Waseda_Meisei_SoftBank at TRECVID 2020: Ad-hoc Video Search

Kazuya Ueki[1,2], Ryo Mutou[1], Takayuki Hori[3,4], Yongbeom Kim[3], and Yuma Suzuki[3]

[1] Department of Information Science, Meisei University,
Room 27-1809, Hodokubo 2-1-1, Hino, Tokyo, 191–8506, Japan
[2] Faculty of Science and Engineering, Waseda University,
Room 40-701, Waseda-machi 27, Shinjuku-ku, Tokyo, 162–0042, Japan
[3] AI Solution Division, 5G & IoT Solution Division, SoftBank Corporation,
1-9-1, Higashi-shimbashi, Minato-ku, Tokyo, 105–0021, Japan
[4] Global Information and Telecommunication Institute, Waseda University,
Room 55-208, Okubo 3-4-1, Shinjuku-ku, Tokyo, 162–0042, Japan
`kazuya.ueki@meisei-u.ac.jp`

Abstract. The Waseda_Meisei_SoftBank team participated in the TRECVID 2020 Ad-hoc Video Search (AVS) task [1]. As with last year's AVS task, we submitted both manually assisted and fully automatic runs this year. Our approach consisted of concept-based video retrieval for manually assisted runs and visual-semantic embedding for fully automatic runs. Our best manually assisted run achieved a mean average precision (mAP) of 25.2%, which ranked the highest among all the manually assisted systems. Our fully automatic run achieved an mAP of 20.0%, which ranked fifth among all participants.

1 System Description

We used two approaches (concept-based approach and visual-semantic embedding approach) for video retrieval from large-scale video data using query sentences. This section introduces how both systems were created.

1.1 Concept-based Approach

For a concept-based approach, we first built a large concept bank comprising of several concept types as shown in Table 1. It contained classifiers such as persons, objects, scenes, and actions to deal with various forms of query sentences. Using this concept bank, all concept scores for all videos were calculated. Here, we explain how to create concept classifiers for each database and pre-trained models.

1. `TRECVID346`, `FCVID239`, `UCF101`, and `ACTIVITYNET200`

   First, a maximum of ten frames from each shot were selected at regular intervals, and the corresponding images were input to the *GoogLeNet* model [11] pre-trained on the ImageNet database [6]. This allowed us to obtain 1,024-dimensional feature vectors from pool5 layers. These feature vectors (maximum ten) were then bound to a single vector using element-wise max-pooling. We trained SVMs using data-given labels of concepts for each database as positive samples and randomly selected images from the TRECVID SIN dataset as negative samples. The shot score for each concept was calculated as the distance to the hyperplane in the SVM model.

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

2. PLACES205/365, HYBRID1183, and IMAGENET1000/4000/4437/8201/12988/21841

   We calculated the concept scores using a scene classification model pre-trained with the Places database [5] or image classification models [12][13][14] pre-trained with the ImageNet database [6]. Since each unit of the convolution neural network (CNN) output layer identifies a concept in a scene/object, the values representing a concept of the CNN output layer (before softmax was applied) were used as concept scores. The maximum concept score for each video was obtained after inputting at most ten images to the CNN.

3. KINETICS400

   We used a *3D-ResNet* model [16][17] pre-trained with the Kinetics database [8]. Sixteen consecutive frames were input into 3D-ResNet and the maximum score of each concept obtained from the output layer was taken as the concept score for each video.

4. ATTRIBUTES300 and RELATIONSHIPS53

   Using annotations of attributes of persons/objects and relationships between persons/objects in the Visual Genome Database [9], 300 types of attributes and 53 types of relationship concepts were created. The attributes used "adjective + noun" concepts such as "blue_sky," "white_plate." As for the relationships, we selected the data whose subject was a person and which had specific verbs such as "wear" (34 types) and "hold/have" (19 types). Finally, we created concepts such as "wear_shirt", "have_ski_pole".

5. FACES40

   First, we trained the 40 face attributes using CelebA dataset [10]. For testing, face regions were cropped using dlib's face detector[5]. Then, the attribute scores were calculated for at most ten images per video, and the maximum score among all images in each video was used as the video's attribute score.

[5] https://github.com/davisking/dlib.git

After calculating the concept scores[6] for every video sequence in advance, we retrieved video using word-based keyword selection through the following 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.
3. For each video, a score is calculated for the query sentence by integrating the scores from multiple concept classifiers.

Given a query sentence, we manually selected some visually important keywords. For example, given the query sentence "a woman sitting on the floor," we picked out the keywords "woman," "sitting," and "floor." We then matched the keywords with concepts using a concept classifier. Semantically similar concepts were also chosen using the word2vec algorithm [18] to select as many concept classifiers as possible.

1.2 Visual-semantic Embedding Approach

In recent years, visual-semantic embedding methods, which map visual and semantic features onto a common space, have been actively researched [19][20]. Visual-semantic embedding approaches were also seen in the TRECVID benchmark [21][22][23], and they achieved relatively high mAPs.

We used the implementation[7] of VSE++ [24] for training. For training the visual-semantic embedding, four image caption datasets, Flickr8k [25], Flickr30k [26], MS COCO [27], and Conceptual Captions [28], were used. The total number of data was 3,559,009, including 65,000 from Flickr8k, 295,070 from Flickr30k, 423,915 from MS COCO, and 2,809,024 from Conceptual Captions[8]. We used gated recurrent unit for feature extraction from query sentences and the *ResNet-50*, *ResNet-101*, and *ResNet-152* models for feature extraction from images. Due to the large amount of training data, 500,000 training data and 50,000 validation data to train visual-semantic embedding models were randomly selected. We repeated this data selection process 32 times for each of the three types of ResNet model, and trained 96 embedding models. Finally, test data were ranked by an average of the scores obtained from the 96 models[9].

2 Submissions

This year we submitted four manually assisted runs (Manual1, Manual2, Manual3, and Manual4) and four fully automatic run (Automatic1, Automatic2, Automatic3, and Automatic4) to the TRECVID 2020 Ad-hoc Video Search (AVS) task as shown in Table 2. For manually assisted runs, Manual4 adapted the concept-based approach alone. Other runs were the fusion of the concept-based and visual-semantic embedding approaches, because we consider the concept-based and visual-semantic embedding approaches to be complementary. These two approaches were combined to re-rank the video retrieval result using reciprocal rank fusion (RRF) [30],

$$RRF_{score} = \sum_{r \in R} \frac{1}{k + r}, \tag{1}$$

[6] The score for each semantic concept was normalized for all test shots iterations using a min-max normalization, that is, the maximum and minimum scores were 1.0 (most probable) and 0.0 (least probable), respectively.

[7] https://github.com/fartashf/vsepp

[8] The total number of data in Conceptual Captions dataset was 3,334,173 including 3,318,333 training data and 15,840 validation data; however, only downloadable data were used.

[9] The score for each model was normalized over all test shots using min-max normalization, that is, the maximum and minimum scores were 1.0 and 0.0, respectively.

Table 2. Our submitted runs for TRECVID 2020.

Run name	Concept-based	Visual-semantic embedding	Additional VSE++ models for Visual-semantic embedding	mAP
Manual1	✓	✓	TREC-VTT-VSE++	24.5
Manual2	✓	✓	TREC-VTT-VSE++, MSR-VTT-VSE++	24.1
Manual3	✓	✓		**25.2**
Manual4	✓			18.3
Automatic1		✓	TREC-VTT-VSE++	19.6
Automatic2		✓	TREC-VTT-VSE++, MSR-VTT-VSE++	19.8
Automatic3		✓		**20.0**
Automatic4		✓	MSR-VTT-VSE++	19.2

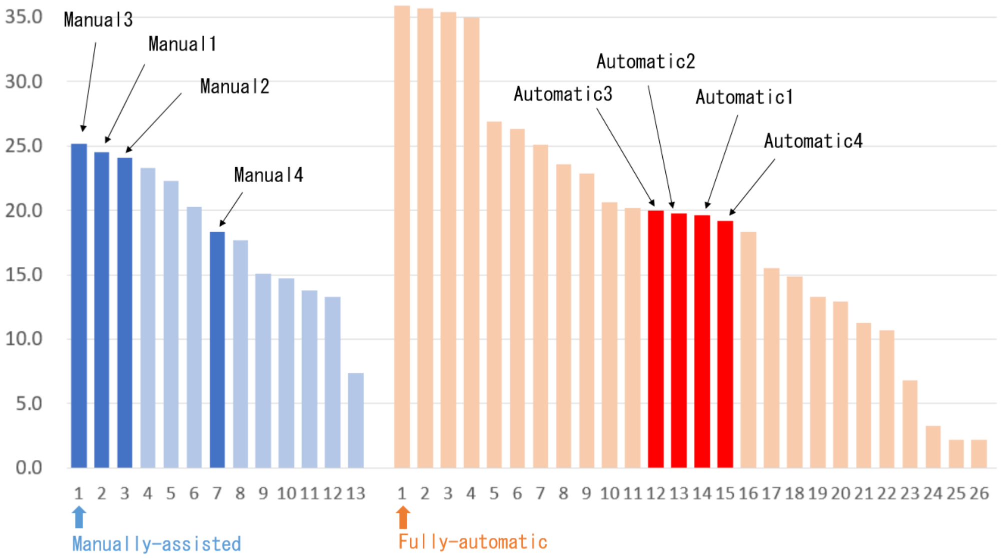

Fig. 1. Comparison of Waseda_Meisei_SoftBank runs with the runs of other teams for all the submitted runs including manually assisted (blue).

where R is the set of ranking, and k is a fixed parameter.

This year we trained other VSE++ models using the additional video captioning dataset (Twitter Vines + Flicker Videos for TRECVID VTT task, and MSR-VTT [29]). Because the number of videos of TRECVID VTT task was not large enough, we trained the VSE++ models using both Flickr30k and TRECVID VTT task dataset. The number of data selected for a VSE++ model was 160,000 for training and 22,667 for validation. In this paper we refer to this model as TREC-VTT-VSE++. We also refer the VSE++ model trained using MSR-VTT as MSR-VTT-VSE++. The number of data selected for training MSE++ models was 154,980 for training and 14,980 for validation.

3 Results

Figure 1 shows the results for all the submitted runs including manually assisted and fully automatic. The mAPs of our manually assisted runs (Manual1, Manual2, Manual3 and Manual4) were 24.5%, 24.1%, 25.2% and 18.3%, respectively, and the best run (Manual3) ranked 1st among all manually assisted systems. However, the best automatic system among all participants achieved a higher mAP of 35.9%, more than that

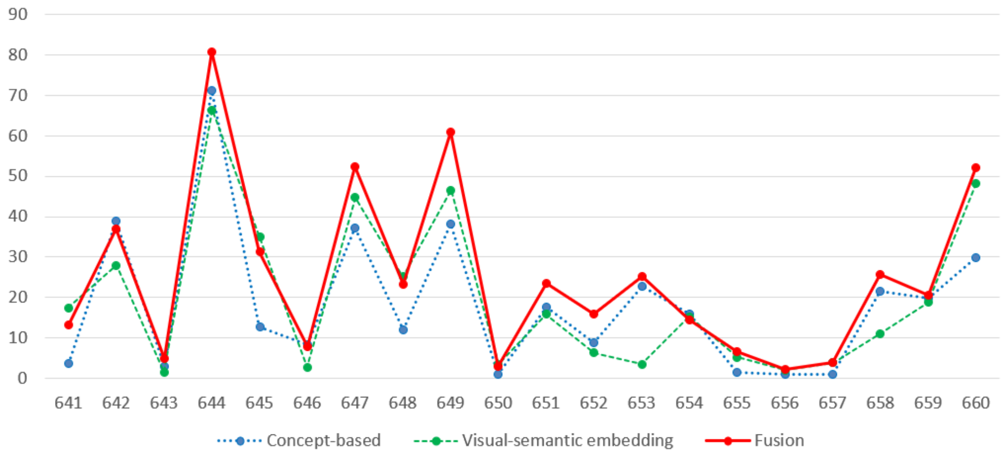

Fig. 2. The average precision for each query sentence. Blue: Concept-based approach alone (Manual4). Green: Visual-semantic embedding approach (Automatic3). Red: The combination of concept-based and visual-semantic embedding approaches (Manual3).

obtained by our manual system. The mAP of our fully automatic run (Automatic3) was 20.0%.

Figure 2 shows the average precision for each query sentence. Average precision for the fusion of concept-based approach combined with the visual-semantic embedding approaches was significantly better than each of the approaches alone. It was confirmed that the video retrieval performance could be improved by integrating these two approaches because of their complementarity.

4 Conclusion

For this year's submissions, we solved the problem of ad-hoc video search using a combination of the concept-based approach and visual-semantic embedding approaches. As these two approaches were complementary, we could improve the video retrieval performance by integrating them.

For future works, we will analyze the advantages and disadvantages of each approach and develop a new method to automatically determine the best approach to be used depending on the query sentence. We will also introduce recently proposed visual-semantic embedding approaches [31][32][33] to improve the video search performance.

Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 17H01831 and 18K11362.

References

1. G. Awad, A. A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Zhang, E. Godard, L. Diduch, J. Liu, A. F. Smeaton, Y. Graham, G. J. F. Jones, W. Kraaij, G. Quénot, "TRECVID 2020: Comprehensive campaign for evaluating video retrieval tasks across multiple application domains," In Proc. of TRECVID 2020, 2020.

2. G. Awad, C. G. M. Snoek, A. F. Smeaton, and G. Quénot, “TRECVid Semantic Indexing of Video: A 6-Year Retrospective,” ITE Trans. on MTA vol.4, no.3, pp.187–208, 2016.
3. Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, S.-F. Chang, “Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks,” arXiv:1502.07209, 2015.
4. K. Soomro, A. R. Zamir, M. Shah, “UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild,” arXiv:1212.0402, 2012.
5. B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” Advances in Neural Information Processing Systems (NIPS), 2014.
6. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” In Proc. of IEEE Computer Vision and Pattern Recognition (CVPR), 2009.
7. F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles “ActivityNet: A large-scale video benchmark for human activity understanding,” In Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015), pp.961–970, 2015.
8. W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T.r Back, P. Natsev, M. Suleyman, and A. Zisserman, “The Kinetics Human Action Video Dataset,” arXiv: 1705.06950, 2017.
9. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalanditis, L.-J. Li, D.A. Shamma, M.S. Bernstein, L. Fei-Fei, Y. Kalantidis, L.-J. Li, D.A. Shamma, M.S. Bernstein, and F.-F. Li, “Visual Genome : Connecting language and vision using crowdsourced dense image annotations,” arXiv:1602.07332, 2016.
10. Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” In Proc. of International Conference on Computer Vision (ICCV), 2015.
11. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. “Going deeper with convolutions,” In Proc. of Computer Vision and Pattern Recognition (CVPR), 2015.
12. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, “Caffe: Convolutional architecture for fast feature embedding”, arXiv:1408.5093, 2014.
13. P. Mettes, D. C. Koelma, and C. G. Snoek, “The ImageNet Shuffle: Reorganized Pretraining for Video Event Detection,” In Proc. of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR), pp.175–182, 2016.
14. S. loffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” arXiv:1502.03167, 2015.
15. J. Redmon and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” arXiv:1612.08242, 2016.
16. K. Hara, H. Kataoka, and YutakaSatoh, “Can Spatiotemporal3DCNNs Retrace the History of 2DCNNs and ImageNet?,” In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2018), pp. 6546-6555, 2018.
17. K. Hara, H. Kataoka, and Y. Satoh, “Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition,” In Proc. of the ICCV Workshop on Action, Gesture, and Emotion Recognition, 2017.
18. T. Mikolov, K. Chen, G. Corrado, and J. Dean. “Efficient estimation of word representations in vector space,” arXiv:1301.3781, 2013.
19. A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, “DeViSE: A Deep Visual-Semantic Embedding Model,” Advances in Neural Information Processing Systems 26, pp.2121–2129, 2013.
20. R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models,” arXiv:1411.2539, 2014.
21. C. G. M. Snoek, X. Li, C. Xu, and D. C. Koelma, “University of Amsterdam and Renmin University at TRECVID 2017: Searching Video, Detecting Events and Describing Video,” Proceedings of TRECVID 2017, 2017.
22. P. A. Nguyen, Q. Li, Z. Cheng, Y. Lu, H. Zhang, and C. Ngo, “VIREO@TRECVID 2017: Video-to-Text, Ad-hoc Video Search, and Video hyperlinking,” Proceedings of TRECVID 2017, 2017.

23. X. Li, J. Dong, C. Xu, J. Cao, X. Wang, and G. Yang, "Renmin University of China and Zhejiang Gongshang University at TRECVID 2018: Deep Cross-Modal Embeddings for Video-Text Retrieval," Proceedings of TRECVID 2018, 2018.
24. F. Faghri, D. J. Fleet, R. Kiros, and S. Fidler, "VSE++: Improved Visual-Semantic Embeddings, arXiv:1707.05612, 2017.
25. C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting Image Annotations Using Amazon's Mechanical Turk," Proc. of the NAACLHLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pp.139–147, 2010.
26. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics. vol.2, pp.67–78, 2014.
27. T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," arXiv:1405.0312, 2014.
28. P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning," Proc. of the 56th Annual Meeting of the Association for Computational Linguistics, pp. 2556–2565, 2018.
29. J. Xu, T. Mei, T. Yao, Y. Rui, "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language," In Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
30. G. V. Cormack, C. L. Clarke, and S. Buettcher, "Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods," Proc. of the 32nd International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp.758–759, 2009.
31. K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, "Stacked Cross Attention for Image-Text Matching," In Proc. of the European Conference on Computer Vision (ECCV), 2018.
32. Y. Wang, H. Yang, X. Qian, L. Ma, J. Lu, B. Li, X. Fan, "Position Focused Attention Network for Image-Text Matching," In Proc. of International Joint Conference on Artificial Intelligence (IJCAI), 2019.
33. C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, Y. Zhang, "Graph Structured Network for Image-Text Matching," In Proc. of Computer Vision and Pattern Recognition (CVPR), 2020.