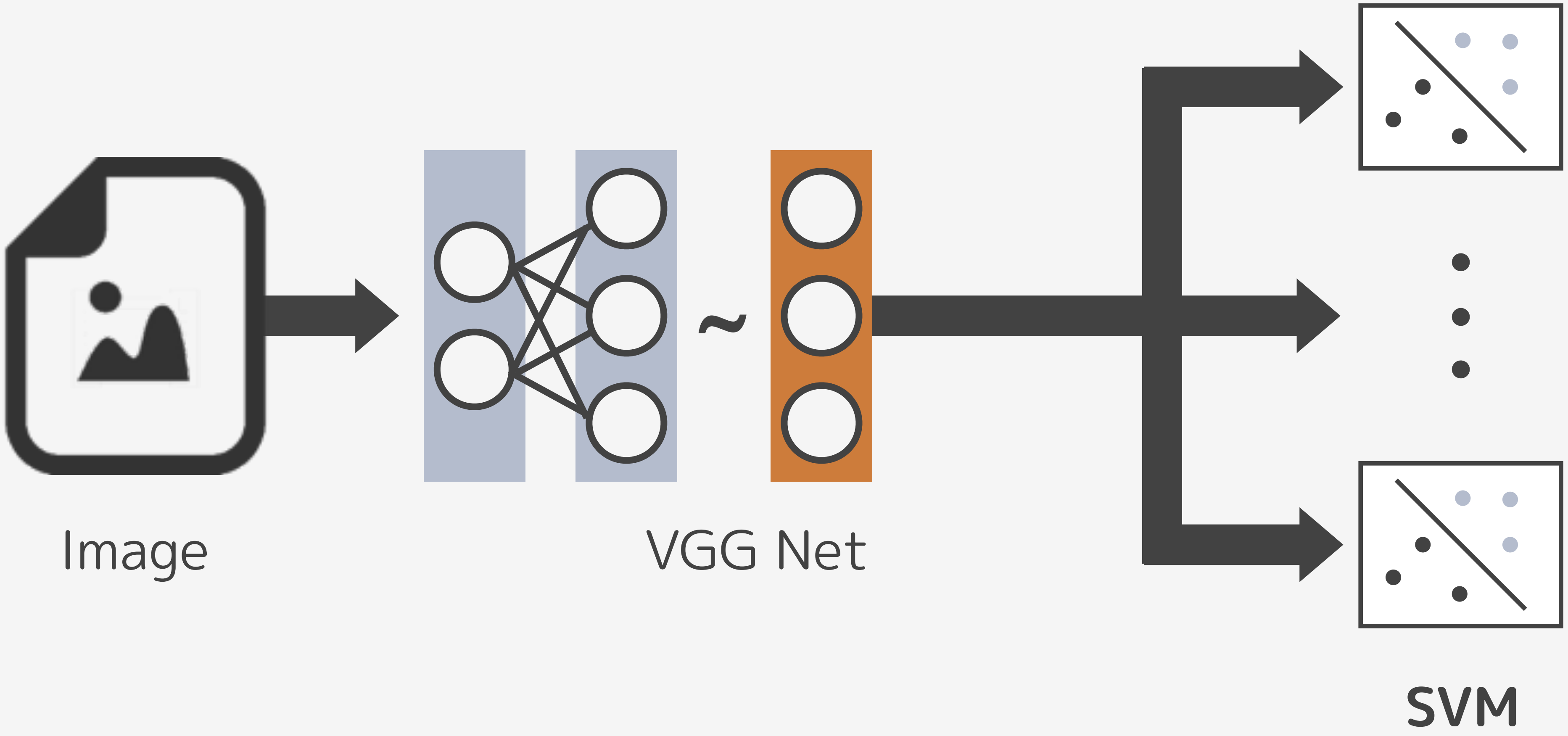


# Previous Approach - SVM Training

Until now . . .

Previous studies have trained classifiers such as SVM by extracted features.

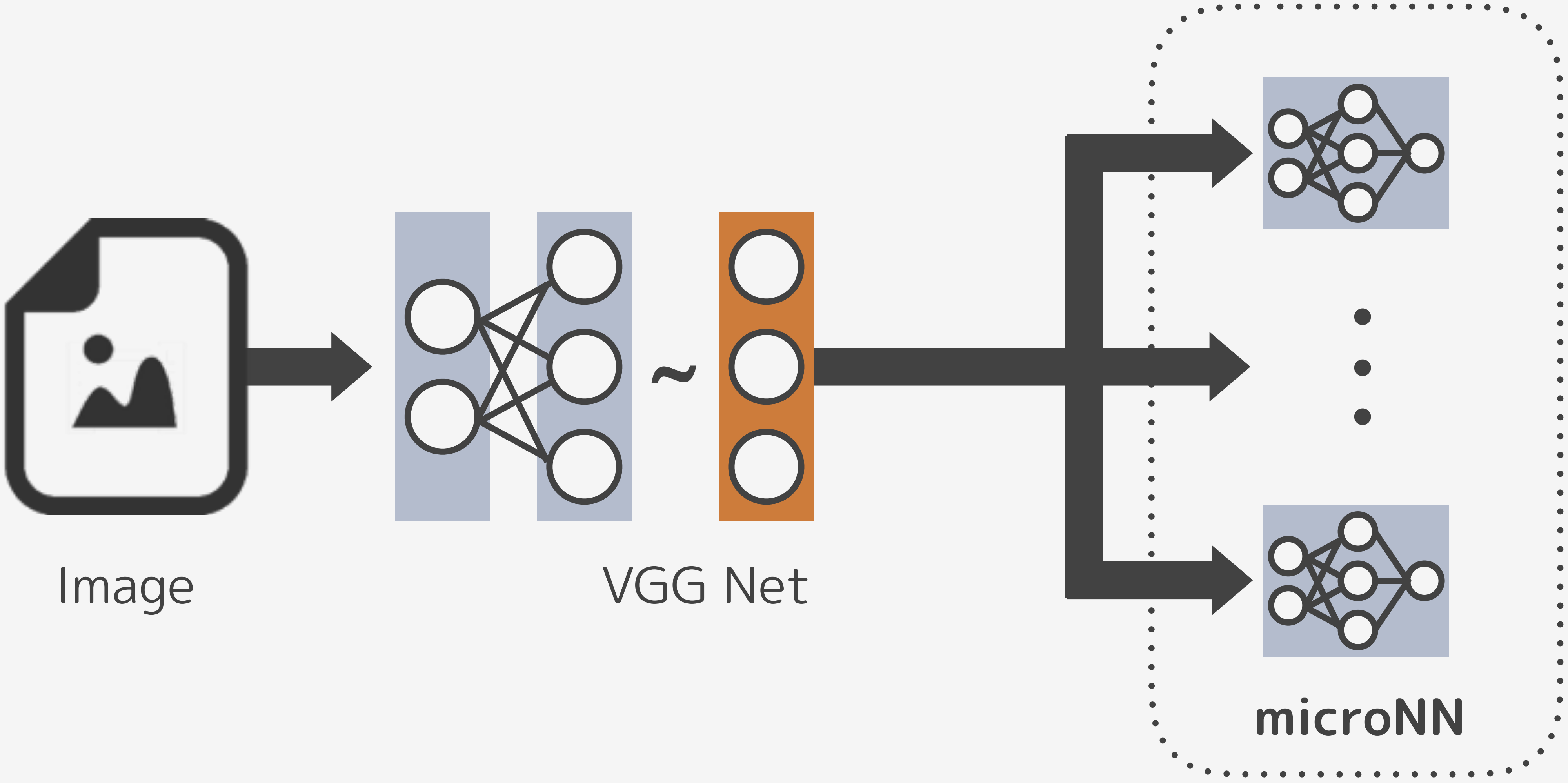
This requires a lot of time.



# Our Approach - MicroNN Training

Perform gradual transfer learning for each concept in the following step

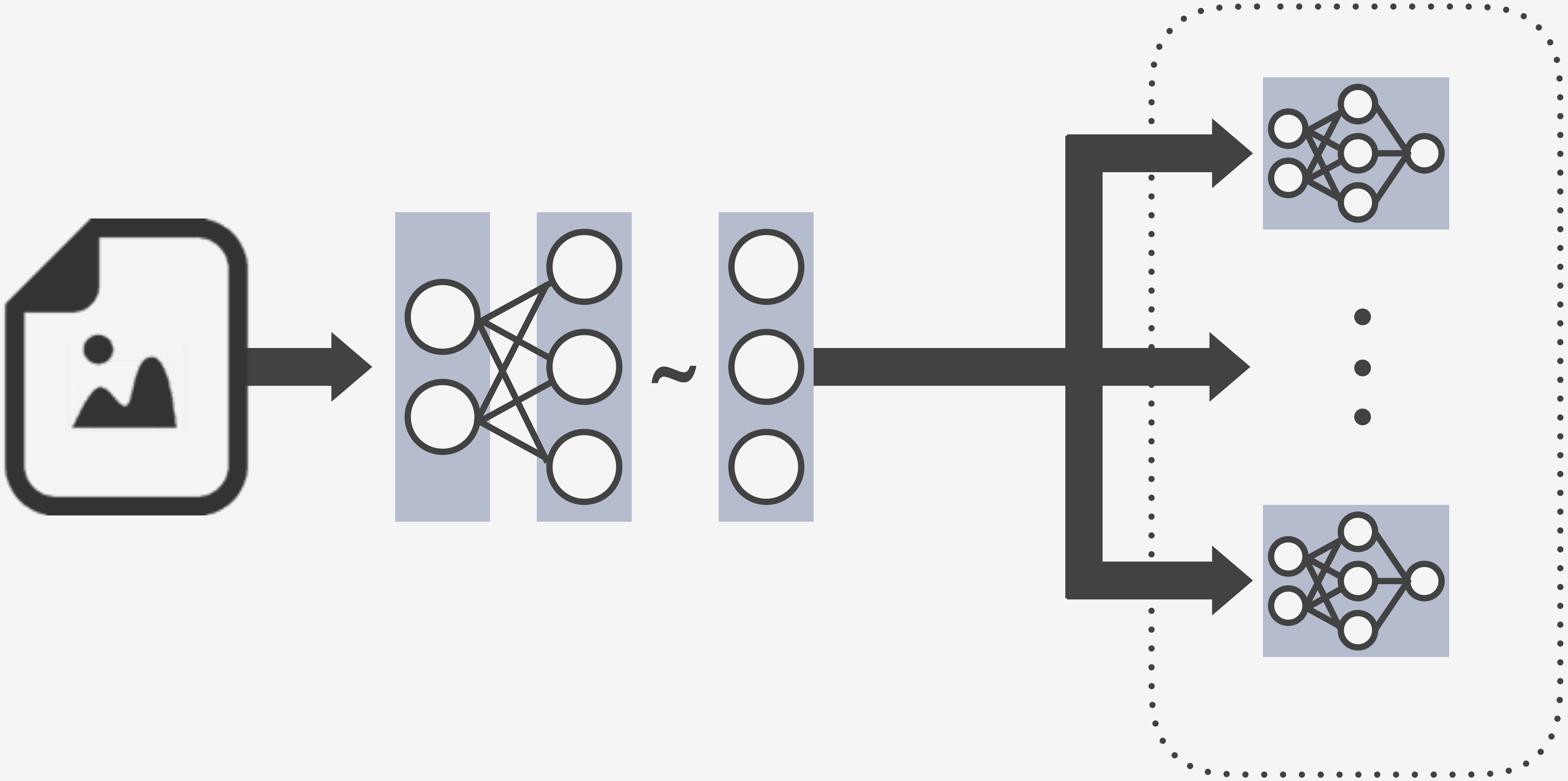
① Start with training microNN using **images**



# Our Approach - MicroNN Training

Perform gradual transfer learning for each concept in the following step

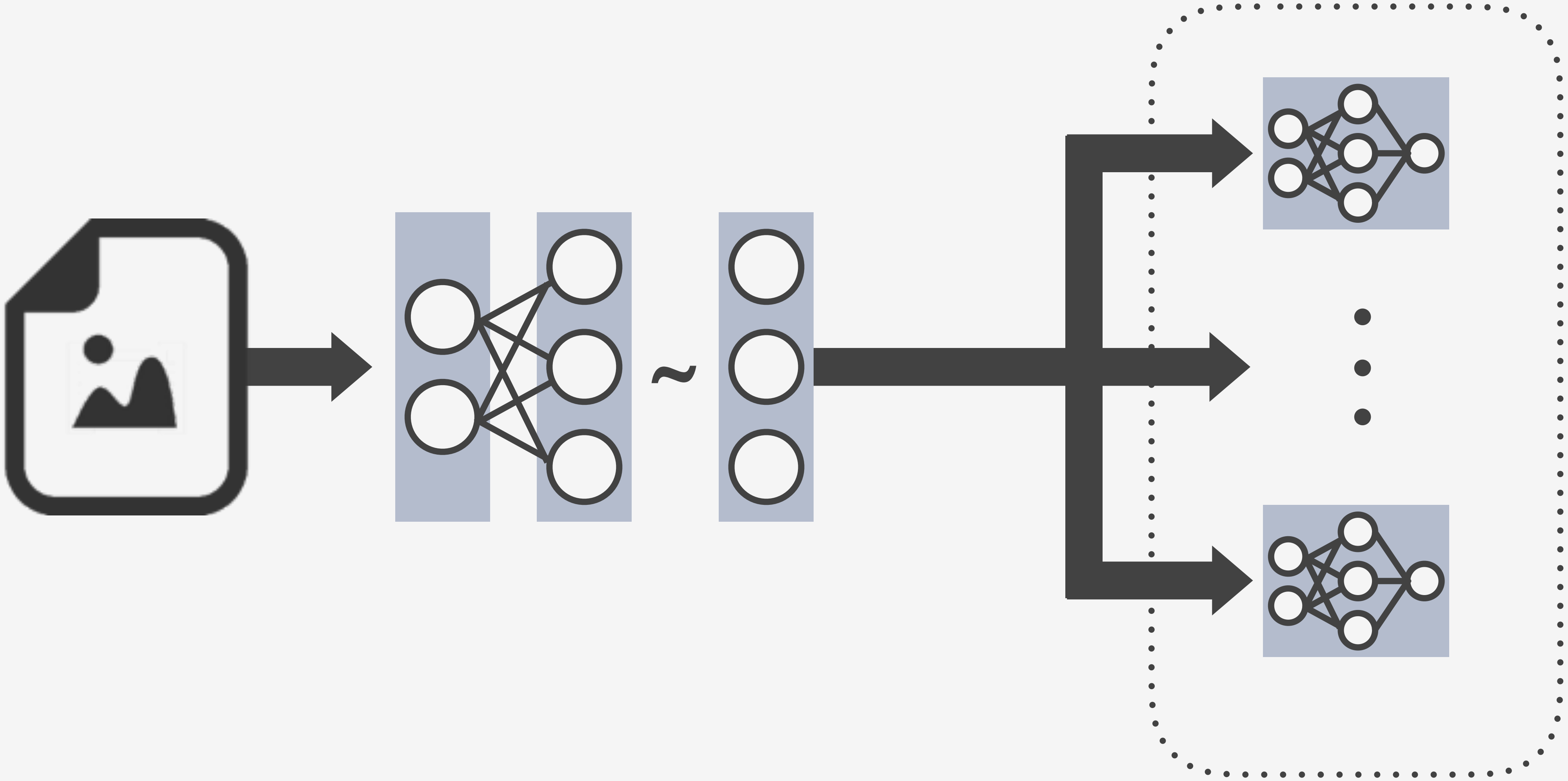
① Start with training microNN using **images**



# Our Approach - MicroNN Training

Perform gradual transfer learning for each concept in the following step

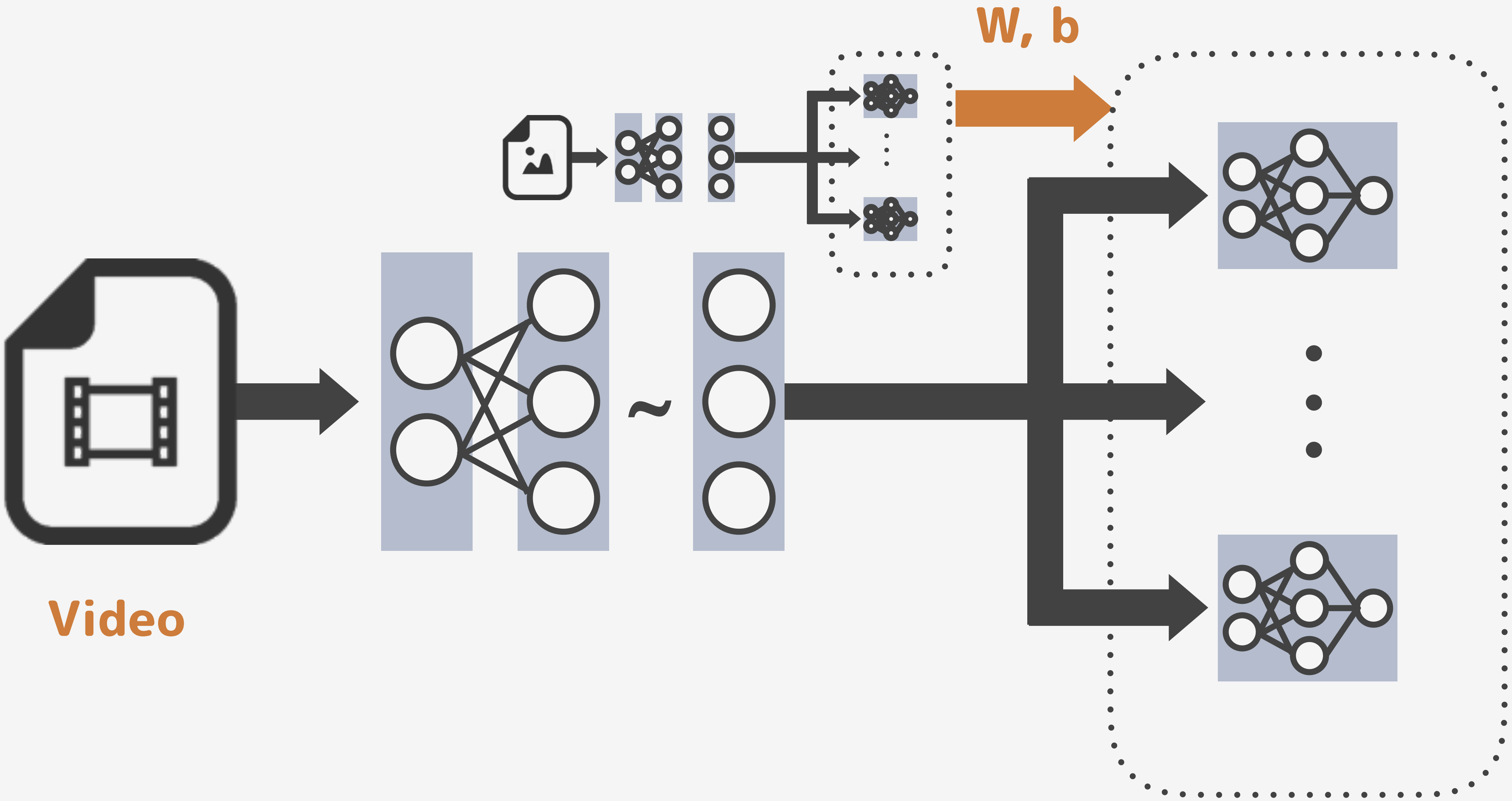
② Refine the microNN using shots in video dataset.



# Our Approach - MicroNN Training

Perform gradual transfer learning for each concept in the following step

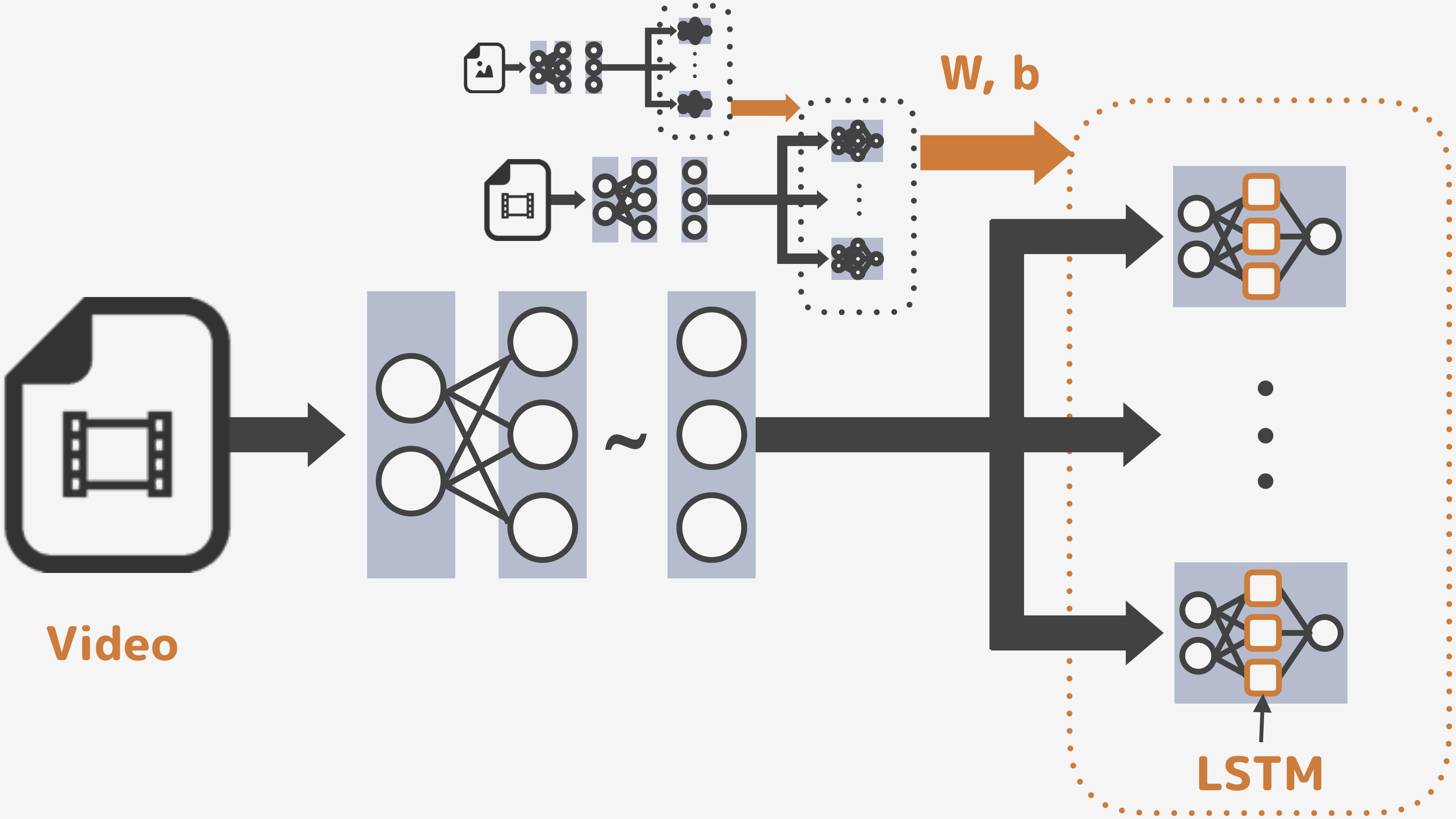
② Refine the microNN using shots in video dataset. The microNN has weight parameters learned at first step as its initial value.



# Our Approach - MicroNN Training

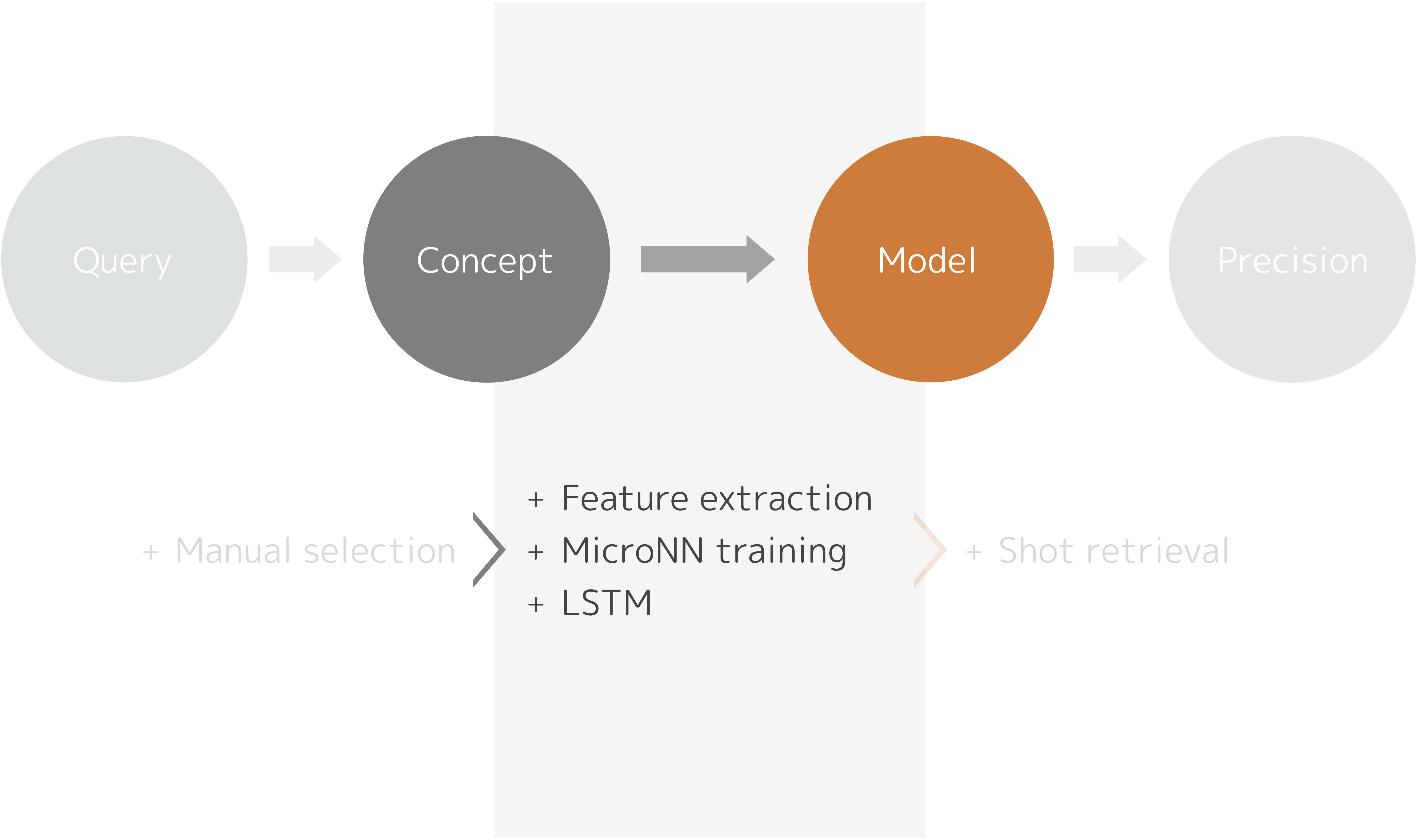
Perform gradual transfer learning for each concept in the following step

③ Futher, hidden layer of microNN is replaced with LSTM for acquiring temporal characteristics. Refine the microNN starting with weight parameters learned at the second step as initial values.



# Our Approach - Overview

Overview of our method for TRECVID 2016 AVS task



Query

Concept

Model

Precision

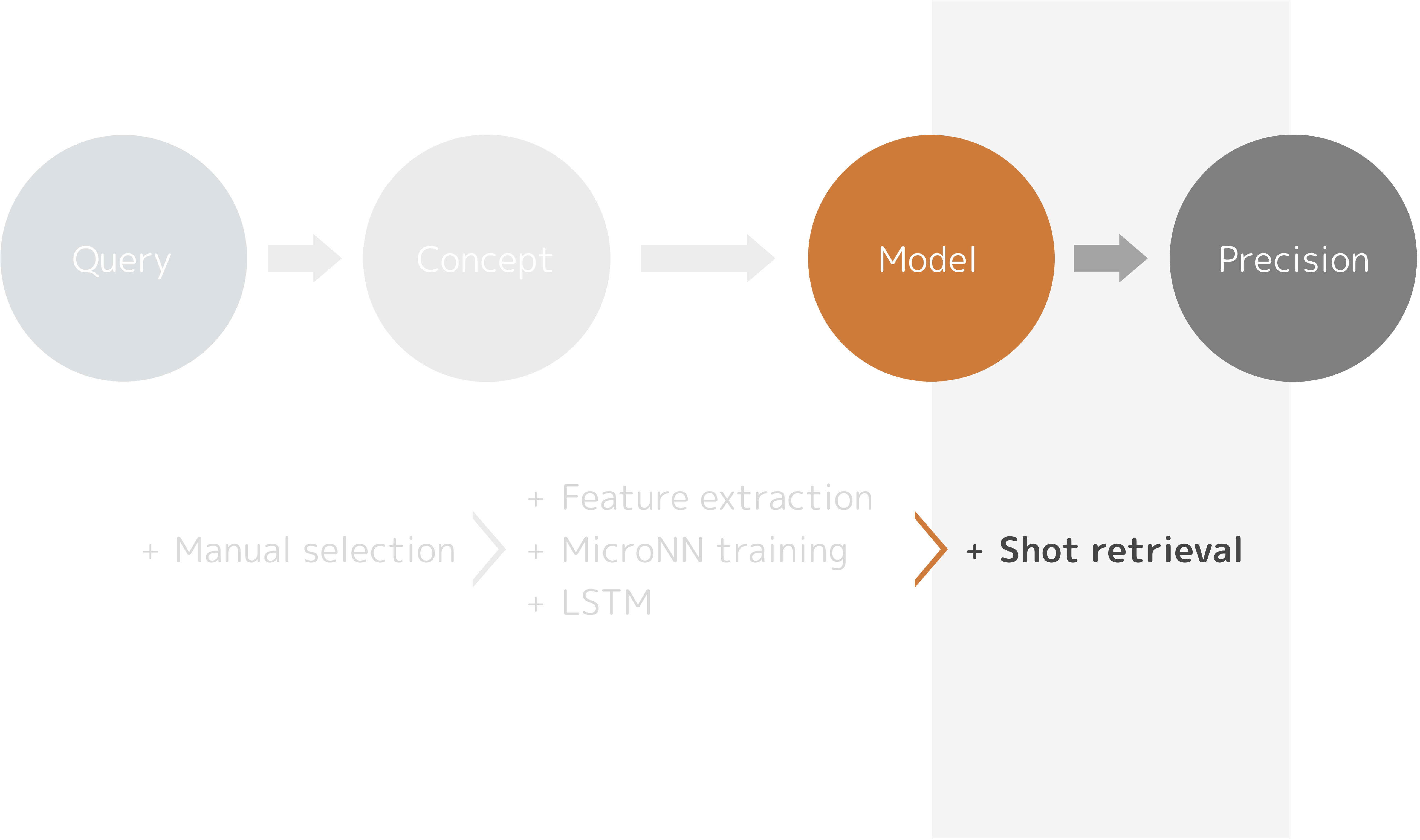
- Manual selection

- Feature extraction
- MicroNN training
- LSTM

- Shot retrieval

# Our Approach - Overview

How we go from a shot's concept relevance to its search score



Query

Concept

Model

Precision

- Manual selection

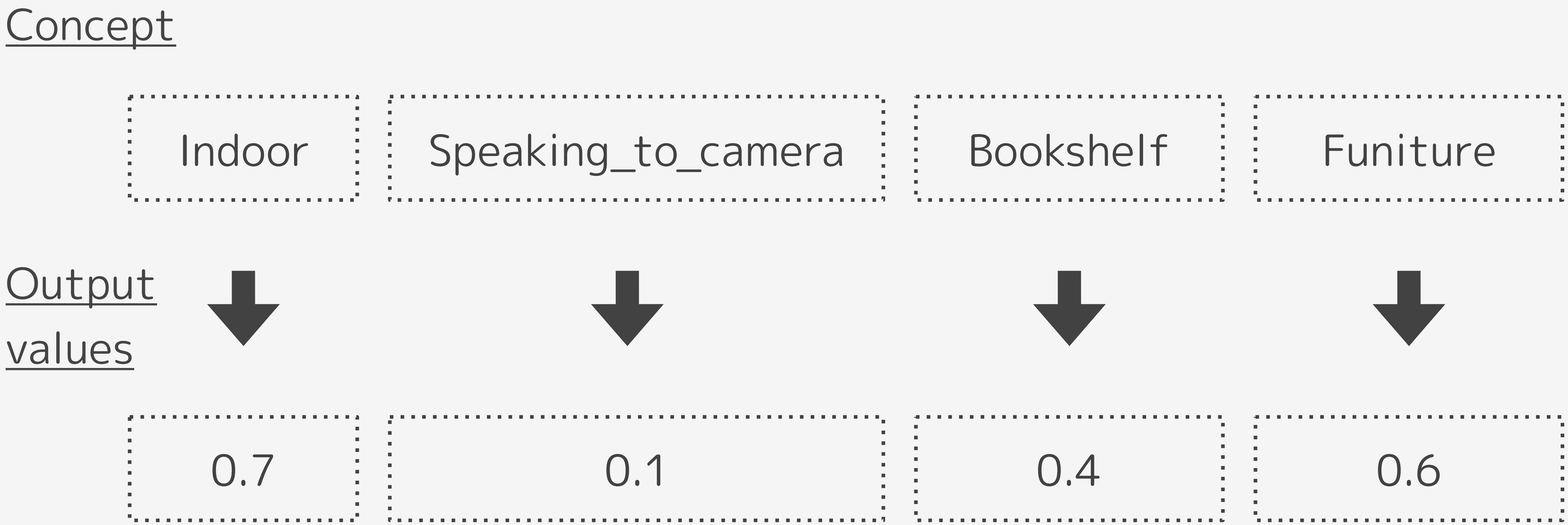
- Feature extraction
- MicroNN training
- LSTM

- **Shot retrieval**

# Our Approach - Shot Retrieval

For each shot, calculate the avarage of output values of microNNs for the selected concepts in a query

MicroNN outputs are normalized to [-1, 1], to balance between different concepts.

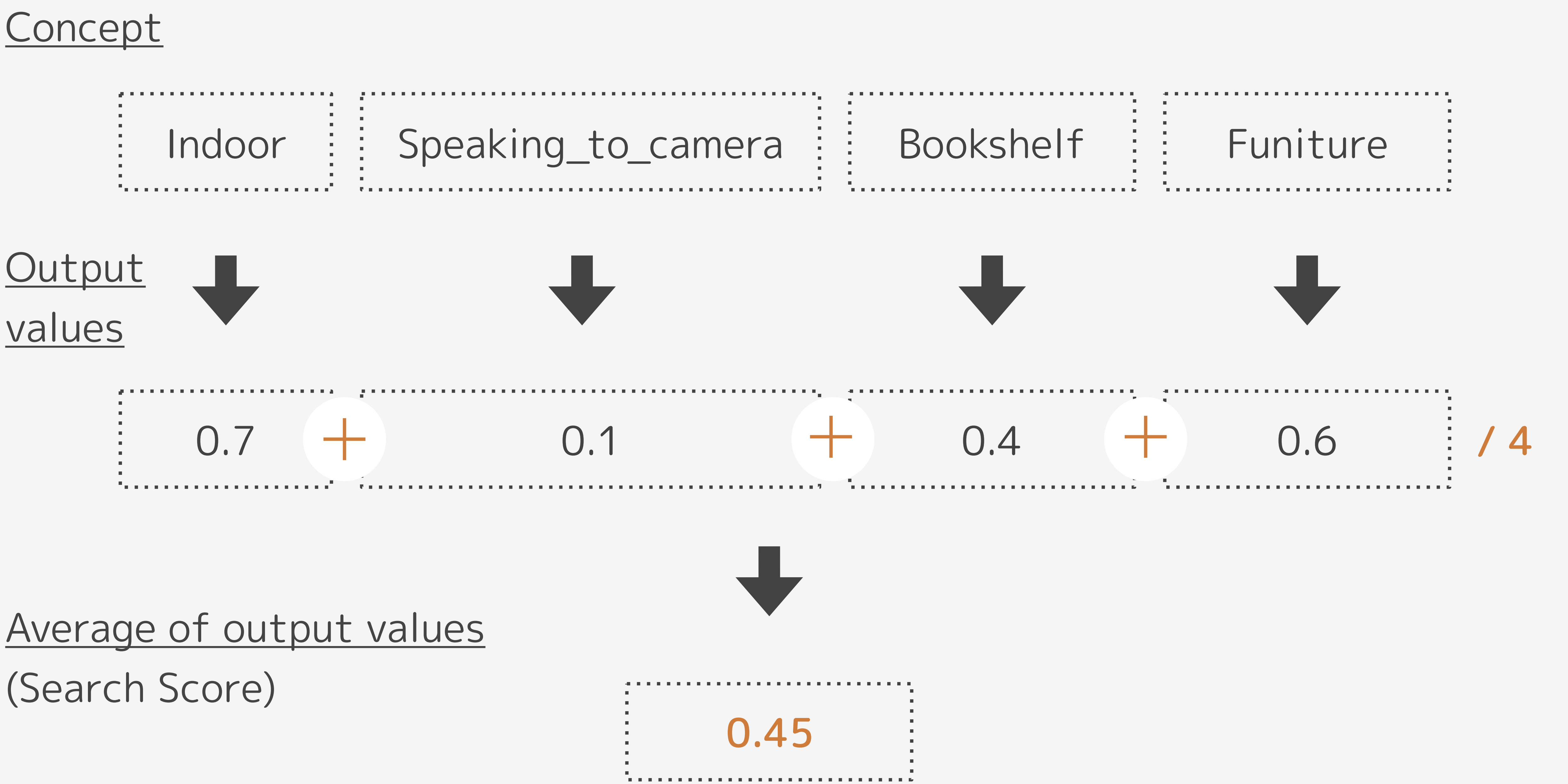


| Concept | Indoor | Speaking_to_camera | Bookshelf | Funiture |
|---|---|---|---|---|
| Output values | 0.7 | 0.1 | 0.4 | 0.6 |

# Our Approach - Shot Retrieval

## How do we compare that with other shots

Calculate the average of output values and use it as overall search score.



Concept

Indoor | Speaking_to_camera | Bookshelf | Funiture

Output values

0.7 + 0.1 + 0.4 + 0.6 / 4

Average of output values (Search Score)

0.45

# Purpose of Experiment

1. Evaluate the learning speed.
2. Evaluate the effectiveness of using LSTM to acquire temporal characteristics.
3. Evaluate wheather using same number of positive and negative examples (“Balanced”) for training improves classification.

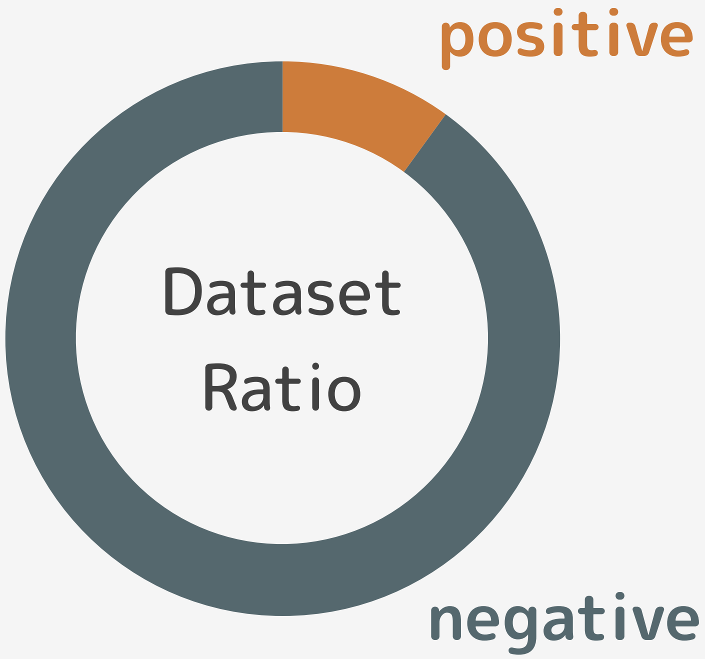
# Experiment - Three Runs

Submitted the following for TRECVID 2016 AVS task

kobe_nict_siegen_D_M_**1**

## Imbalanced

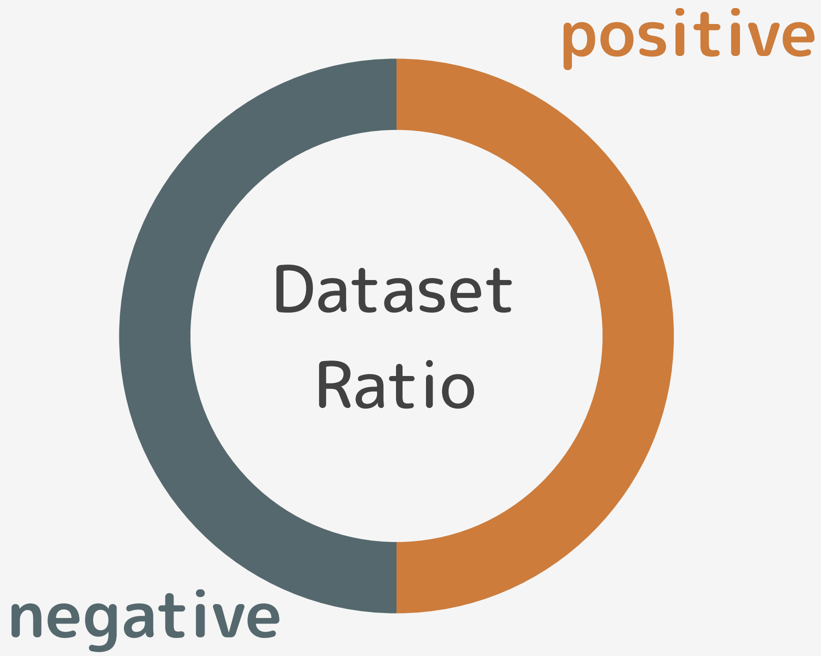
Fine-tuning is carried out using **imbalanced** numbers of positive and negative examples. (30,000 total)



kobe_nict_siegen_D_M_**2**

## Balanced

Fine-tuning is carried out using **balanced** numbers of positive and negative examples. (30,000 total)



kobe_nict_siegen_D_M_**3**

## (Imbalanced) LSTM

Unlike max-pooling, LSTM obtains temporal characteristics. LSTM-based microNNs are trained only for 14 concepts for which temporal relations among video frames are important

Only 14 concepts

# Experiment - Dataset

Used in this study

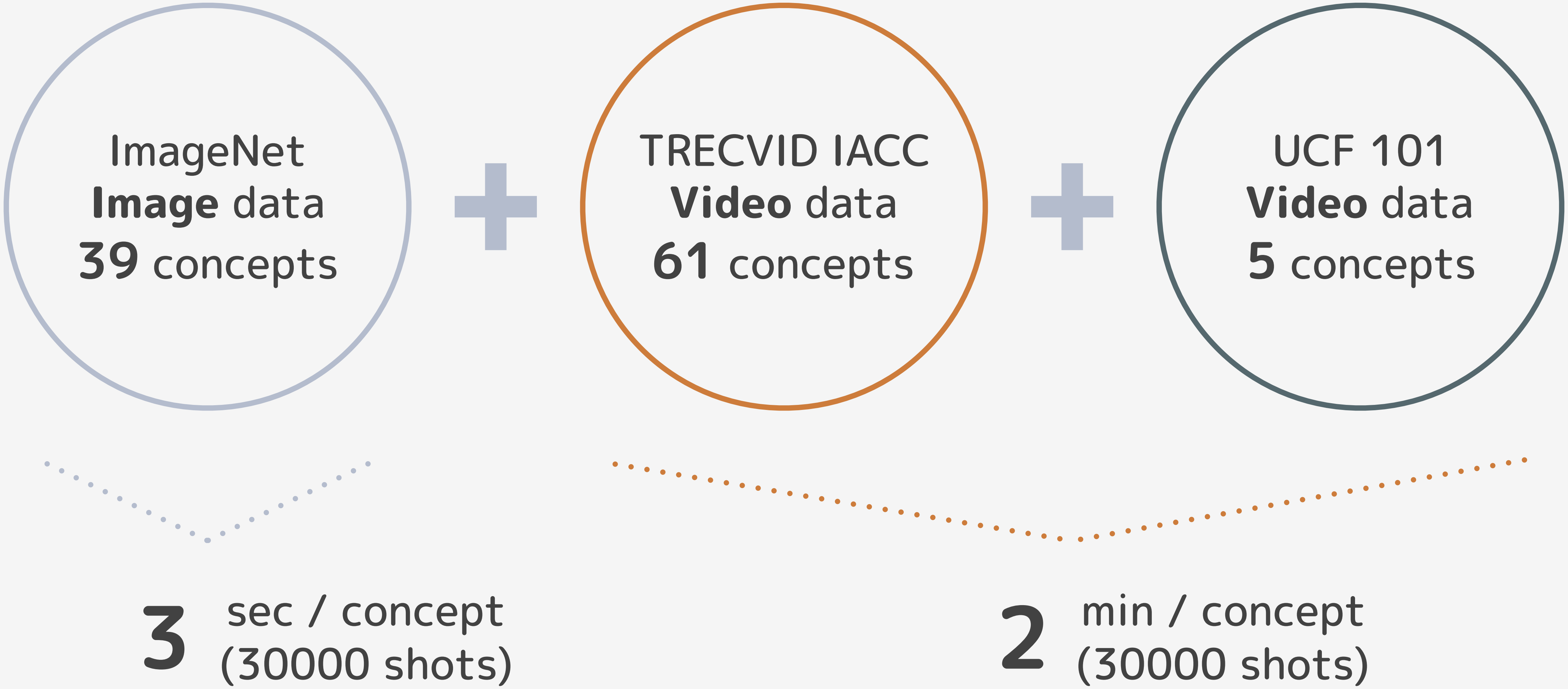
ImageNet
**Image** data
**39** concepts

TRECVID IACC
**Video** data
**61** concepts

UCF 101
**Video** data
**5** concepts

# Experiment - Dataset

## Training time



ImageNet
**Image** data
**39** concepts

TRECVID IACC
**Video** data
**61** concepts

UCF 101
**Video** data
**5** concepts

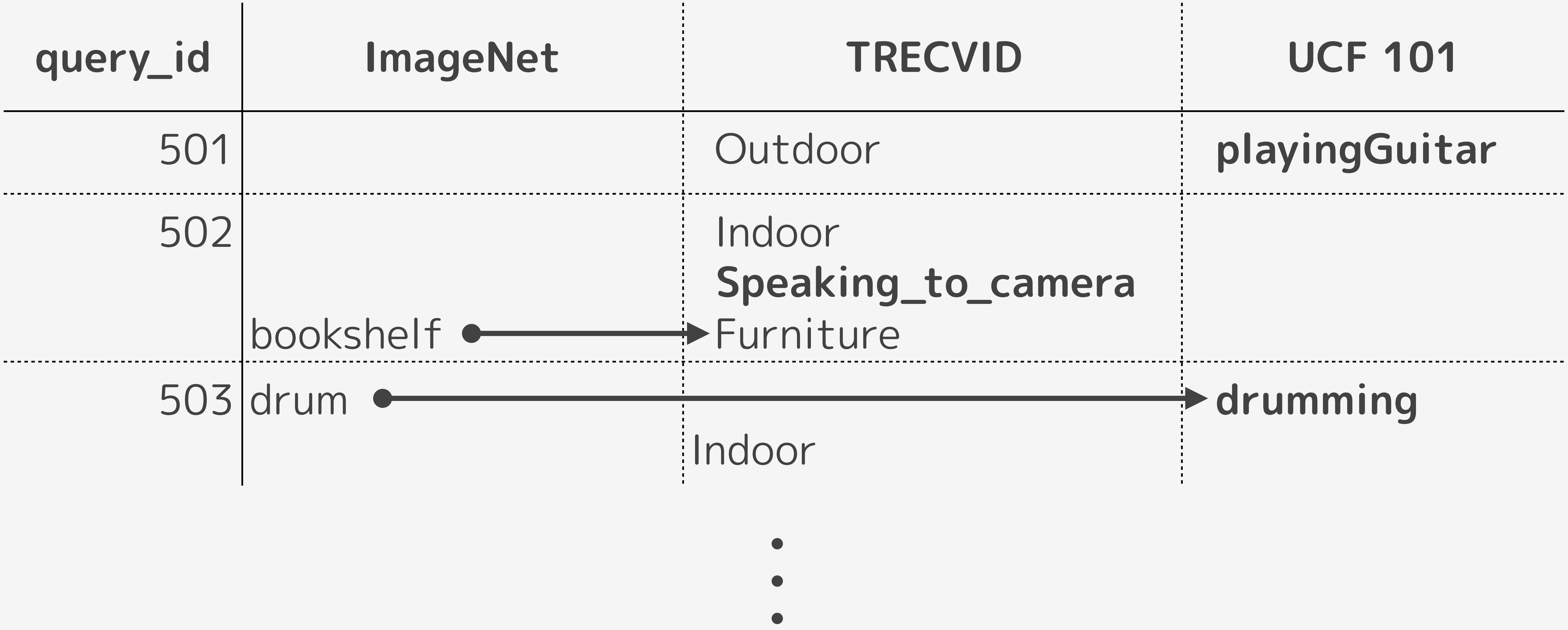
**3** sec / concept
(30000 shots)

**2** min / concept
(30000 shots)

# Experiment - Dataset

## Used in this study

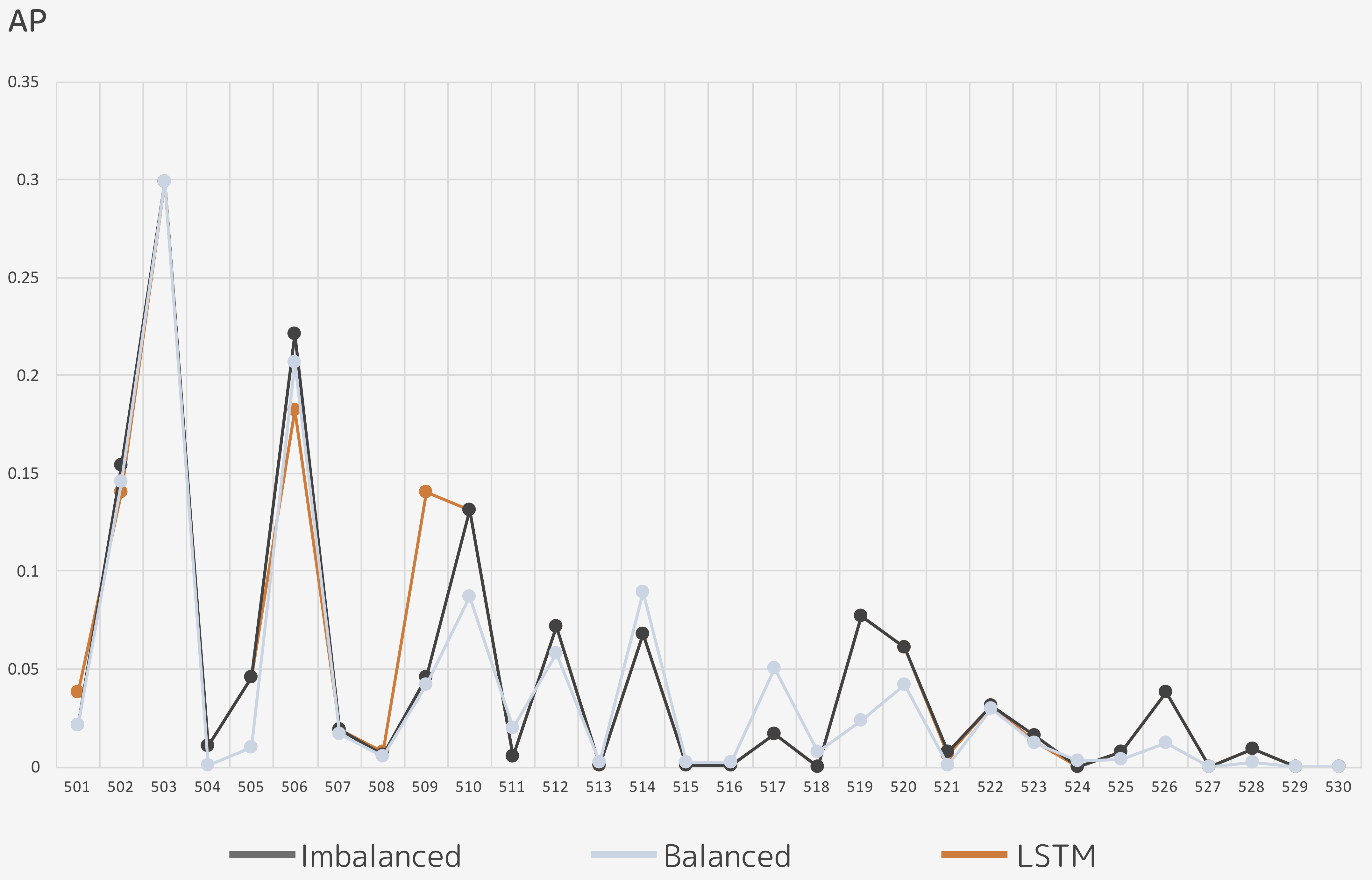
List of some concepts selected for each query



| query_id | ImageNet | TRECVID | UCF 101 |
|---|---|---|---|
| 501 | | Outdoor | **playingGuitar** |
| 502 | bookshelf | Indoor<br>**Speaking_to_camera**<br>Furniture | |
| 503 | drum | Indoor | **drumming** |

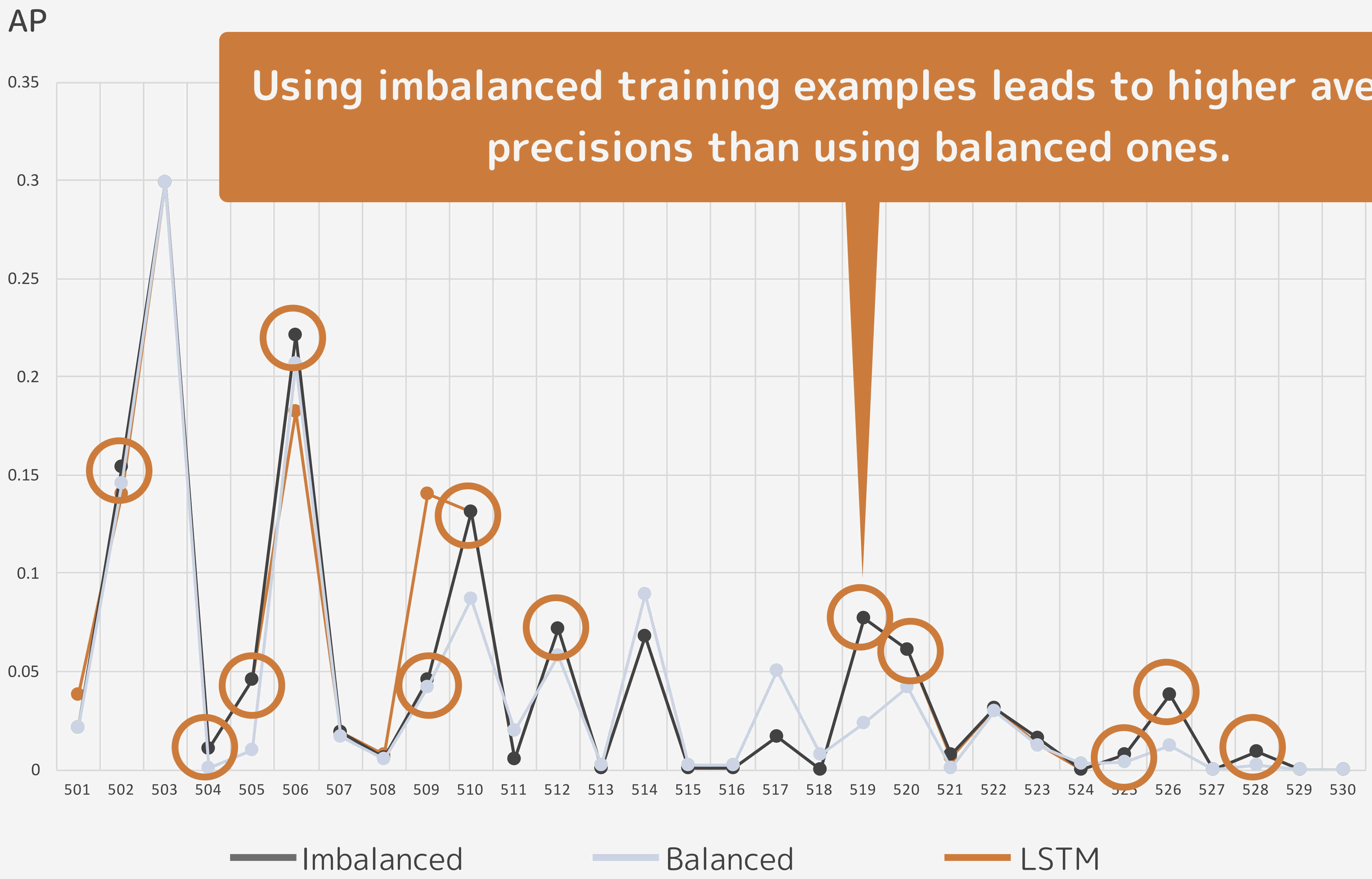
# Experiment - Result

Performance comparison between Imbalanced, Balanced and LSTM on each of the 30 queries



# Experiment - Result

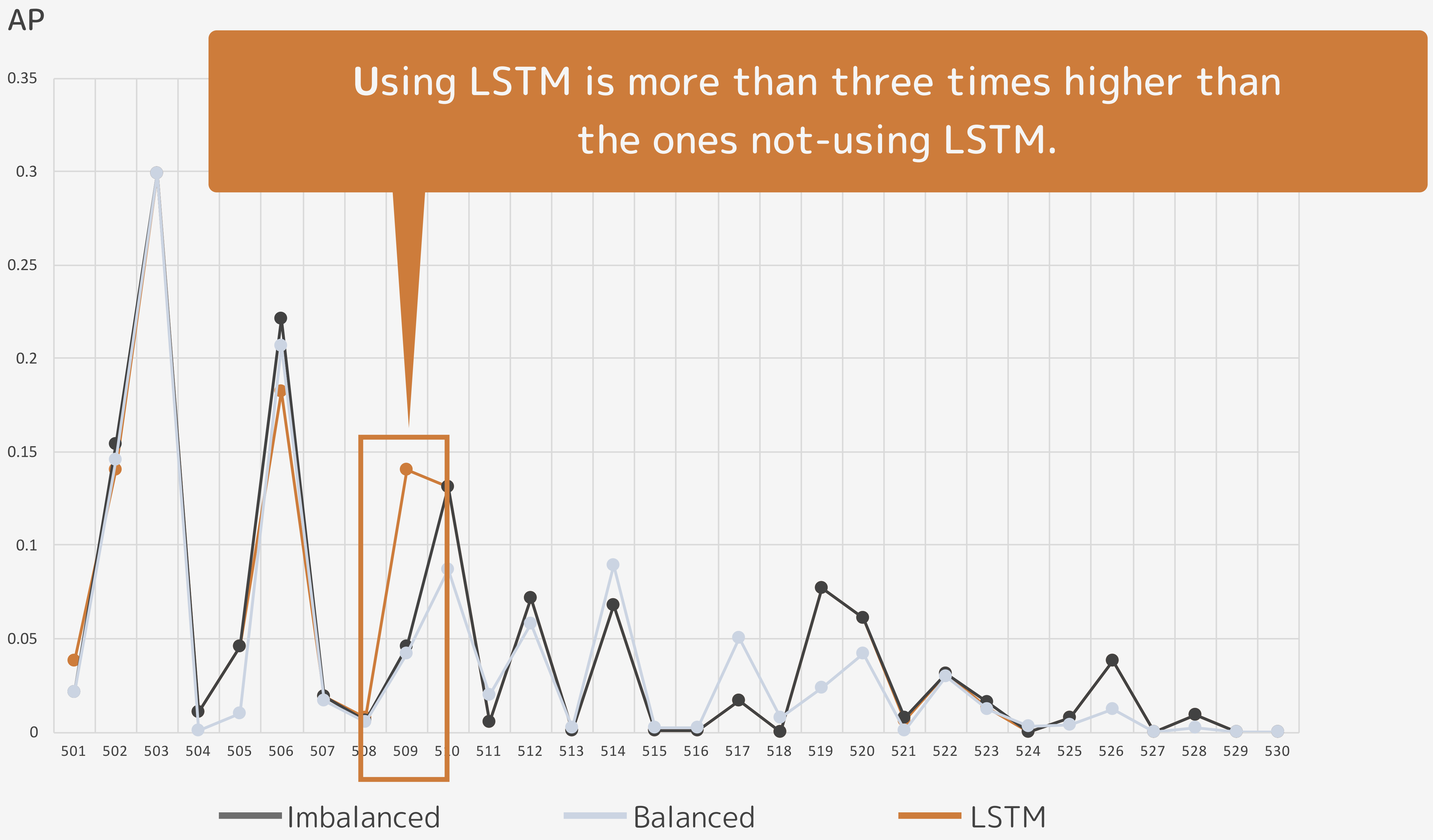
Performance comparison between Imbalanced, Balanced and LSTM on each of the 30 queries



Using imbalanced training examples leads to higher average precisions than using balanced ones.

# Experiment - Result

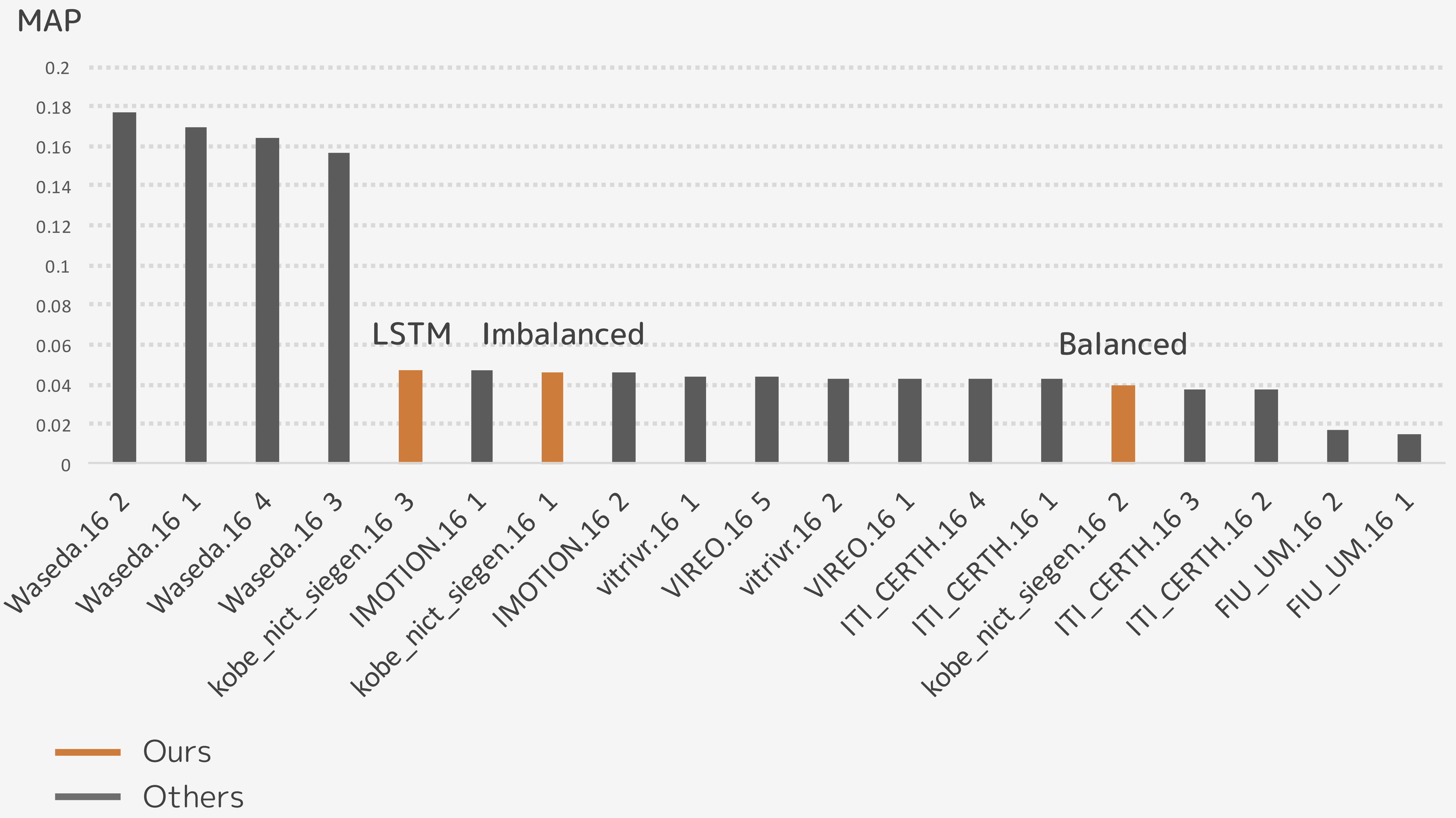
Performance comparison between Imbalanced, Balanced and LSTM on each of the 30 queries



Using LSTM is more than three times higher than the ones not-using LSTM.

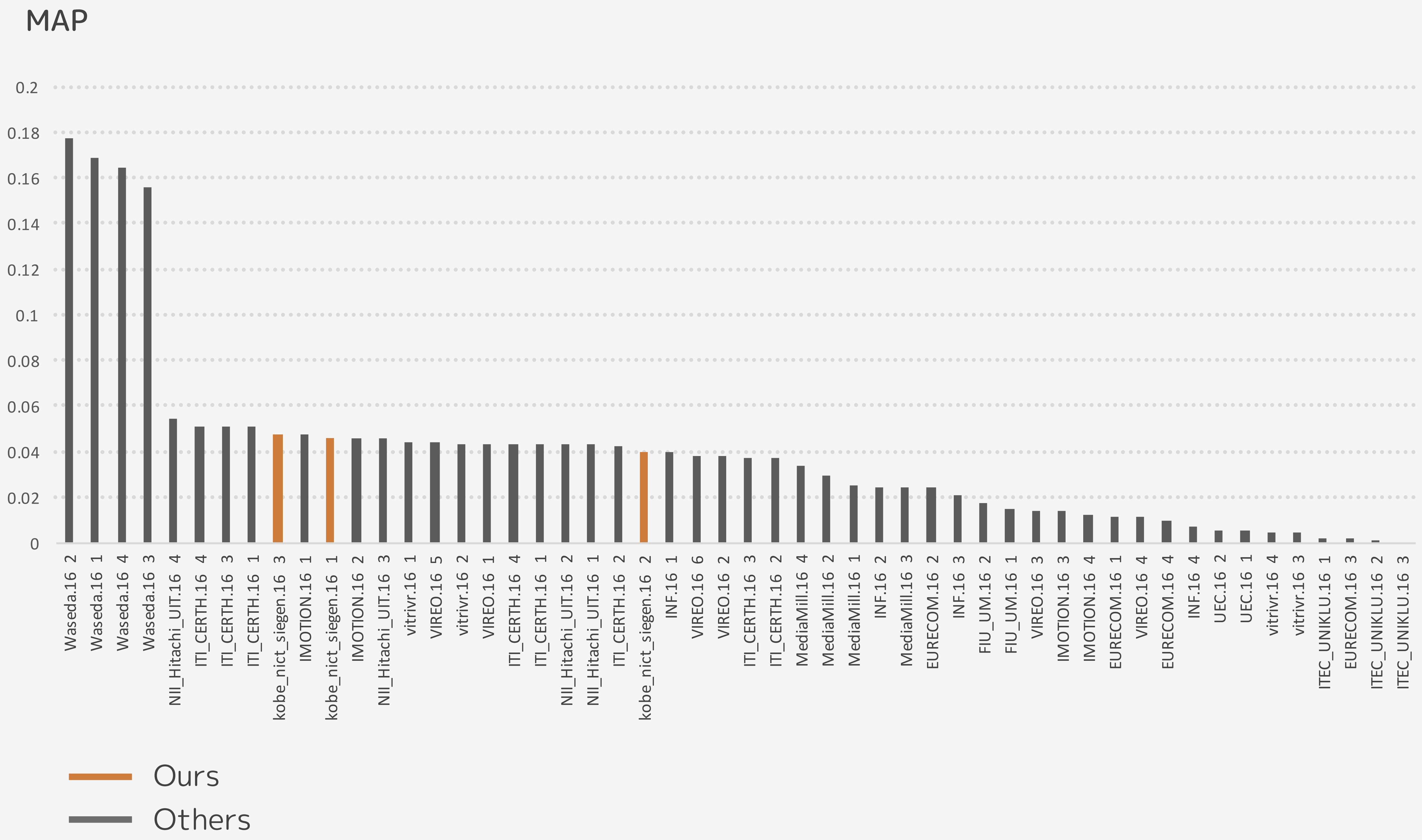
# Experiment - Result

Performance comparison between our method and the other methods developed for the manually-assisted category in AVS task



# Experiment - Result

## Performance comparison between our method and the other methods developed for the AVS task



# Conclusion

Video search through efficient transfer learning using microNN

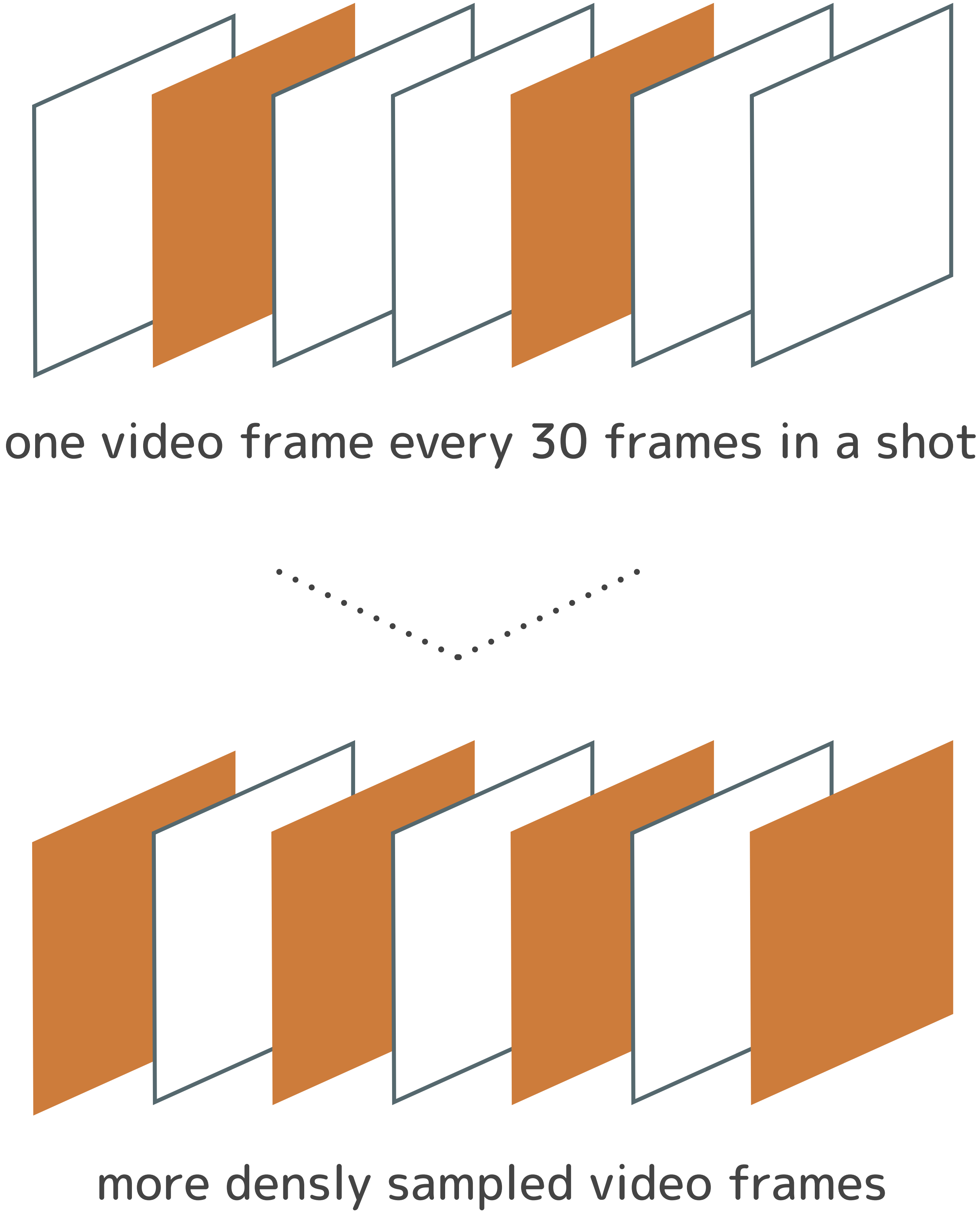
- fast
- flexibile

Imbalanced examples are more useful than balanced examples

Validity of acquired temporal characteristics by LSTM

# Future work

Further experiments by using LSTM on reduced frame interval.



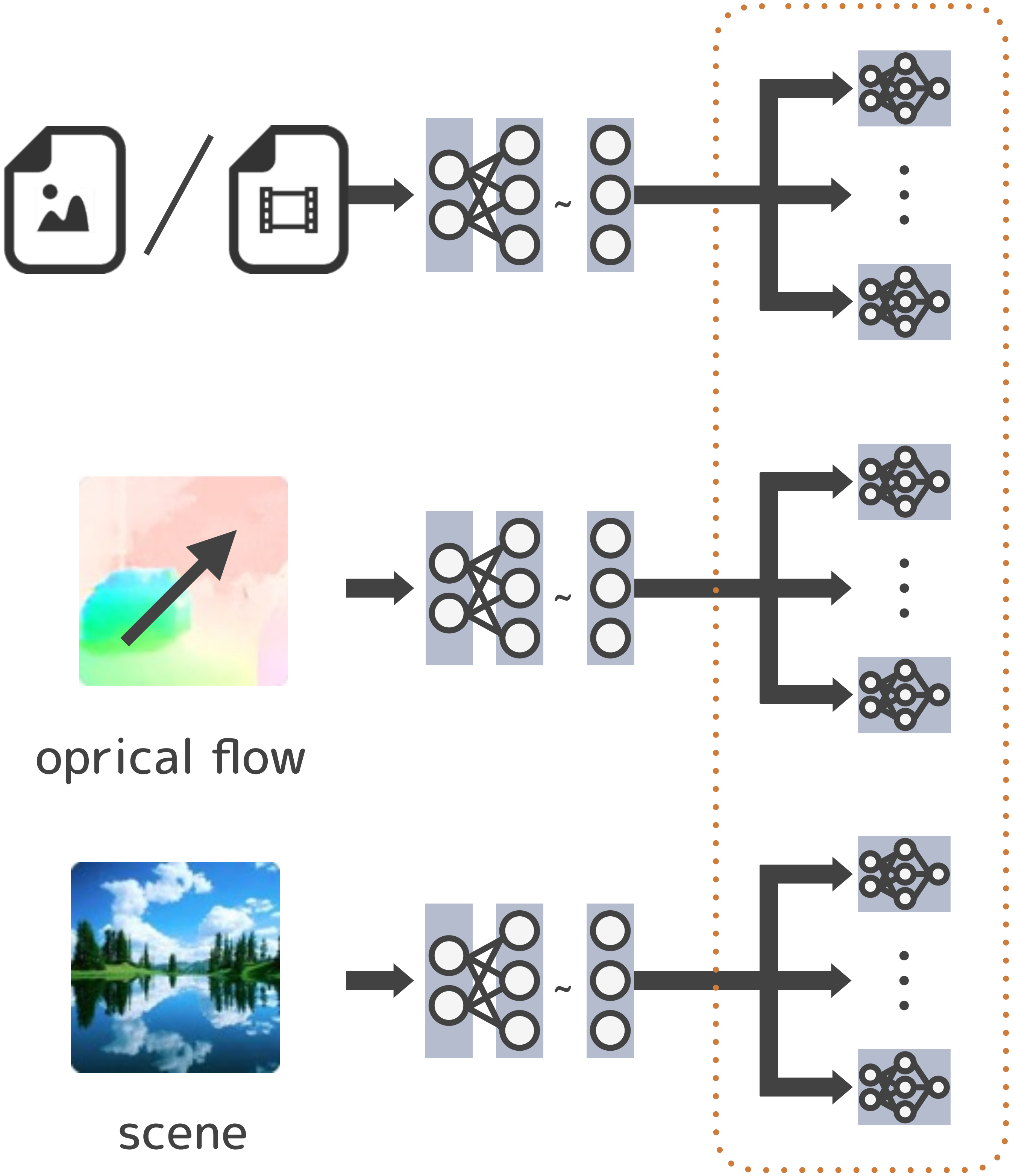
one video frame every 30 frames in a shot

more densly sampled video frames

# Future work

Acquiring temporal characteristics using optical flow.

Before detecting objects in a scene, we can first classify its environment to improve the performance.


oprical flow
scene