Kindai University, Osaka Gakuin University and Osaka University at TRECVID 2022 AVS Task

Kimiaki Shirahama*, Kazuma Fujioka†, Taichi Shinno‡,
Takashi Matsubara§ and Kuniaki Uehara¶
* Faculty of Informatics, Kindai University
† Graduate School of Science and Engineering, Kindai University
‡ Department of Informatics, Kindai University
§ Graduate School of Engineering Science, Osaka University
¶ Department of Business Administration, Osaka Gakuin University
Contact: shirahama@info.kindai.ac.jp

***Abstract*—This paper presents our method developed for Ad-hoc Video Search (AVS) task in TRECVID 2022. Main models (SCAN models) in our method are trained on Conceptual Captions dataset [1]. Our method additionally uses pre-trained models (CLIP models) that are trained using WIT (WebImageText) dataset [2], LAION-400M and LAION-2B datasets [3]. Furthermore, our method employs object-centric features extracted by an object detection model pre-trained on the combination of MS-COCO, OpenImages, Object365 and Visual Genome datasets [4].**

Our method performs retrieval by fusing the following two types of component models: The first type is based on Stacked Cross Attention Network (SCAN) [5] to align regions in a shot with words and phrases in a topic (or its narrative). The second type follows Contrastive Language-Image Pre-training (CLIP) that conducts global alignment between a shot and a topic (or its narrative). Specifically, our four submitted runs are configured by the following fusions of component models:

1) **F_M_C_D_kindai_ogu_osaka.22_1: This run adopts a weighted fusion of three variants of SCAN characterised by different region-word/phrase alignment approaches and different architectures, and four variants of CLIP pre-trained with different architectures and training datasets. The weight of each model is determined based on grid search on V3C1 dataset for 20 topics in 2021 AVS task.**
2) **F_M_C_D_kindai_ogu_osaka.22_2: In addition to the seven models used in F_M_C_D_kindai_ogu_osaka.22_1, this run fuses one more variant of SCAN that performs region-word/phrase alignment by following the constituency tree of a topic (or its narrative).**
3) **F_M_C_D_kindai_ogu_osaka.22_3: This run is the same to F_M_C_D_kindai_ogu_osaka.22_2 except for that the eight models are fused with the same weight.**
4) **F_M_C_D_kindai_ogu_osaka.22_4: This run is the same to F_M_C_D_kindai_ogu_osaka.22_1 except for that the seven models are fused with the same weight.**

The evaluation results show that the MAP of F_M_C_D_kindai_ogu_osaka.21_1 is 0.199 while the MAP of its counterpart F_M_C_D_kindai_ogu_osaka.22_4 is 0.186. In addition, the MAP of F_M_C_D_kindai_ogu_osaka.21_2 and the one of its counterpart F_M_C_D_kindai_ogu_osaka.22_3 are 0.195 and 0.177, respectively. These results verifies the effectiveness of the weighted fusion. Also, the fact that F_M_C_D_kindai_ogu_osaka.21_1 outperforms F_M_C_D_kindai_ogu_osaka.21_2 and F_M_C_D_kindai_ogu_osaka.21_4 outperforms F_M_C_D_kindai_ogu_osaka.21_3 indicates the ineffectiveness of the additional SCAN model based on the constituency tree-based alignment approach. One possible reason is the insufficient parameter optimisation and hyper-parameter tuning due to the time limitation until the submission deadline. We will present the finalised performance of the additional SCAN model at the workshop.

I. Introduction

We are continuously participating in TRECVID to objectively compare the performance of our system to those of systems developed all over the world [6]. This year we participated in Ad-hoc Video Search (AVS) [7] in order to address the following two issues:

The first issue is to devise a high-quality fine-grained matching between visual features of a shot and textural features of a topic by considering the linguistic structure of the topic. This is inspired by our past experiences to apply Stacked Cross Attention Network (SCAN) [5] to AVS task [8], [9], [10]. Roughly speaking, SCAN was used to compute the relevance of a shot to a topic by aligning regions in the shot with words in the topic. However, this alignment does not fit human's perception because he/she checks not only whether regions corresponding to words exist in the shot, but also whether those regions suit to phrases involving multiple words. Last year we attempted to overcome this by extending SCAN to align regions in a shot with words and phrases obtained by extracting the constituency tree of a topic [9]. However, this approach lacked examining the consistency of aligned regions. For example, different regions are aligned with the phrase "red dress" and its component noun "dress". To avoid this kind of inconsistent alignment, our SCAN model is further extended to perform region-word/phrase alignment by following the hierarchical relations represented by the constituency tree of the topic.

The second issue is to examine the effectiveness of models that are pre-trained using a vast amount of image-caption pairs collected from the Internet. Compared to traditional training data that are labelled for a predetermined, limited set of visual concepts, directly using raw text not only eases the collection of a very large amount of training data, but also offers supervision on a much wider range of visual concepts. Thus, a model pre-trained on such image-caption pairs has a generality

to flexibly deduce the relevance of a shot's keyframe to a linguistically complex topic. In particular, our AVS method employs four pre-trained models, the first and second models that are trained on a set of 400 million image-caption pairs [2], and the third and fourth model that are respectively trained on another set of 400 million image-caption pairs and a set of two billion pairs [3].

II. Our AVS Method

Our AVS method measures the relevance of a shot to a topic is computed as the (weighted) sum of scores obtained by each of component models. Below, those component models are described by categorising them into two types, SCAN and CLIP.

A. SCAN Models

This section presents three SCAN models fused in our AVS method. The first model is the original SCAN that performs alignment between regions in a shot and words in a topic (or its narrative). The second model is an extended SCAN where the alignment additionally considers phrases obtained by extracting the constituency tree of a topic (or its narrative). The last model is a further extension of SCAN to conduct hierarchical alignment between regions and words/phrases by following the constituency tree of a topic (or its narrative) in a bottom-up fashion. For simplicity, the first, second and third SCAN models are termed *SCAN_org*, *SCAN_phrase* and *SCAN_hier*, respectively. Below, these models are sequentially explained.

SCAN_org aligns regions in an image with words in a caption based on an attention mechanism. Roughly speaking, an attention between a region and a word is computed as their normalised similarity lying between 0 and 1. This attention represents a probabilistic relevance of matching the region with the word. Then, the "word-level" relevance indicating an overall suitability of all regions for the word is computed as the cosine similarity between the word's feature and the average of regions' features weighted by their attentions to the word. Finally, the "caption-level" relevance of the image is calculated as the average of word-level relevances over all words in the caption. In this framework, the FC layer in the region encoder, and the word embedding layer and the bidirectional GRU in the word encoder are optimised so that the caption-level relevances are high and low for relevant and irrelevant image-caption pairs, respectively. This optimisation consequently leads to semantically meaningful alignment between regions and words. We firstly train *SCAN_org* using Conceptual Captions dataset [1] containing about three million image-caption pairs, and apply it to AVS task by regarding an image and a caption as the keyframe of a shot and a topic (or its narrative), respectively. For more details, please refer to the original paper of SCAN [5] and our notebook paper in 2019 [8].

This year one extension is made on *SCAN_org*. Until last year *SCAN_org* used a bottom-up attention model [11] to extract a fixed number of salient regions (i.e., 36) from an image and encode each of them into a 2048-dimensional visual feature. This bottom-up attention model is implemented with Faster R-CNN based on ResNet101 backbone and trained on Visual Genome dataset [12]. However, this model is now more than five years old, and it has been empirically proven that visual features extracted by VinVL [4] offer significantly higher performances on different tasks. The reason for this is that compared to the bottom-up attention model, VinVL uses Faster R-CNN based on a more advanced model (ResNeXt-152) and trained on a much larger training dataset, which is the combination of MS-COCO, OpenImages, Objects365 and Visual Genome [4]. Thus, for *SCAN_org* as well as *SCAN_phrase* and *SCAN_hier*, the bottom-up attention model is replaced with VinVL. Since VinVL extracts different numbers of regions for different images, the alignment process is also extended to align a variable number of regions in an image with words (and also phrases) in a caption.

SCAN_phrase is an extended version of *SCAN_org* so as to include phrases in a caption into the alignment process. In other words, phrases are treated as additional words to be aligned with regions. Except this, *SCAN_phrase* is exactly the same to *SCAN_org*. Thus, in what follows we describe how to obtain phrases in a caption and how to extract the feature of each phrase as well as the one of each word. First, phrases are acquired by extracting a constituency tree of a caption with the parser developed in [13]. Roughly speaking, the parser starts with an empty tree and sequentially adds each word to the tree by selecting the action to attach the word as a child node of an existing node or the action to juxtapose the word as a sibling to an existing node by creating a shared parent node. A neural network for this action selection is trained on Wall Street Journal part of Penn Treebank (PTB) dataset [13]. Fig. 1 (a) shows an example of constituency tree extracted for "a hang glider floating in the sky on a sunny day".

Words and phrases are encoded using a Tree-LSTM that is an extended LSTM to propagate hidden states and memory cells based on the topology of a tree [14]. In particular, a Child-Sum Tree-LSTM is used to perform the following bottom-up propagation: Given a constituency tree of a caption, words corresponding to leaf nodes are firstly encoded into 300-dimensional vectors via a word embedding layer. These vectors are then used to compute 1024-dimensional hidden states and memory cells for leaf nodes. The process for each internal node starts with defining an "overall" hidden state from its child nodes as the sum of their hidden states. The overall hidden state is used to compute values of the input and output gates. A forget gate value is separately calculated for each child node to signify whether it is strongly related to the node or not. Afterwords, a hidden state and a memory cell for the node are obtained using the gate values. This way, hidden states and memory cells are propagated from leaf nodes to the root node corresponding to the whole caption. Note that no external input like word embedding features exist for internal nodes. In other words, the propagation for each internal node is based only on hidden states from its child nodes. The hidden state of each node is regarded as the feature

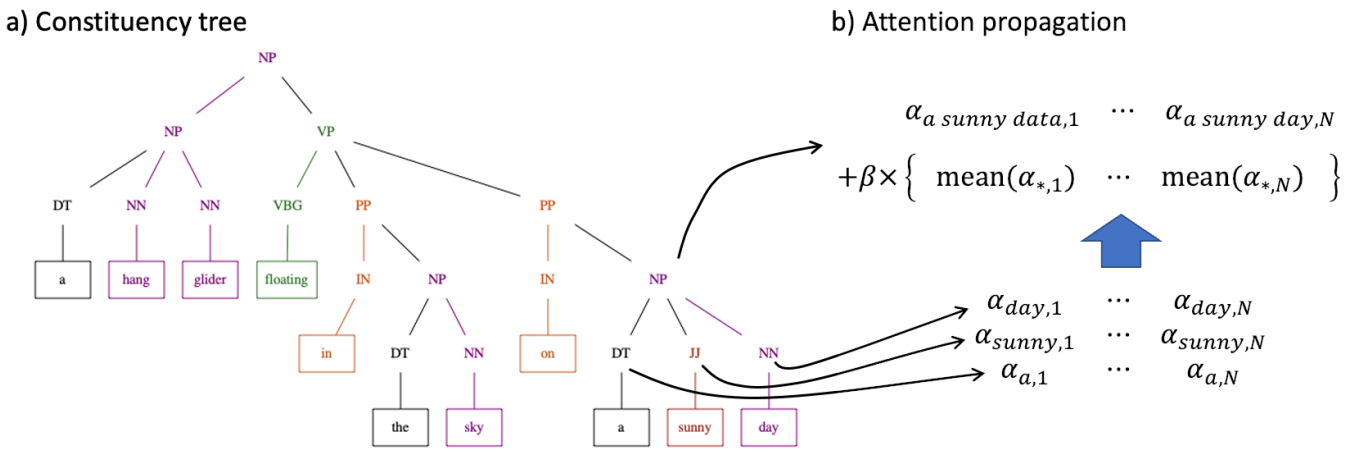

Fig. 1. An illustration of a constituency tree extracted for "hang glider floating in the sky on a sunny day" (a) and the attention propagation in *SCAN_hier* (b).

of the corresponding word or phrase.

To simplify the explanation of *SCAN_hier*, we collectively call words and phrases "tokens" as long as there is no need to distinguish between them. *SCAN_hier* is an extension of *SCAN_phrase* that independently align each token with regions. In other words, although the constituency tree of a caption indicates the hierarchical relations among tokens, they are ignored in *SCAN_phrase*'s alignment process. As a result, different regions are inconsistently aligned with a phrase and its component words. To resolve this, *SCAN_hier* performs region-token alignment based on the constituency tree of a caption, as illustrated in Fig. 1 (b). Let us assume that a token t consisting of M child tokens $\{t'_m\}_{m=1}^{M}$ is now being aligned with N regions in an image. In Fig. 1 (b), t is "a sunny day" and t'_1, t'_2 and t'_3 are "a", "sunny" and "day", respectively. Based on [5], attentions on N regions for t and those for t'_m $(1 \leq m \leq M)$ are computed as $\alpha_{t,1}, \cdots, \alpha_{t,N}$ and $\alpha_{t'_m,1}, \cdots, \alpha_{t'_m,N}$, respectively. Under the above-mentioned setting, for the nth region $(1 \leq n \leq N)$, we aim to propagate the attentions $\{\alpha_{t'_m,n}\}_{m=1}^{M}$ for M child tokens to the attention calculation for t. Thereby, it is more likely to align t with regions similar to the ones aligned with child tokens. More specifically, the following equation means that the attention $\alpha_{t,n}$ on the n region for t is fused with the mean of attentions $\{\alpha_{t'_m,n}\}_{m=1}^{M}$ for M child tokens:

$$\alpha'_{t,n} = \alpha_{t,n} + \beta \frac{1}{M} \sum_{m=1}^{M} \alpha_{t'_m,n} \ , \tag{1}$$

where β is a weight to control the influence of the attention propagation from child nodes. This way attentions for tokens are propagated to the root token corresponding to the whole caption in a bottom-up fashion. Except for using the updated attention $\alpha'_{t,n}$ for each token, *SCAN_hier* is the same to *SCAN_phrase*.

B. CLIP Models

A Contrastive Language-Image Pre-training (CLIP) model is a transformer-based image-caption embedding model that is trained using a very large dataset containing 400 million image-caption pairs [2]. A CLIP model consists of an image encoder and a text encoder that are jointly trained to encode images and captions into embeddings (feature vectors), in terms of which cosine similarities for relevant image-caption pairs are high while those for irrelevant pairs are low. Such a CLIP model is used to compute an embedding of a topic, and an embedding of each shot's keyframe. Then, retrieval is performed by ranking shots in descending order of cosine similarities of shots' embeddings to the embedding of the topic. We use two CLIP models, *ViT-B/32* and *ViT-L/14*, each of which is based on a vision transformer with a different architecture and input patch size. In addition, we employ two more CLIP models, *LAION-400M* and *LAION-2B*, which are models structured as *ViT-B/32* and trained using a set of 400 million image-caption pairs and a set of 2 billion pairs, respectively [3]. Since the above-mentioned four CLIP models produce different retrieval results, their fusion is expected to yield an improved performance.

III. Results

Our submitted four runs are configured by combining *SCAN_org*, *SCAN_phrase*, *SCAN_hier*, *ViT-B/32*, *ViT-L/14*, *LAION-400M* and *LAION-2B* as follows:

1) F_M_C_D_kindai_ogu_osaka.22_1: This runs performs the fusion of the seven models, *SCAN_org-1024*, *SCAN_org_2048*, *SCAN_phrase*, *ViT-B/32*, *ViT-L/14*, *LAION-400M* and *LAION-2B*, which are weighted by 0.4, 0.6, 0.2, 0.7, 0.7, 0.5 and 0.6, respectively. Here, *SCAN_org-1024* differs from *SCAN_org_2048* only in the dimensionality of image and caption embeddings, that is, the former and latter encode images and captions into 1024- and 2048-dimensional vectors, respectively.

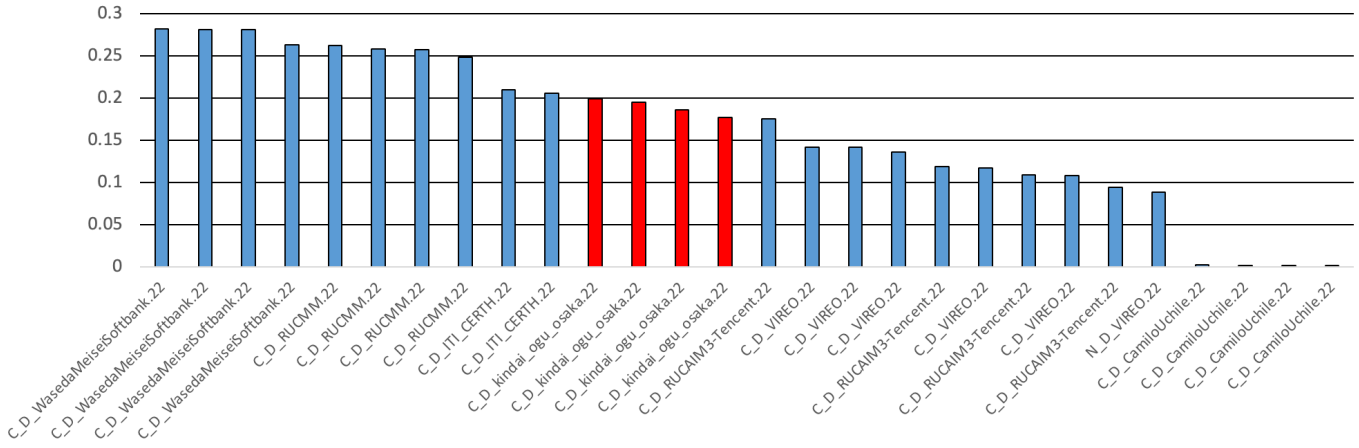

Fig. 2. Ranking list of the runs submitted to for the main AVS task.

In addition, the fusion weights are determined by performing grid search on V3C1 dataset with 20 topics at 2021 AVS task.

2) F_M_C_D_kindai_ogu_osaka.22_2: In addition to the above-mentioned seven models, this run further fuses *SCAN_hier* with a weight of 0.2.
3) F_M_C_D_kindai_ogu_osaka.22_3: This run is a counterpart of F_M_C_D_kindai_ogu_osaka.22_2 and fuses the eight models with equal weights. That is, this run is used to examine the effectiveness of the weighted fusion.
4) F_M_C_D_kindai_ogu_osaka.22_4: This run is a counterpart of F_M_C_D_kindai_ogu_osaka.22_1 and fuses the seven models with equal weights in order to check the effectiveness of the weighted fusion.

Note that the result of each component model is produced by fusing the retrieval result using a topic and the one using its narrative. That is, the topic and narrative are treated as separate textual descriptions for each of which retrieval is performed. Afterwords, retrieval results using the topic and narrative are fused with equal weights.

Fig. 2 shows the ranking list of the runs submitted to the main AVS task. As depicted by the four red-coloured bars, our team is ranked at the fourth position among the seven teams. In particular, among our four submitted runs, F_M_C_D_kindai_ogu_osaka.22_1 achieves the highest MAP 0.199 and its counterpart using non-weighted fusion yields the MAP 0.186. Similarly, F_M_C_D_kindai_ogu_osaka.21_2 attains the second highest MAP 0.195 while the MAP of its counterpart F_M_C_D_kindai_ogu_osaka.22_3 is 0.177. These results validate the effectiveness of the weighted fusion used in F_M_C_D_kindai_ogu_osaka.22_1 and F_M_C_D_kindai_ogu_osaka.22_2. Regarding performances for each topic, F_M_C_D_kindai_ogu_osaka.22_1 achieved the best AP 0.133 among all the submitted runs for topic 710. Moreover, F_M_C_D_kindai_ogu_osaka.22_3 gains the best AP 0.098 for 707.

Also, the ineffectiveness of *SCAN_hier* is indicated by the comparison between F_M_C_D_kindai_ogu_osaka.22_1 and F_M_C_D_kindai_ogu_osaka.22_2 and the one between F_M_C_D_kindai_ogu_osaka.22_3 and F_M_C_D_kindai_ogu_osaka.22_4 in Fig. 2. For both of the comparisons, the performances are degraded by adding *SCAN_hier*. One possible reason is the insufficient training of *SCAN_hier*. Due to the massiveness of Conceptual Captions and the time limitation until the submission deadline, the number of epochs to train *SCAN_hier* is much smaller than the numbers of epochs to train *SCAN_org* and *SCAN_phrase*. We will present the finalised performance of *SCAN_hier* at the workshop.

IV. Conclusion and Future Work

This paper introduced our method developed for TRECVID 2022 AVS task. It fuses seven or eight models belonging to one of the two types, SCAN or CLIP. Compared to our last year's method, the main advance is the development of *SCAN_hier* that performs image-word/phrase alignment by following hierarchical relations represented in the constituency tree of a topic. Although we submitted the four runs that were expected to yield the best performances and saw the effectiveness of the weighted fusion, the evaluation to examine the effectiveness of *SCAN_hier* has not yet been finished. Our urgent future work is to finish the performance evaluation of *SCAN_hier* including its hyper-parameter tuning.

References

[1] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. of ACL 2018*, 2018, p. 2556–2565.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. of ICML 2021*, 2021, pp. 8748–8763.

[3] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," in *Proc. of DCAI 2021*, 2021.

[4] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "Vinvl: Revisiting visual representations in vision-language models," in *Proc. of CVPR 2021*, 2021, pp. 5575–5584.

[5] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. of ECCV 2018*, 2018, pp. 212–228.

[6] G. Awad, K. Curtis, A. A. Butt, J. Fiscus, A. Godil, Y. Lee, A. Delgado, J. Zhang, E. Godard, B. Chocot, L. Diduch, J. Liu, Y. Graham, , and G. Quénot, "An overview on the evaluated video retrieval tasks at trecvid 2022," in *Proc. of TRECVID 2022*, 2022.

[7] J. Lokoč, W. Bailer, K. Schoeffmann, B. Muenzer, and G. Awad, "On influential trends in interactive video retrieval: Video browser showdown 2015–2017," *IEEE Transactions on Multimedia*, vol. 20, no. 12, p. 3361–3376, 2018.

[8] K. Shirahama, D. Sakurai, M. Takashi, and K. Uehara, "Kindai university and kobe university at TRECVID 2019 avs task," in *Proc. of TRECVID 2019*, 2019.

[9] D. Mukai, R. Utsunomiya, S. Utsuki, K. Shirahama, M. Takashi, and K. Uehara, "Kindai university and osaka gakuin university at trecvid 2020 avs and actev tasks," in *Proc. of TRECVID 2020*, 2020.

[10] K. Shirahama, T. Sato, N. Yamawaki, T. Matsubara, and K. Uehara, "Kindai university and osaka gakuin university and osaka university at trecvid 2021 avs task," in *Proc. of TRECVID 2021*, 2021.

[11] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. of CVPR 2018*, 2018, pp. 6077–6086.

[12] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 32–73, 2017.

[13] K. Yang and J. Deng, "Strongly incremental constituency parsing with graph neural networks," in *Proc. of NeurIPS 2020*, 2020.

[14] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. of ACL 2015*, 2015, pp. 1556–1566.