Zero-shot video retrieval using concept-based and visual-semantic embedding approaches

Kazuya Ueki[1,2], Ryo Mutou[1], Takayuki Hori[2,3], Yongbeom Kim[3], and Yuma Suzuki[3]

[1] Meisei University [2] Waseda University [3] SoftBank Corporation

Concept-based approach

Huge concept bank based on pretrained models

Name	Database	# Concepts	Concept Type(s)	Models
TRECVID346	TRECVID SIN	346	Person, Object, Scene, Action	GoogLeNet + SVM
FCVID239	FCVID	239	Person, Object, Scene, Action	GoogLeNet + SVM
UCF101	UCF101	101	Action	GoogLeNet + SVM
PLACES205	Places	205	Scene	AlexNet
PLACES365	Places	365	Scene	GoogLeNet
HYBRID1183	Places, ImageNet	1,183	Person, Object, Scene	AlexNet
IMAGENET1000	ImageNet	1,000	Person, Object	GoogLeNet
IMAGENET4000	ImageNet	4,000	Person, Object	GoogLeNet
IMAGENET4437	ImageNet	4,437	Person, Object	GoogLeNet
IMAGENET8201	ImageNet	8,201	Person, Object	GoogLeNet
IMAGENET12988	ImageNet	12,988	Person, Object	GoogLeNet
IMAGENET21841	ImageNet	21,841	Person, Object	GoogLeNet
ACTIVITYNET200	ActivityNet	200	Action	GoogLeNet + SVM
KINETICS400	Kinetics	400	Action	3D-ResNet
ATTRIBUTES300	Visual Genome	300	Attributes of persons/objects	GoogLeNet + SVM
RELATIONSHIPS53	Visual Genome	53	Relationships b/w persons/objects	GoogLeNet + SVM
FACES40	CelebA	40	Face Attributes	face detector + CNN

concepts > 50,000

Video retrieval pipeline

1. Extract one or more keywords from a query sentence.
2. Select one or more concept classifiers related to a keyword. The corresponding concept may not exist in the concept bank.
   → **Word2vec** to obtain more concepts
3. For each video, a score is calculated for the query sentence by integrating the scores from multiple concept classifiers.

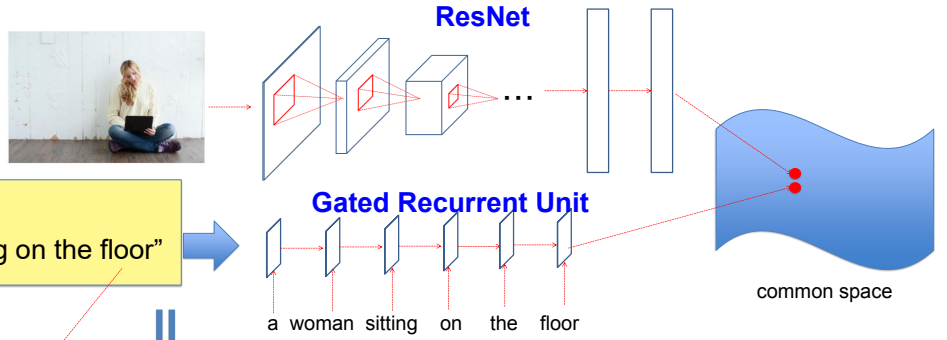
Query:
"a woman sitting on the floor"

Keywords:
"woman," "sitting," "floor"

score of "woman" + score of "sitting" + score of "floor"

Visual-semantic embedding approach

VSE++ "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives" [F. Faghri, et.al, 2017]



$$\ell_{MH}(i,c) = \max_{c'}\left[\alpha + s(i,c') - s(i,c)\right]_+ + \max_{i'}\left[\alpha + s(i',c) - s(i,c)\right]_+$$

The use of hard negatives in structured prediction, and ranking loss functions

Models were trained using four image caption datasets:

	# training
• Flickr8k	65k
• Flickr30k	295k
• MS COCO	424k
• Conceptual Captions	2809k

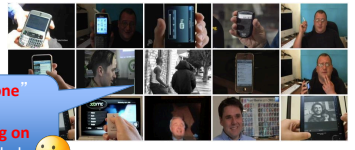
Example of retrieval results using past test data

ID: 534
"a person talking behind a podium wearing a suit outdoors during daytime"

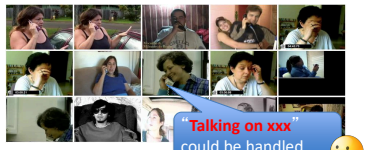
"Person," "podium," "suit," "outdoors," and "daytime" were captured and the average precision was high.

"Podium" and "outdoors" were not regarded as important, and many were not searched correctly.

ID: 553
"a person talking on a cell phone"

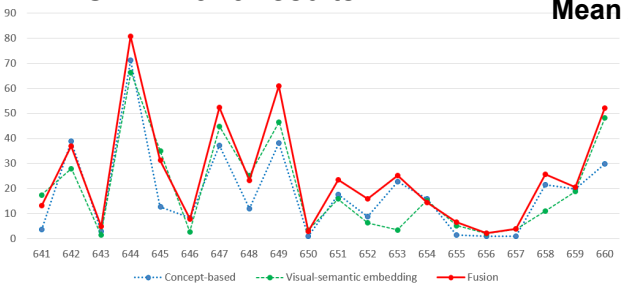


"Person" and "cell phone" could be captured but phrases such as "talking on xxx" could not be handled.



"Talking on xxx" could be handled

TRECVID 2020 results



Mean Average Precision

- 25.2 Fusion
- 20.0 Visual-semantic embedding
- 18.3 Concept-base

The video retrieval performance could be improved by integrating these two approaches because of their complementarity.