Analysis of the Complementarity of Latent and Concept Spaces for Cross-Modal Video Search

Varsha Devi, Philippe Mulhem, Georges Quénot

Analysis of the Complementarity of Latent and Concept Spaces for Cross-Modal Video Search

Varsha Devi, Georges Quénot, Philippe Mulhem
Univ. Grenoble Alpes, CNRS,
Grenoble INP (Institute of Engineering Univ. Grenoble Alpes), LIG,
Grenoble, France
firstname.lastname@univ-grenoble-alpes.fr

***Abstract* – This paper focuses on studying the complementarity between the spaces from hybrid cross-modal state-of-the-art systems for video retrieval like [5]. We aim at investigating if these spaces really convey different features, or if they are representing the same things. We use PCA (Principal Component Analysis) to study the optimal dimensions, CCA (Canonical Correlation Analysis) to assess the similarity of the spaces, and check if such approach is in fact similar to ensemble learning. We achieve experiments on the MST-VTT corpus, and show that in fact these two spaces are indeed very similar, paving the way for new models that could enforce more dissimilar spaces.**

***Keywords* –**Latent Space, Concept Space, Correlation, CCA, Ensemble Learning, Cross Modal Retrieval

I. Introduction

Cross-Modal Video-Text retrieval makes use of both natural language textual representations and video content. The video and the textual parts are encoded into a powerful representation of their own and are mapped to one common space in which the similarity of different modalities reflects the semantic closeness between them. The cross-modal retrieval task is challenging due to the semantic gap between high level semantics expressed in terms of textual query and low level video features. The mainstream approaches for cross-modal retrieval are *Concept Based* [9], [17], [22], [25]–[28] or *Concept Free* [18], [24], [31], [33].

In a Concept Based approach, the textual query and videos are mapped to visual concepts (actions, objects, places etc.) using some linguistic rules and pre-trained Convolutional Neural Networks (CNNs) respectively. The similarity between videos and text is computed based on the similar visual concepts between them. This similarity is able to support explanation of retrieved results. Even though having the advantage of explainability, these systems have a major drawback, i.e. selection of concepts for queries and reliability of concept classifiers used to index videos.

On the other hand, Concept Free approaches represent videos and text into their respective features, and then map those features onto the learned latent space for similarity computation. The similarity between videos and textual queries in this case is not explainable. Examples of such models are Visual Semantic Embedding (VSE++) [8], Word2VisualVec [23], Dual Encoding [4]. Most of the latent space models are commonly trained using a pair-wise ranking loss function [4], [19].

In addition to Concept Based and Concept Free models, the late fusion of both spaces has become the norm in recent state-of-the-art and TRECVid benchmarking [5], [10], [13], [25], [29], [30]. The results of such hybrid approaches have shown that the combination of concept free and concept based achieves better performance [5],[30]. So, it is fair to assess that Concept Based and Concept Free are complementary to each other. However, to our knowledge, no specific study has been done on the analysis of the complementarity or redundancy of such hybrid approaches.

This paper presents a detailed analysis of hybrid approaches, including various visual and statistical methods, to explore the intra and inter relationship of latent space and concept space. In Section II, we present related works on concept based, concept free, and hybrid approaches, along with some ways to explore single or multiple embedding spaces. Section III lists the three research questions that we are focusing on. Section IV describes the experiments we performed. We present and discuss the results in section V before concluding in Section VI.

II. Related Work

A. Cross-Modal Retrieval

A Cross-Modal retrieval system processes Text-to-Video (TTV) tasks, where a textual query is used to retrieve the videos, and Video-To-Text (VTT) tasks, in which the queries are in the form of videos and text captions are retrieved. *Concept-Based approaches* use concepts from large pre-defined concept vocabularies for mapping visual concept to a query and/or videos [9], [13], [17], [22], [25]–[29]. They perform better in TTV tasks when the concepts are accurately identified, but in practice, human intervention is often required in-order to filter concepts after automatic mapping [15], [22].

To tackle the problem of concept ambiguity, *Concept-Free approaches* map encoded videos and textual queries to high dimensional latent spaces. They are effective [4], [18], [24], [31], [33], but lack interpretability.

These latent spaces are trained using a pair-wise ranking loss function, which yields latent spaces effective for retrieval, and its optimization allows incorporation of domain knowledge in the loss function. *Yan et al.* [34] introduced Deep Canonical Correlation Analysis (DCCA) in order to learn latent space, able to find low-level semantic correlation.

In [6], a pair-wise ranking loss function and DCCA were combined in a way that maintains their strengths. The authors

proposed the Canonical Correlation Analysis Layer (CCAL) in a neural network to produce maximally correlated latent representation of two modalities that lead to good results.

Hybrid approaches, that apply a late fusion of Concept-Based and Concept-Free approaches, have become the norm in TRECVid benchmarking [5], [10], [13], [25], [29], [30]. These approaches show that both concept free and concept based are complementary to each other. *Wu et al.* [30] proposed a hybrid model based on [4] for the training of visual decoding network using the proposed class sensitive binary cross entropy (BCE) loss function, and [5] extended [4] to train the concept space using traditional BCE loss function. Such a model is simple, effective and end-to-end. Besides the increase in retrieval accuracy, the results of hybrid approaches are interpretable.

B. Ensemble Learning

Ensemble learning [16], [35] is a general and reliable technique, mostly used for classification, which co-ordinates the outputs of multiple supervised learning algorithms with the same architecture trained with different initializations using diversified data. Different initializations allow the machine learning models to have different learning paths, reducing the overall error by averaging out the individual error due to diversity of results and errors. There are two ways to design ensemble learning algorithms. The first approach is to train the machine learning models independently several times in such a way that the resulting set of models are accurate and diverse. The second approach [2] designs the ensemble algorithm in a coupled fashion, where the models are trained jointly and weighted scores for each model gives a good fit to the data [7], [11]. As our analysis is based on a dual encoding pipeline, which consists of CNN, hence, the ensemble learning technique in our case is applied to CNN-based models.

C. Visual Analysis of High Dimensional Latent and Concept Spaces

The video-text features extracted using neural networks are not human-interpretable [12], [21]. However, In order to interpret the latent space representation, many visual methods are proposed. One of the widely used methods [21] is based on Principal Component Analysis (PCA) for projection of a high dimensional latent vector on low dimensional basis vector. One can then observe if the properties of the high dimensional latent vectors are preserved in the reduced dimensional latent vectors. Domain specific methods may also be adopted to visualize semantic meanings in latent space. *Liu et al.* [20] maps attribute vectors created from opposing concepts, whereas [14] clusters latent variables according to hierarchical structures. In [1], the authors project both images and text on the latent space to demonstrate the relationship between them.

III. Research Questions

We aim at studying to which extent the concept-based and concept-free models, typically from [4], [5], are complementary.

The first aspect of this study is related to the optimal dimensions of each of these space, taken independently. The point here is to investigate if these two spaces, when learned independently, have similar optimal dimensions. If so, we may consider that these two spaces have similar representation capabilities, leading to consider that their complementarity does not come from strong intrinsic differences. This leads to our first research question:

R1: Is the number of optimum dimension the same in both the concept and latent spaces for two subtasks, Text-to-Video (TTV) and Video-to-Text (VTT)?

To answer **R1**, we will used side by side the initial spaces learned. In a second study related to this question, we use Principal Component Analysis (PCA) to explore the salient linear dimensions of these high dimensional spaces, and to study their variance.

Our other research questions are related to studying the complementarity of these spaces taken together, from two points of view:

R2: Do the latent and concept spaces represent complementary information?

R3: Does ensemble learning exhibit complementarity on the latent and concept spaces?

To answer **R2**, we study the correlation between spaces. We know that the latent space has free axes in the former and that the axes of the concept space are constrained by the concepts: the analysis of correlation between two different high dimension features spaces using Canonical Correlation Analysis (CCA) will be able to answer this question. The answer to question R3 will compare the performances of these two spaces used jointly, using ensemble learning (cf. section IV.D): if the results using ensemble learning are the same, then the two spaces are in fact similar.

IV. Experiments

We describe here the context of our experiments, and the experiments conducted (dimensions study, correlation study using CCA and ensemble learning).

A. Experimental Context

1) Dataset: We performed all of our experiments on MSR-VTT dataset [32]. It consists of 10K video clips, each being described by 20 captions (in total of 200K natural language sentences). In all the experiments reported here, we use the official split of MSR-VTT i.e. 6,513 video clips for training of the dual encoding model, 497 for validation, and 2,990 clips for testing.

2) Evaluation Metrics: We report $R@K$ ($K = 1, 5, 10$), Median rank (Med r) and mean Average Precision (mAP), where $R@K$ is the percentage of test queries for which at least one relevant item is found among the top-K retrieved results. Med r is the median rank of the first relevant item in the search results. The performance will improve as the value of $R@K$ and mAP increases, and the value of Med r decreases. All-average mAP is the average of mAP in TTV and VTT. For overall comparison, we also report the sum of all recalls in some cases.

3) Implementation Details: We chose the dual encoding model proposed by *Dong et al.* [5] for the mapping of visual

and textual representation in hybrid space, i.e. shared latent and concept space, as it achieves state-of-the-art performance. We use the PyTorch code provided by the dual coding model[1] to set up the basic architecture of a visual encoding network and a textual encoding network. We employ the frame-level video features of 4,096 dimensions for each video frame provided by the authors, extracted using the pre-trained ResNet-152 [3] and ResNext-101 [26]. For the video-text concept features, the concept list is compiled from the training set captions of MSR-VTT. We use a learning rate of 0.0001 and Adam optimizer to train the model, with the batch size of 128.

B. Optimal Dimensions Study

We explore the latent and concept spaces in order to evaluate the number of optimum dimensions for both spaces and their performances in their respective optimum regions. The aim is to find the optimum regions by evaluating the dual encoding model for varying numbers of dimensions in latent space and concept space, and to compare both spaces in order to see if both the spaces are similar with respect to their optimum dimensions.

For the latent space, after the extraction of video and text latent features using the dual encoding model [5], we map different dimensional latent features to the latent spaces and analyze the performance of the dual encoding model to explore the optimum regions.

For the concept space, we studied the performance of dual encoding model trained with concept space only while varying the number of dimensions[2]. For that, we first extract video and text concept features using the dual encoding model [5] and map them onto latent space with varying number of dimensions, in order to find the optimum regions for a concept space.

C. Dimensionality Study using PCA

We study the dimensionality of latent space and concept space using PCA to explore the distribution of data of these high dimensional features space, in a way to explore the salient linear dimensions of these high dimensional spaces. We project the high dimensional latent features and concept features onto low-dimensional space by employing PCA. By estimating the variance of the top $\mathscr{K}$ principal components of latent space, and concept spaces, we can estimate that whether the optimum dimension for both spaces correspond to the same number as the one obtained in the optimal dimension study above in Section B.

D. Complementarity Study using CCA and Ensemble Learning

We study here the complementarity between the latent and concept spaces. This is done using two approaches. The first one is to analyze the correlation between two different high dimensional feature spaces using Canonical Correlation Analysis (CCA). This will evaluate the similarity in their representations. The second one is to compare the performances of these two spaces used independently and jointly, using ensemble learning: if the results using ensemble learning are the same than assuming complementarity, this will show that the two spaces are similar.

1) Correlation between Vector Spaces using CCA:
Using the notation of [5], consider the set of video features $f(v)$ and text features $f(s)$ in latent space as $\mathscr{X}_l \in \mathscr{R}^{N\times p}$ and the set of video features $g(v)$ and text features in concept space $g(s)$ as $\mathscr{X}_c \in \mathscr{R}^{N\times q}$, with dimensions p and q respectively, with N number of videos and text/captions in the dataset. The feature vector for i^{th} video or text in latent and concept spaces can be denoted as $\mathscr{X}_l^i$ and $\mathscr{X}_c^i$ respectively. In this section, we investigate the relationships between the feature vectors of all video-text in latent space($\mathscr{X}_l$) and in concept space ($\mathscr{X}_c$) using Canonical Correlation Analysis (CCA).

Consider $\mathscr{M}$-Dimensional CCA transformed latent space features $\mathscr{X}_l^i \in \mathscr{X}_l$ and concept space features $\mathscr{X}_c^i \in \mathscr{X}_c$ for i^{th} video-text as $\tilde{\mathscr{X}}_l^i \in \mathbb{R}^{N\times\mathscr{M}}$ and $\tilde{\mathscr{X}}_c^i \in \mathbb{R}^{N\times\mathscr{M}}$ respectively, where $\mathscr{M}$ being chosen as minimum of the latent space dimension p and concept space dimension q, mathematically $\mathscr{M} = min(p,q)$. We define the highly correlated feature vectors $(\tilde{\mathscr{X}}_l, \tilde{\mathscr{X}}_c)$ as the projection of $\mathscr{X}_l$ and $\mathscr{X}_c$ onto CCA basis vectors, along with which the correlation was above the threshold $\mathscr{T}_h$. Let us denote two linear transformation matrices corresponding to these i^{th} correlated basis vectors ($\tilde{\mathscr{X}}_l^i$ and $\tilde{\mathscr{X}}_c^i$) by A_l^i and A_c^i respectively for latent space and concept space. The correlated projections of $\tilde{\mathscr{X}}_l^i$, and $\tilde{\mathscr{X}}_c^i$ are given by:

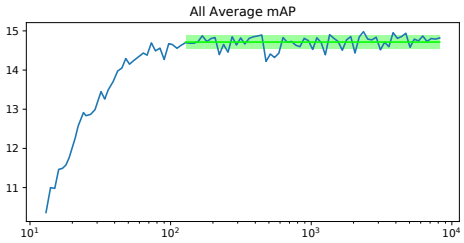
$$\begin{aligned}\tilde{\mathscr{X}}_l^i &= A_l^i \mathscr{X}_l^i \\ \tilde{\mathscr{X}}_c^i &= A_c^i \mathscr{X}_c^i\end{aligned} \tag{1}$$

Here $\tilde{\mathscr{X}}_l^i$, and $\tilde{\mathscr{X}}_c^i$ can be considered as correlated components embedded in $\mathscr{X}_l^i$, and $\mathscr{X}_c^i$. Hence, using these correlated components, we want to observe the correlation measure between variables of latent space vectors and concept space vectors. If there are groups with high correlation amongst variables which cover a good amount of variance, then there might be overlapped feature representation amongst the spaces, which answers **R2**.

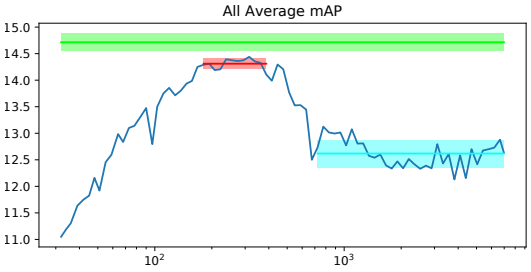
2) Ensemble Learning:
We analyze in detail the model from [5] with ensemble learning approaches. The analysis consists of training the dual encoding model in three different ways: (i) *Homogeneous non-coupled model*: where the latent space model and concept space model are trained independently with same configuration and hyper-parameters, then the retrieval accuracy is combined using weighted average of two models; (ii) *Homogeneous coupled model*: where the latent space and concept space are trained jointly with same hyper-parameters; and (iii) *Heterogeneous coupled model*: in which the latent space and concept space are trained jointly with different and the best chosen hyper-parameter for each (similar to the hyper-parameter setting in baseline model [5]). Hence, with all these model training scenarios, we want to observe the differences in the behaviour of latent space and concept space when learned in similar or different settings. We also want to see if the performance gain in TTV and VTT task is either due

[1] `https://github.com/danieljf24/hybrid_space`

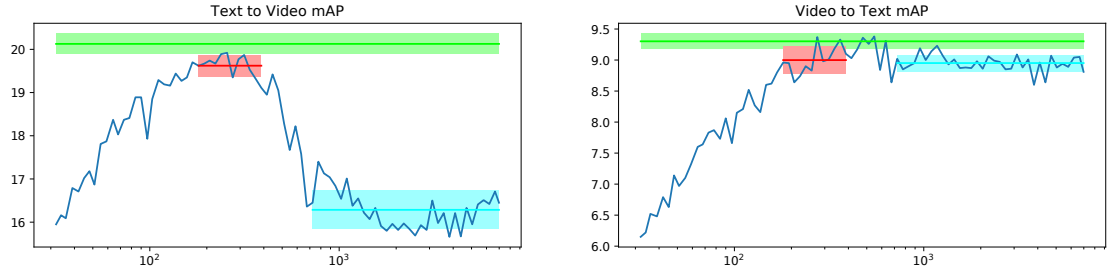
[2] The code provided by the authors was modified in order to do concept space training/testing only



(a) All average mAP for latent space



(b) All average mAP for concept space



(c) mAP for Text-to-Video (TTV) and Video-to-Text (VTT) separately for concept space

Fig. 1. **Optimum dimension study on MSR-VTT for latent and concept space**. X-axis represents the number of dimensions in latent space (a) and concept space (b, c) in log-scale, while Y-axis represents the mAP in %. Green line with bar represents the mean and standard deviation in the stable region of the latent space. The mean and standard deviations in the optimal regions for concept space are shown red line and bar for Text-To-Video (TTV) task, and cyan line and bar for Video-To-Text (VTT) task, respectively.

to ensemble learning or something else.

V. Results & Discussion

We answer here our three research questions.

R1: Is the number of optimum dimension the same in both the concept and latent spaces for two subtasks, TTV and VTT?

To evaluate the answer of this research question, we vary the dimensions of latent space and concept space in order to find the optimum regions (see Section B). Figure 1(a) shows the evolution of the average of the mAP for the TTV and VTT retrieval tasks as a function of the number of latent dimensions. This number is sampled from 16-D to 65536-D on a log scale with 10 samples per octave (16, 17, 18, 19, 21, 22, 24, 25, 27, 29, 32, 34 . . . 46340, 49667, 53231, 57052, 61147, 65536). The same plots for the TTV and VTT tasks separately are exactly similar (not shown). The performance reaches a plateau around 200 dimensions, and it is then stable until about 8000 dimensions, after which it decreases very slowly (not shown).

Figure 1(b) shows the evolution of the average of the mAP for the TTV and VTT retrieval tasks as a function of the number of concept dimensions. This number is sampled between from 16-D to 6983-D (i.e., the maximum number of selected concepts), also on a log scale with 10 samples per octave. This time, the same plots for the TTV and VTT tasks are very different and are shown in Figure 1(c). There are different optimal regions for the TTV and VTT tasks and these optimal regions are much narrower, around 200 for the TTV task and overall and beyond 500 for the VTT task.

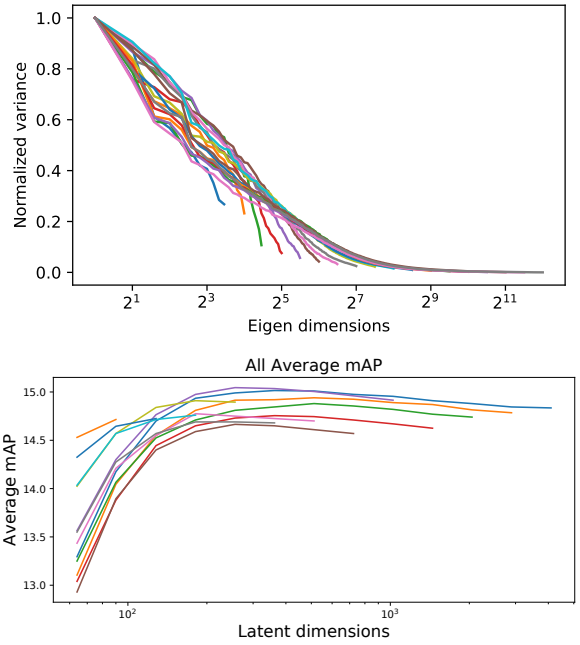


Fig. 2. PCA performance analysis for latent space. The X-axis represents the number of principal components, whereas Y-axis represents the normalized variance (top) and the all-average mAP in % (bottom).

TABLE I

Ensemble learning Experiments on MSR-VTT. Larger R@{1,5,10}, mAP and smaller Med r indicate better performance.

Method	Text-to-Video Retrieval					Video-to-Text Retrieval					SumR
	R@1	R@5	R@10	Med r	mAP	R@1	R@5	R@10	Med r	mAP	
Evaluation on a single space:											
Latent independent 1536 (10)	10.94	29.12	39.73	19.30	20.21	19.00	42.60	54.81	8.10	9.26	196.20
Latent independent 512 (20)	10.88	29.06	39.73	19.25	20.15	19.42	42.91	55.20	7.95	9.30	197.21
Latent-latent coupled homogeneous (10)	11.17	29.83	40.58	18.10	20.63	19.95	43.77	56.31	7.60	9.57	201.61
Latent-latent coupled heterogeneous (10)	11.37	30.25	41.11	17.80	20.93	19.85	43.70	56.26	7.60	9.63	202.54
Latent-concept coupled heterogeneous (10)	11.42	30.29	41.16	17.70	21.00	19.65	43.24	55.84	7.90	9.57	201.60
Evaluation on two spaces:											
Latent-latent indep. homogeneous (10)	11.50	30.30	41.22	17.55	21.04	20.92	44.78	56.99	7.25	9.89	205.71
Latent-latent coupled homogeneous (10)	11.41	30.31	41.16	17.70	20.98	20.30	44.57	56.85	7.10	9.80	204.60
Latent-latent coupled heterogeneous (10)	11.78	31.12	42.28	16.20	21.60	21.23	45.65	58.08	7.00	10.36	210.14
Latent-concept coupled heterogeneous (10)	11.76	30.98	42.10	16.40	21.52	20.25	44.74	57.48	7.10	10.09	207.31

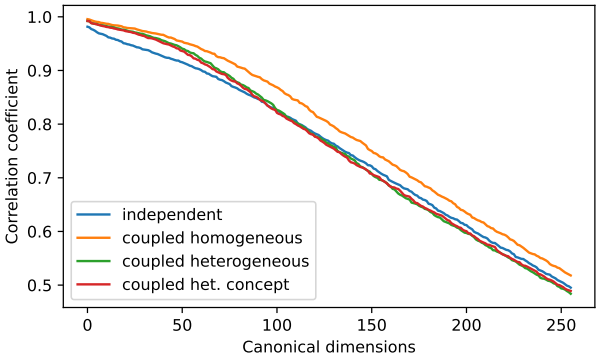


Fig. 3. CCA analysis: Top 256 canonical correlation of independent (non-coupled), coupled homogeneous, coupled heterogeneous and coupled heterogeneous concept training's. Most correlated is coupled homogeneous training of two spaces; whereas least correlated is independent training of two spaces.

Discussion Overall, the optimum number of dimensions is around 200 for both the latent and concept spaces, and this is much lower than the numbers used in [5] (1536-D and 512-D). The performance using concepts is a bit lower, but not much for the optimal values. The fact that the values are lower and even more outside of the optimal region can be explained by the fact that the classification task associated with the concept space places strong constraints on it. The asymmetry between the TTV and VTT behaviours can be explained by the asymmetry in the ratio of captions to videos. On all curves, fluctuations can be observed. These are due to the effect of the random initialization in the training, and they are at the same level as what is observed when running the same experiments multiple times.

The research question R1 is also studied using PCA to re-verify the number of optimum dimensions in latent space, as mentioned in Section C. For the sake of simplicity, we only analyse here the latent space. Figure 2(top) shows the decrease of the values of the normalized variances as a function of the eigen dimensions for multiple latent space trainings with a variable number of latent dimensions, ranging from 11 to 4096 on a log scale and with two samples per octave (11, 16, 22, 32, 45, 64 ... 1024, 1448, 2048, 2896, 4096). Curves are shown with different colors and the number of latent dimensions used for training can be inferred from the point at which the curve stops on the right side. As can be seen, whatever the initial number of latent dimensions, there is no significant variances beyond a few hundred eigenvalues (around $2^7 - 2^9$ eigenvalues). The cumulative variance continues to increase beyond (not shown) but this likely corresponds to noise.

Figure 2(bottom) shows the evolution of the performance (all averag mAP) of the same multiple trainings when reducing the size of the latent representations by keeping only the components with the highest variances. The difference in the initial performance (at the point which is most on the right for each curve) likely comes again from the random initialization and is also in the standard deviation obtained with multiple identical experiments. However, the variation of performance on each single curve corresponds to a same initialization and is expected to be significant. We see that for those starting with a high number of dimensions there is a slight increase in performance, confirming that the components with lowest variances contain mostly noise. The best performance seems to be reached by training with a number of dimensions larger than the optimum value found either directly (without PCA) or indirectly (with PCA) and then reducing the space size using PCA, *e.g.*, 4096 $\rightarrow$ 256.

Discussion As the number of optimum dimensions with and without PCA are very close for the latent space, these experiments show that even with dimensionality reduction of latent representation, the optimum regions of both spaces are still the same with the slight increase in performance due to noise reduction, which shows that (i) the properties of original high dimensional latent space are preserved in reduced dimensional space, and (ii) the observation holds that these two spaces may have a lot in common, leading to answer **yes to R1**.

R2: Do the latent and concept spaces represent complementary information?

To answer this question R2, we rely on a CCA study, in order to find out the correlation and complementarity between two spaces. One large difference between the latent space

and the concept space is that the concept space is associated with a classification task, while the latent space is not. If the classification task is removed, the concept space becomes just a second latent space with different characteristics (e.g., using a Jaccard similarity instead of a cosine). So, in our analysis, we consider four possible combinations of latent and/or concept spaces: (i) two identical latent spaces independently trained and lately fused (i.e., homogeneous non-coupled), (ii) two identical latent spaces jointly trained (i.e., homogeneous coupled), (iii) two different latent spaces (respectively with cosine and Jaccard similarities) jointly trained (i.e., heterogeneous coupled), and (iv) the same with additionally a classification task on the second latent space, which turns it into a concept space, and the overall system as the regular hybrid one (coupled heterogeneous concept). For better comparisons, we used 512 as the dimension for all spaces.

Figure 3 plots the top-256 canonical correlations of the latent-latent or latent-concept mappings in the four configuration just mentioned. The correlations are all quite high with similar profiles but still small differences. The independent training is the least correlated, the coupled homogeneous one is the most correlated, the other two, coupled heterogeneous and coupled heterogeneous concept (hybrid) are in between and almost identical, indicating that the classification task makes no difference in the correlation.

R3: Does ensemble learning exhibit complementarity on the latent and concept spaces?

To answer this question, we report quantitative evaluation of the four combinations described above used in ensemble learning experiments. The second part of table I shows the performance of the four combinations on the fused spaces using the standard MSR-VTT metrics, the last row corresponding to the regular dual encoding hybrid system [5]. The values correspond to the average of 10 identical runs with different random initialization so that we can get an estimate of the statistical significance of the differences between the experiments using a Z-test. When fused, there is no statistically significant difference between the independent and coupled trainings for the homogeneous latent spaces. The main best performance is achieved by the latent-latent coupled heterogeneous method that uses latent spaces of different types (with cosine and Jaccard similarities). The experiments on latent-latent coupled homogeneous underperform latent-latent coupled heterogeneous: there is a decrease in performance if cosine similarity is used in concept space. There is a slight decrease in performance when adding the classification task but the statistical significance is marginal. The first part of the table shows the performance when performing the task using the first latent space only and the first line is inserted for showing that there is no statistically significant difference between a 1536-D latent space and a 512-D one.

Discussion The experiments with CCA and ensemble learning showed that there is high correlation between the latent space and concept space when considering the same hyperparameters, for instance same distance metrics for calculating similarity in two spaces. There is the least correlation when considering independent training of two spaces, as the spaces are not optimized with the constraints present in the other space. Overall, we can answer **No to R2**. The ensemble learning experiments show that there is no significance difference in performance of two independent latent space with different dimensions. The significant improvement in retrieval comes from training two latent spaces with different similarity computation techniques. This analysis leads us also to answer **No to R3**.

VI. Conclusion

In this paper, we explored the complementarity of the latent and concept spaces of hybrid approaches for cross-modal video search, achieving state-of-the-art retrieval performance as well as providing some support for explanation. We reused the framework proposed in [5] to explore this problem from 3 different perspectives: a) the dimensionality of these spaces taken independently, b) their complementarity from the data content using CCA, and c) their complementarity for retrieval using ensemble learning. From these three perspectives, our experiments show that they represent the same information and the performance increase in retrieval is because of the heterogeneity of the similarities used in two spaces, i.e., Cosine vs. Jaccard. To our knowledge, this is the first time that such study is conducted. The side effect from our results is that, even though these latent and concept spaces are supposed to convey different information, they do not.

Our findings open the way for many future works. Among them, the first one is to explore ways to inspect approaches that may enforce stronger complementary of these spaces, leading to new hybrid approaches. Another direction for our work could concentrate on frameworks that support the study of spaces complementarity, for hybrid spaces in other contexts. Such framework could help the community to detail the behaviours of any hybrid spaces. In future, we would like to use the nonlinear decomposition for the analysis of latent and concept space and correlation between the two, considering the complexity of the inputs.

REFERENCES

[1] Jorge E Camargo and Fabio A Gonzalez. Multimodal visualization based on latent topic analysis. In *2014 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2014.

[2] Thomas G Dietterich et al. Ensemble learning. *The handbook of brain theory and neural networks*, 2(1):110–125, 2002.

[3] Jianfeng Dong, Shaoli Huang, Duanqing Xu, and Dacheng Tao. Dl-61-86 at trecvid 2017: Video-to-text description. In *TRECVID*, 2017.

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

[5] Jianfeng Dong, Xirong Li, Chaoxi Xu, Xun Yang, Gang Yang, Xun Wang, and Meng Wang. Dual encoding for video retrieval by text. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[6] Matthias Dorfer, Jan Schlüter, Andreu Vall, Filip Korzeniowski, and Gerhard Widmer. End-to-end cross-modality retrieval with cca projections and pairwise ranking loss. *International Journal of Multimedia Information Retrieval*, 7(2):117–128, 2018.

[7] Anuvabh Dutt, Denis Pellerin, and Georges Quénot. Coupled ensembles of neural networks. *Neurocomputing*, 396:346–357, 2020.

[8] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2017.

[9] Markatopoulou Foteini, Moumtzidou Anastasia, Galanopoulos Damianos, Mironidis Theodoros, Kaltsa Vagia, Ioannidou Anastasia, and Spyridon Symeonidis. Iti-certh participation in trecvid 2016. In *TRECVID 2016 Workshop*, 2016.

[10] Danny Francis, Phuong Anh Nguyen, Benoit Huet, and Chong-Wah Ngo. Eurecom at trecvid avs 2019. In *TRECVID*, 2019.

[11] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[12] Yordan Hristov, Alex Lascarides, and Subramanian Ramamoorthy. Interpretable latent spaces for learning from demonstration. In *Conference on Robot Learning*, pages 957–968. PMLR, 2018.

[13] Po-Yao Huang, Junwei Liang, Vaibhav Vaibhav, Xiaojun Chang, and Alexander Hauptmann. Informedia@ trecvid 2018: Ad-hoc video search with discrete and continuous representations. In *TRECVID Proceedings*, volume 70, 2018.

[14] Xiaonan Ji, Han-Wei Shen, Alan Ritter, Raghu Machiraju, and Po-Yin Yen. Visual exploration of neural document embedding in information retrieval: Semantics and feature selection. *IEEE transactions on visualization and computer graphics*, 25(6):2181–2192, 2019.

[15] Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 27–34, 2015.

[16] Bartosz Krawczyk, Leandro L Minku, João Gama, Jerzy Stefanowski, and Michał Woźniak. Ensemble learning for data stream analysis: A survey. *Information Fusion*, 37:132–156, 2017.

[17] Duy-Dinh Le, Sang Phan, Vinh-Tiep Nguyen, Benjamin Renoust, Tuan A Nguyen, Van-Nam Hoang, Thanh Duc Ngo, Minh-Triet Tran, Yuki Watanabe, Martin Klinkigt, et al. Nii-hitachi-uit at trecvid 2016. In *TRECVID*, volume 25, 2016.

[18] Xirong Li, Chaoxi Xu, Gang Yang, Zhineng Chen, and Jianfeng Dong. W2vv++ fully deep learning for ad-hoc video search. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 1786–1794, 2019.

[19] Xirong Li, Fangming Zhou, Chaoxi Xu, Jiaqi Ji, and Gang Yang. Sea: Sentence encoder assembly for video retrieval by textual queries. *IEEE Transactions on Multimedia*, 2020.

[20] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts, 2019.

[21] Yang Liu, Eunice Jun, Qisheng Li, and Jeffrey Heer. Latent space cartography: Visual analysis of vector space embeddings. *Computer Graphics Forum*, 38(3):67–78, 2019.

[22] Yi-Jie Lu, Hao Zhang, Maaike de Boer, and Chong-Wah Ngo. Event detection with zero example: Select the right and suppress the wrong concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 127–134, 2016.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[24] Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27, 2018.

[25] Phuong Anh Nguyen, Qing Li, Zhi-Qi Cheng, Yi-Jie Lu, Hao Zhang, Xiao Wu, and Chong-Wah Ngo. Vireo@ trecvid 2017: Video-to-text, ad-hoc video search, and video hyperlinking. In *TRECVID*, 2017.

[26] Cees GM Snoek, Xirong Li, Chaoxi Xu, and Dennis C Koelma. University of amsterdam and renmin university at trecvid 2017: Searching video, detecting events and describing video. In *TRECVID*, 2017.

[27] Cees GM Snoek and Marcel Worring. *Concept-based video retrieval*. Now Publishers Inc, 2009.

[28] Kazuya Ueki, Koji Hirakawa, Kotaro Kikuchi, Tetsuji Ogawa, and Tetsunori Kobayashi. Waseda_meisei at trecvid 2017: Ad-hoc video search. In *TRECVID*, 2017.

[29] Kazuya Ueki, Takayuki Hori, and Tetsunori Kobayashi. Waseda_meisei_softbank at trecvid 2019: Ad-hoc video search. In *TRECVID*, 2019.

[30] Jiaxin Wu and Chong-Wah Ngo. *Interpretable Embedding for Ad-Hoc Video Search*, page 3357–3366. Association for Computing Machinery, New York, NY, USA, 2020.

[31] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A graph-based framework to bridge movies and synopses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4592–4601, 2019.

[32] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[33] Ran Xu, Caiming Xiong, Wei Chen, and Jason J. Corso. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2346–2352. AAAI Press, 2015.

[34] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3441–3450, 2015.

[35] Xiao-Lei Zhang and DeLiang Wang. A deep ensemble learning method for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 24(5):967–977, 2016.