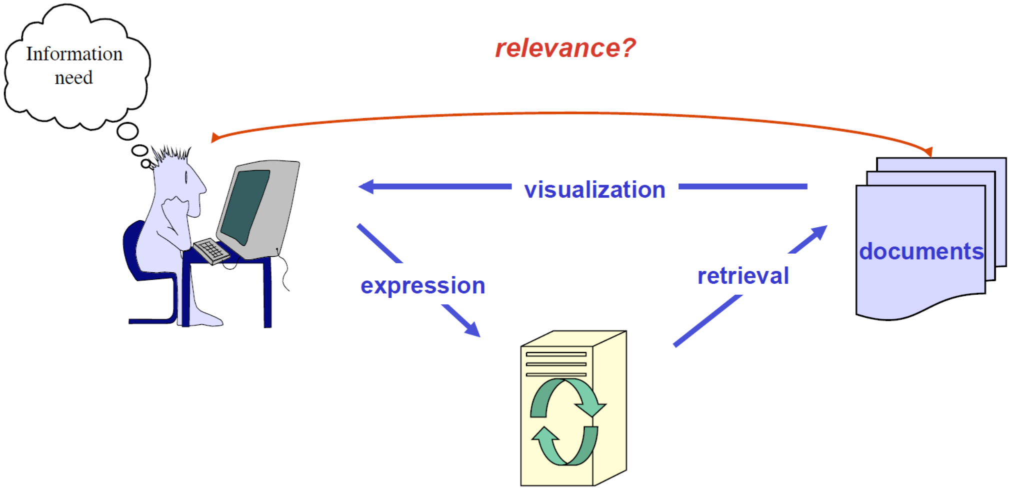
Advanced Machine Learning
Lecture 12 – Multimedia Retrieval

Georges Quénot

Univ. Grenoble Alpes, CNRS, Grenoble-INP, LIG, F-38000 Grenoble France

December, 2023

IR basics - Satisfaction of a user information need



IR basics - “Classical” IR and beyond

- Classical Multimedia IR
  - Text retrieval, image retrieval, video retrieval, music retrieval, audio retrieval . . .
  - A query, a collection of documents → a ranked list of results
  - Plus a “ground truth” (reference) and a metric → performance evaluation
- Information (stream) filtering
- Recommendation systems
- Personalized, mobile, in context search
- Question answering, multimedia question answering
- Relatively new: justification, explainability, transparency, fairness
  - Why this document? Why not this other one? Diversity, long tail blindness
  - European Commission’s GDPR: right to explanation
  - Avoid undesirable results, *e.g.*, “ImageNet roulette”

IR basics - Classical Multimedia IR

- Represent the query and documents in a vector representation (descriptors)
  - Color histograms (color distribution)
  - Gabor transforms (texture distribution)
  - Points of interest: SIFT, STIP, SURF ... (local representations)
  - Bags of Visual Words (clustering and histogramming of points of interest)
  - Fisher Vectors, VLADs, VLATs ...
  - Block, pyramidal decompositions
  - ...
  - CNN features
- Metric between representation vectors: Euclidean distance, cosine similarity ...
- Plus: metric learning
- Note: usually same representations for retrieval and for classification

Progress in multimedia IR through a few major papers

- CNN Features off-the-shelf: an Astounding Baseline for Recognition (Razavian et al., 2014)
- Deep Image Retrieval: Learning Global Representations for Image Search (Gordo et al., 2016)
- Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models (Kiros et al., 2014)
- VSE++: Improving Visual-Semantic Embeddings with Hard Negatives (Faghri et al., 2018)
- Dual Encoding for Zero-Example Video Retrieval (Dong et al., 2019)
- Waseda Meisei SoftBank at TRECVID 2020: Ad-hoc Video Search (Ueki et al., 2020)
- Interpretable Embedding for Ad-Hoc Video Search (Wu and Ngo, 2020)

Use of "off-the-shelf" CNN Features (Razavian et al., 2014)

- "CNN Features off-the-shelf: an Astounding Baseline for Recognition"
- Mostly about classification but one section about object (instance) retrieval
- Use of the publicly available trained CNN called OverFeat (variant of AlexNet)
- Use of the L_2 normalized output of the first fully connected layer as representation
- Variants with spatial search and data augmentation with dimensionality reduction
- Comparison with 5 state-of-the-art descriptors: VLAD (Vector of Locally Aggregated Descriptors), BoW (Bag of Visual Words), IFV (Fisher Vectors), Hamming Embedding, and BoB (Bag of Boundaries).
- Results on 5 image retrieval test collections: Oxford5k buildings, Paris6k buildings, Sculptures6k, Holidays dataset, and UKbench

Use of “off-the-shelf” CNN Features (Razavian et al., 2014)

	Dim	Oxford5k	Paris6k	Sculp6k	Holidays	UKBench
BoB[3]	N/A	N/A	N/A	**45.4**[3]	N/A	N/A
BoW	200k	36.4[20]	46.0[35]	8.1[3]	54.0[4]	70.3[20]
IFV[33]	2k	41.8[20]	-	-	62.6[20]	83.8[20]
VLAD[4]	32k	55.5 [4]	-	-	64.6[4]	-
CVLAD[52]	64k	47.8[52]	-	-	81.9[52]	89.3[52]
HE+burst[17]	64k	64.5[42]	-	-	78.0[42]	-
AHE+burst[17]	64k	66.6[42]	-	-	79.4[42]	-
Fine vocab[26]	64k	74.2[26]	74.9[26]	-	74.9[26]	-
ASMK*+MA[42]	64k	80.4[42]	77.0[42]	-	81.0[42]	-
ASMK+MA[42]	64k	**81.7**[42]	78.2[42]	-	82.2[42]	-
CNN	4k	32.2	49.5	24.1	64.2	76.0
CNN-ss	32-120k	55.6	69.7	31.1	76.9	86.9
CNNaug-ss	4-15k	**68.0**	**79.5**	42.3	**84.3**	**91.1**
CNN+BOW[16]	2k	-	-	-	**80.2**	-

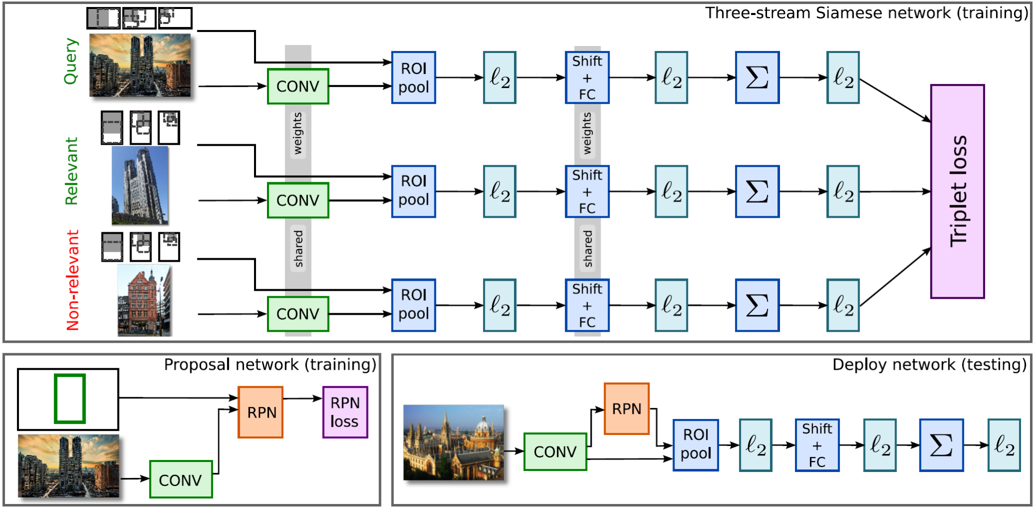
Use of "off-the-shelf" CNN Features (Razavian et al., 2014)

- Less clear results than for classification tasks
- Best only with spatial search and with data augmentation with dimensionality reduction
- Not always the best but representation size smaller than with other approaches
- Very good baseline anyway for automatically built (learned) descriptors used "as is"
- Classical "handcrafted" global descriptors are obsolete for visual IR tasks
- For doing better:
  - Use a more powerful "backbone", *e.g.*, ResNet
  - Do fine tuning on the backbone weights
  - Do metric learning

Metric learning: Siamese networks (Gordo et al., 2016)

- *Instance-level* retrieval (monuments in the evaluation)
- Use of a pre-trained CNN (VGG16) backbone for "raw" feature (descriptor) extraction
- Add a fully connected layer (and some normalization steps) for mapping the raw descriptor to a final one
- Use Euclidean distance for estimating the closeness of a candidate image to the query
- Use a three-stream Siamese network architecture with a triplet loss function
- Use of a region proposal network
- Fine tuning
- Hard negative mining: more effective if negative samples are chosen as close to the query as possible
- Need for a training set for the metric learning

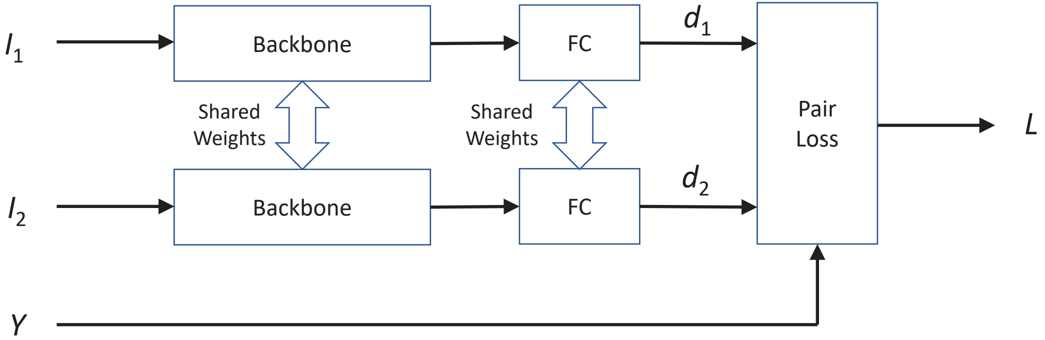
Metric learning: Siamese networks (Gordo et al., 2016)



Metric learning: Siamese networks, pair and triplet losses

- Contrastive (pair) loss:
  $L(Y, I_1, I_2) = (1 - Y) \|d_1 - d_2\|^2 - (Y) \|d_1 - d_2\|^2$
  with $Y = 0$ if I_1 and I_2 are similar and $Y = 1$ if I_1 and I_2 are dissimilar
- Contrastive (pair) loss with margin:
  $L(Y, I_1, I_2) = (1 - Y) \|d_1 - d_2\|^2 + (Y) \max(0, m - \|d_1 - d_2\|)^2$
  We don't care about putting the dissimilar as far away as possible, just far enough, no effect if I_1 and I_2 are dissimilar and $\|d_1 - d_2\|$ is greater than m
- Triplet loss with margin:
  $L(I_q, I^+, I^-) = \max(0, m + \|q - d^+\|^2 - \|q - d^-\|^2)$
  We don't care about putting the negatives as far away from the query as possible, just farther than the positive enough, no effect if $\|q - d^-\|^2$ is greater than $m + \|q - d^+\|^2$
- Significant improvement over the previous state of the art (see paper)

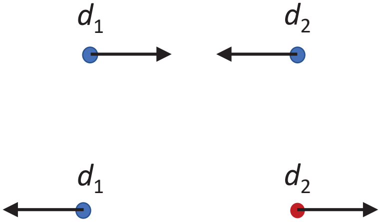
Metric learning: Two-Stream Siamese Network



Metric learning: Contrastive Pair Loss

Similar: Y = 0

Dissimilar: Y = 1



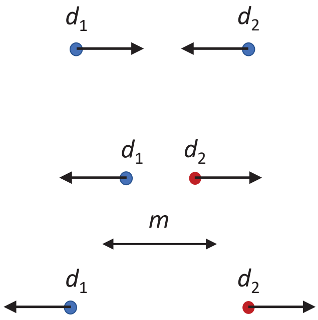
$L = +\|d_1 - d_2\|^2$

$L = -\|d_1 - d_2\|^2$

$$L(Y, I_1, I_2) = (1 - Y)\,\|d_1 - d_2\|^2 - (Y)\,\|d_1 - d_2\|^2$$

with $Y = 0$ if I_1 and I_2 are similar and $Y = 1$ if I_1 and I_2 are dissimilar

Metric learning: Contrastive Pair Loss with Margin



Similar: Y = 0 $\quad L = +\|d_1 - d_2\|^2$

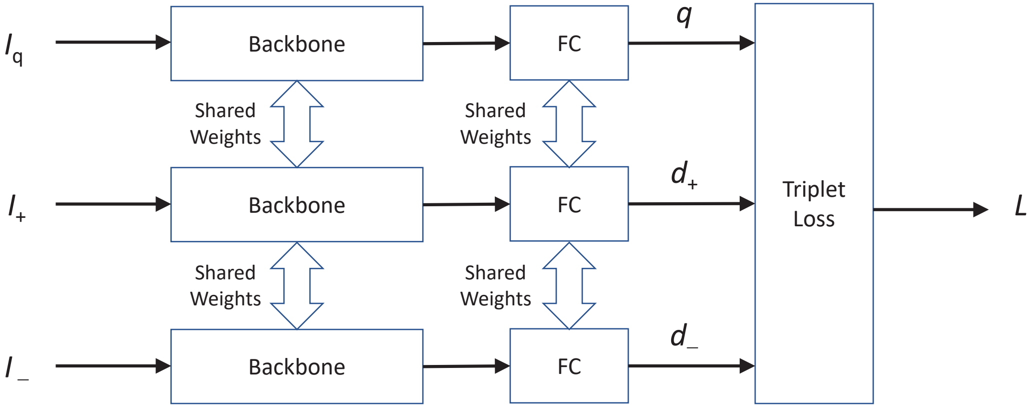
Dissimilar: Y = 1 $\quad L = \max(0, m - \|d_1 - d_2\|^2)$

Dissimilar: Y = 1 $\quad L = 0$

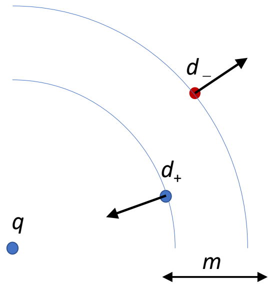
$$L(Y, I_1, I_2) = (1 - Y)\,\|d_1 - d_2\|^2 + (Y)\max(0, m - \|d_1 - d_2\|)^2$$

We don't care about putting the dissimilar as far away as possible, just far enough,
no effect if I_1 and I_2 are dissimilar and $\|d_1 - d_2\|$ is greater than m

Metric learning: Three-Stream Siamese Network



Metric learning: Triplet Loss with Margin



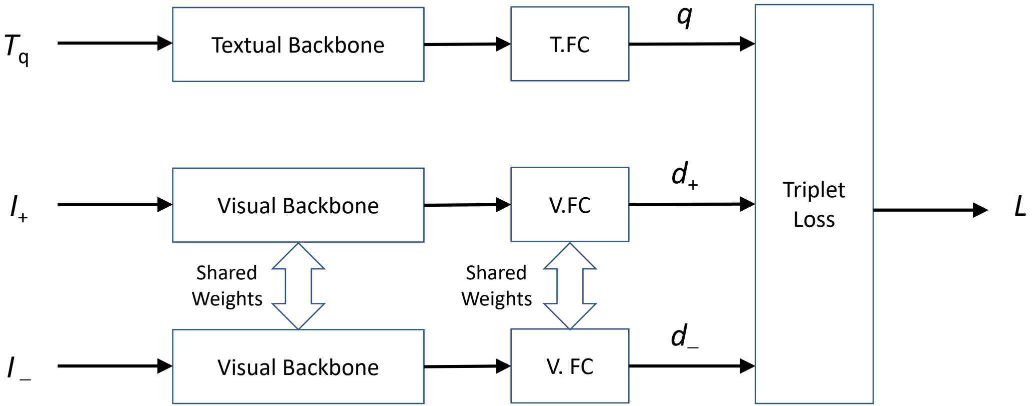
$$L(I_q, I^+, I^-) = \max(0, m + \|q - d^+\|^2 - \|q - d^-\|^2)$$

We don't care about putting the negatives as far away from the query as possible, just farther than the positive enough, no effect if $\|q - d^-\|^2$ is greater than $m + \|q - d^+\|^2$

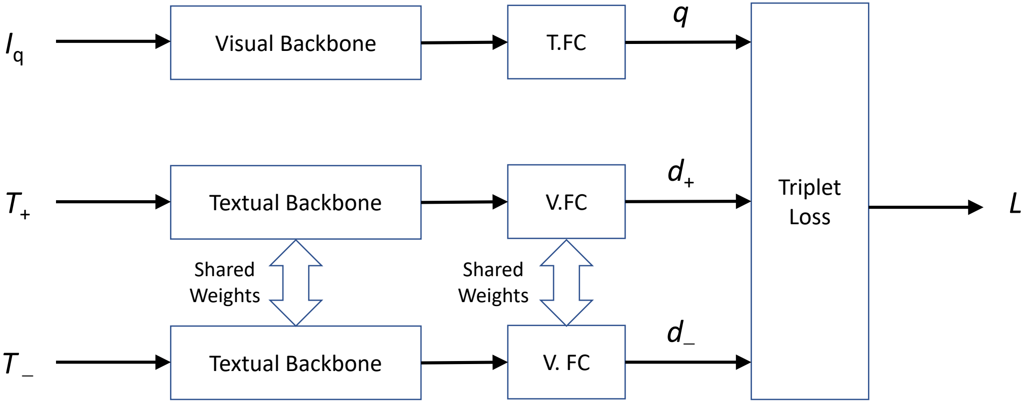
Visual-Semantic embedding (Kiros et al., 2014) (Faghri et al., 2018)

- Similar to metric learning with Siamese networks (though older) with a main difference that the query and the documents are from different modalities: text to image or image to text retrieval tasks
- Training with a large collection of (image, caption) pairs
- Modality-specific backbones:
  - CNN for image streams and LSTM for text streams
- Modality-specific transformation matrices

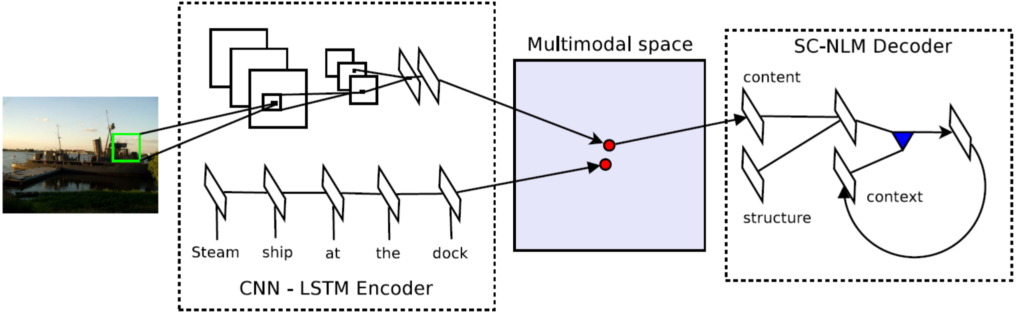
Metric learning: Cross-Modal Three-Stream Siamese Network



Metric learning: Cross-Modal Three-Stream Siamese Network

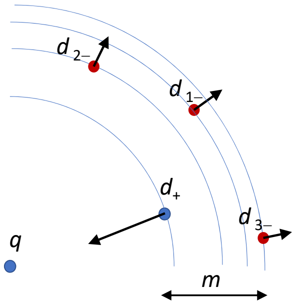


Visual-Semantic embedding (Kiros et al., 2014) (Faghri et al., 2018)



The VSE network architecture. The right part is not relevant for the retrieval tasks.

Metric learning: Triplet Loss with Margin on a batch, average

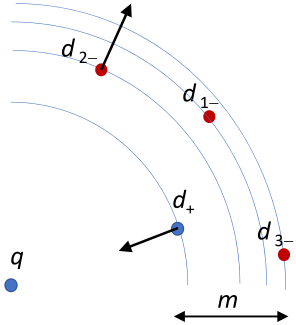


One triplet: $L = \max(0, m + \|q - d^+\|^2 - \|q - d_k^-\|^2)$

$N - 1$ triplets with the same query: $L = \Sigma_k \max(0, m + \|q - d^+\|^2 - \|q - d_k^-\|^2)$

$N(N-1)$ triplets with the multiple queries: $L = \Sigma_q \Sigma_k \max(0, m + \|q - d^+\|^2 - \|q - d_k^-\|^2)$

Metric learning: Triplet Loss with Margin on a batch, maximum



One triplet: $L = \max(0, m + \|q - d^+\|^2 - \|q - d_k^-\|^2)$

$N - 1$ triplets with the same query: $L = \max_k \max(0, m + \|q - d^+\|^2 - \|q - d_k^-\|^2)$

$N(N-1)$ triplets with the multiple queries: $L = \Sigma_q \max_k \max(0, m + \|q - d^+\|^2 - \|q - d_k^-\|^2)$

Visual-Semantic embedding (Kiros et al., 2014) (Faghri et al., 2018)

- Sum of triplet loss mith margin over modalities, samples per modality, and set of negative samples for each positive sample:

$$L = \sum_x \sum_k \max(0, \alpha - \|s(x,v)\|^2 + \|s(x,v_k)\|^2) + \sum_v \sum_k \max(0, \alpha - \|s(x,v)\|^2 + \|s(x_k,v)\|^2)$$

where v_k is a contrastive (non-associated) sentence for image x, and vice-versa with x_k, in practice, the v_k and x_k are taken only in the same batch

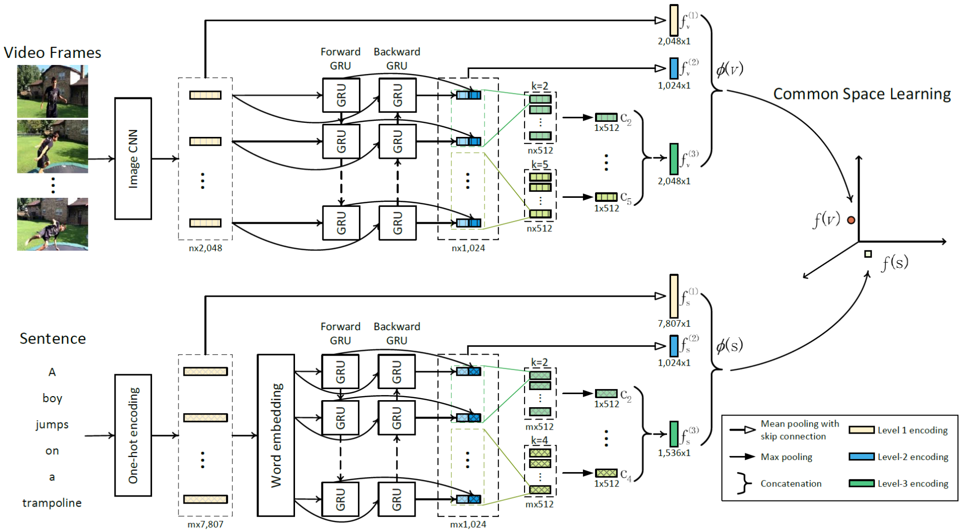
- In VSE++, focus on hard negatives by replacing the second sum by the max operator, keeping only the one negative which is closest (within the batch) to the query while the original VSE averaged them:

$$L = \sum_x \max_k \max(0, \alpha - \|s(x,v)\|^2 + \|s(x,v_k)\|^2) + \sum_v \max_k \max(0, \alpha - \|s(x,v)\|^2 + \|s(x_k,v)\|^2)$$

Dual Encoding for Zero-Example Video Retrieval (Dong et al., 2019)

- Similar to VSE++ but with video instead of images and more elaborated encoding
- Challenges:
  - Video retrieval, TRECVid Ad'hoc Video Search (AVS) task
  - Text to video and video to Text, TRECVid VTT and MSR VTT tasks
- Word embedding instead of LSTM
- Two additional levels for taking into account the sequence aspects:
  - Bi-directional GRU (sequence to sequence)
  - 1D CNNs on Bi-GRU output vector sequences
  - Very similar for both streams
- Concatenation and classical common space learning then

Dual Encoding for Zero-Example Video Retrieval (Dong et al., 2019)



Dual Encoding for Zero-Example Video Retrieval (Dong et al., 2019)

Table 2. **Ablation study on MSR-VTT**. The overall performance, as indicated by **Sum of Recalls**, goes up as more encoding layers are added. Dual encoding exploiting all the three levels is the best.

Encoding strategy	Text-to-Video Retrieval					Video-to-Text Retrieval					Sum of Recalls
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP	
Level 1 (Mean pooling)	6.4	18.8	27.3	47	0.132	11.5	27.7	38.2	22	0.054	129.9
Level 2 (biGRU)	6.3	19.4	28.5	38	0.136	10.1	26.8	37.7	20	0.057	128.8
Level 3 (biGRU-CNN)	7.3	21.5	31.2	32	0.150	10.6	27.3	38.5	20	0.061	136.4
Level 1 + 2	6.9	20.4	29.1	41	0.142	11.6	29.6	40.7	18	0.058	138.3
Level 1 + 3	7.5	21.6	31.2	33	0.151	11.9	30.5	41.7	16	0.062	144.4
Level 2 + 3	7.6	**22.4**	**32.2**	**31**	**0.155**	11.9	**30.9**	42.7	16	**0.066**	147.7
Level 1 + 2 + 3	**7.7**	22.0	31.8	32	**0.155**	**13.0**	30.8	**43.3**	**15**	0.065	**148.6**

Concept-based retrieval approaches (Ueki et al., 2020)

- Gather as many pre-trained visual classifiers "concept banks" as possible : concepts, places, faces, activities . . .
- Build and merge them and apply them to video key frames or to video shots
- Get a vector of the probability of presence for each concept in a visual unit (video shot)
- Identify which concepts are present in or associated to the query text using NLP techniques and make a probability vector and or a Boolean expression from it
- Score the similarity between the query and video shots representations using a vector space model and/or a Boolean expression
- Rank results accordingly
- Actually, quite old approach, at least since 2009
- Concept-based approaches are usually less good than concept-free ones
- Support for explainability

Concept banks for Video Retrieval (Ueki et al., 2020)

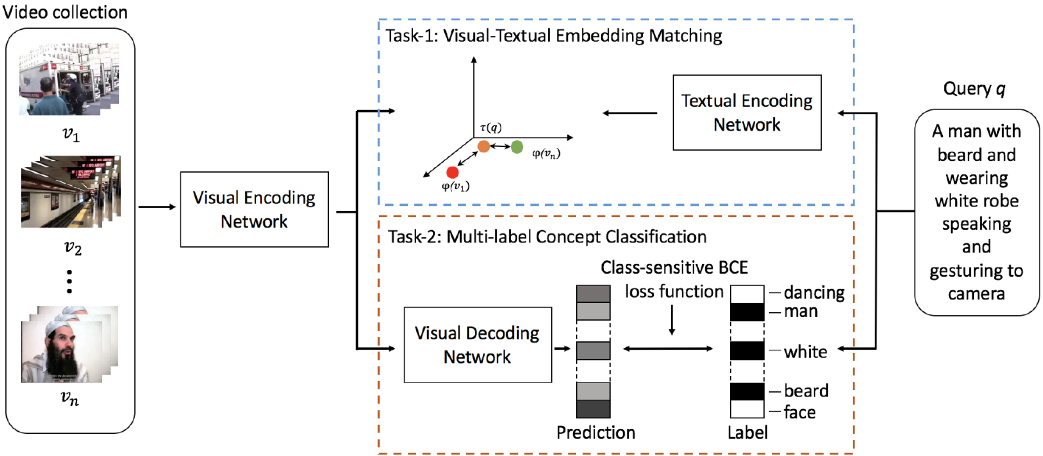
Table 1. Concept bank used in our systems.

Name	Database	# Concepts	Concept Type(s)	Models
TRECVID346	TRECVID SIN [2]	346	Person, Object, Scene, Action	GoogLeNet + SVM
FCVID239	FCVID [3]	239	Person, Object, Scene, Action	GoogLeNet + SVM
UCF101	UCF101 [4]	101	Action	GoogLeNet + SVM
PLACES205	Places [5]	205	Scene	AlexNet
PLACES365	Places	365	Scene	GoogLeNet
HYBRID1183	Places, ImageNet [6]	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 [7]	200	Action	GoogLeNet + SVM
KINETICS400	Kinetics [8]	400	Action	3D-ResNet
ATTRIBUTES300	Visual Genome [9]	300	Attributes of persons/objects	GoogLeNet + SVM
RELATIONSHIPS53	Visual Genome	53	Relationships b/w persons/objects	GoogLeNet + SVM
FACES40	CelebA [10]	40	Face Attributes	face detector + CNN

Interpretable Embedding for Ad-Hoc Video Search (Wu and Ngo, 2020)

- Two tasks:
  - Visual-Textual Embedding Matching, similar to dual encoding an VSE++
  - Multi-label Concept Classification, with collection-specific concepts
- Combination of concept-based and concept-free approaches
- Significantly better performance for the hybrid method
- Some elements of explainability from the concept-based side

Interpretable Embedding for Ad-Hoc Video Search (Wu and Ngo, 2020)



Interpretable Embedding for Ad-Hoc Video Search (Wu and Ngo, 2020)

Datasets	IACC.3			V3C1
Query sets	tv16	tv17	tv18	tv19
TRECVid top results:				
Rank 1	0.054 [36]	0.206* [41]	0.121 [21]	0.163 [49]
Rank 2	0.051 [30]	0.159 [44]	0.087* [18]	0.160 [22]
Rank 3	0.040 [24]	0.120* [33]	0.082 [6]	0.123 [45]
Embedding only:				
VideoStory [17]	0.087	0.150	/	/
VSE++ [15]	0.123	0.154	0.074	/
W2VV [13]	0.050	0.081	0.013	/
W2VV++ [21]	0.163	0.196	0.115	0.127
Dual coding [14]	0.165	0.228	0.117	0.152
Concept only:				
QKR [29]	0.064	/	/	/
ConBank (auto)	/	0.159 [44]	0.060 [47]	/
ConBank (manual)	0.177 [46]	0.216 [44]	0.106 [47]	0.114 [45]
Dual-task:				
$DT_{concept}$	0.148	0.147	0.091	0.115
$DT_{embedding}$	0.163	0.232	0.118	0.168
$DT_{combined}$	**0.185***	**0.241***	**0.123***	**0.185***

Conclusion

- Concept-based, concept-free, and hybrid approaches
- Represent both queries and documents using neural networks, either when of the same modality or not
- LSTM, word embeddings, transformers (BERT) for texts
- 2D CNNs for still images and 3D CNNs for videos
- Multiple levels
- Projection in a common embedding space using Siamese networks with triplet losses
- Fine tuning of pre-trained networks
- (Moderate) explainability using hybrid approaches
- Multimedia Transformers (MMT)

References

Jianfeng Dong, Xirong Li, Chaoxi Xu, Shouling Ji, Yuan He, Gang Yang, and Xun Wang. Dual encoding for zero-example video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pages 9346–9355, 2019.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2018.

Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *Computer Vision – ECCV 2016*, pages 241–257, 2016.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models, 2014.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. CNN Features off-the-shelf: an Astounding Baseline for Recognition, 2014.

Kazuya Ueki, Ryo Mutou1, Takayuki Hori, Yongbeom Kim, and Yuma Suzuki. Waseda meisei softbank at trecvid 2020: Ad-hoc video search. In *Proceedings of TRECVid*, 2020.

Jiaxin Wu and Chong-Wah Ngo. Interpretable embedding for ad-hoc video search. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 3357–3366. ACM, 2020.