Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2759–2770
Marseille, 20-25 June 2022

Modality Alignment between Deep Representations for Effective Video-and-Language Learning

Hyeongu Yun[12*], **Yongil Kim**[12*], **Kyomin Jung**[12†]
[1]Automation and Systems Research Institute, [2]Seoul National University
{youaredead,miles94,kjung}@snu.ac.kr

Abstract

Video-and-Language learning, such as video question answering or video captioning, is the next challenge in the deep learning society, as it pursues the way how human intelligence perceives everyday life. These tasks require the ability of multi-modal reasoning which is to handle both visual information and text information simultaneously across time. In this point of view, a cross-modality attention module that fuses video representation and text representation takes a critical role in most recent approaches. However, existing Video-and-Language models merely compute the attention weights without considering the different characteristics of video modality and text modality. Such naïve attention module hinders the current models to fully enjoy the strength of cross-modality. In this paper, we propose a novel Modality Alignment method that benefits the cross-modality attention module by guiding it to easily amalgamate multiple modalities. Specifically, we exploit Centered Kernel Alignment (CKA) which was originally proposed to measure the similarity between two deep representations. Our method directly optimizes CKA to make an alignment between video and text embedding representations, hence it aids the cross-modality attention module to combine information over different modalities. Experiments on real-world Video QA tasks demonstrate that our method outperforms conventional multi-modal methods significantly with +3.57% accuracy increment compared to the baseline in a popular benchmark dataset. Additionally, in a synthetic data environment, we show that learning the alignment with our method boosts the performance of the cross-modality attention.

Keywords: Multi-modality, Video-and-Language Learning, Cross-modality Attention

1. Introduction

For deep learning researchers, multi-modality recently became an important keyword as multi-modal models have shown the ability to collate plentiful information scattered over various modalities (Cadene et al., 2019; Li et al., 2020b; Shen et al., 2018). In particular, Video-and-Language learning which includes both video modality and text modality is attracting a huge attention (Lei et al., 2020a; Zellers et al., 2021; Li et al., 2020a; Miech et al., 2019; Yu et al., 2019; Yang et al., 2021). Specifically, Video-and-Language learning such as video captioning or video question answering require the ability of reasoning over both time and multiple modality. For example, a video question answering model should be able to find appropriate visual information in a video frame sequence with a given question. That is to say, capturing the relationship between video information and text information is important for a multi-modal model for Video-and-Language learning.

Cross modality attention module which combines correlation over different modalities becomes a critical component for Video-and-Language learning (Ye et al., 2019; Lu et al., 2019; Chen et al., 2020b).Generally, the attention mechanism induces a model to learn the most important representation among the whole sequence with a given query. For single modality models, the attention module finds crucial parts to concentrate, greatly improves the performance of the model (Vaswani et al., 2017). However, the cross-modal attention mechanism in multi-modal models is less effective than in single modality models because of the noticeable differences characteristics between multiple modalities. Existing Video-and-Language models do not take this into account and merely utilize the attention mechanism as the same way as in single modality models, which hinders the models to fully enjoy the strength of the attention mechanism.

In this paper, we propose a novel Modality Alignment method that optimizes the alignment between representation structures of the video modality and the text modality. Our method leverages Centered Kernel Alignment (CKA) as an auxiliary objective to be maximized. As training the auxiliary loss via gradient descent frameworks, the embedding representation structures of both modalities are also trained to be similar. Therefore, our Modality Alignment method enhances the cross modality attention module inside a multi-modal model to be more aware of correlated information, eventually improving the final performance.

CKA was originally designed to measure similarity between neural networks representations (Cristianini et al., 2002). Recently, Kornblith et al. (2019) discovered the robustness of CKA, which comes from the invariance to orthogonal transformations and isotropic scaling. In this work, we reveal another desirable property of CKA that can be directly optimized through gradient descent frameworks. With the robustness and trainability of CKA, we utilize CKA in order to align

[*] Equal contribution
[†] Corresponding authors

multi-modal representations. As far as we know, this is the first attempt to exploit CKA as a training objective in handling multi-modality. Also, our Modality Alignment method can be easily applied to various multimodal tasks.

We validate our proposed method through experiments in three steps. Firstly, we show that aligning the embedding representations through maximizng CKA can effectively boost the performance on cosine similarity learning, which is a basis of the attention mechanism. Secondly, our experiments in Image Captioning task demonstrate that our Modality Alignment method helps especially the cross modality attention module where the attention score is computed on cosine similarity. Finally, in the real world Video QA task, we empirically demonstrate that out method makes a multimodal model to effectively learn the cross-modal attention. For TVQA (Lei et al., 2018) and TVQA+ (Lei et al., 2020a), which are challenging benchmarks in Video QA, the models applied with our method outperforms the baseline models.

Namely, our contributions are listed as followings:

- We show that Centered Kernel Alignment, a similarity measurement between neural network representations, can be exploited to align two embedding representations from different modalities.
- We demonstrate that our Modality Alignment method which optimizes the similarity between embedding representations is helpful for the cross-modal attention.
- We examine that our Modality Alignment method, which can be easily applied to existing models, improves the performance in various multi-modal tasks through extensive experiments.

2. Related Works

2.1. Representation Similarity between Modalities

Several works have attempted to analyze the similarity between representation in neural networks to achieve interpretability. The most fundamental measurements that can be used with this neural network similarity are correlation and Canonical Correlation Analysis (CCA) (Hardoon et al., 2004).

An alignment method using the correlation of neuron responses has been proposed to share core representations between different networks (Li et al., 2015). Similarly, Singular Vector Canonical Correlation Analysis (SVCCA) (Raghu et al., 2017) has been introduced in order to pick out perturbing directions from representations with applying CCA as a similarity measure. Morcos et al. (2018) subsequently have proposed Projection Weighted CCA (PWCCA) which is more reflective to subspaces of representations via projection. More recently, Kornblith et al. (2019) have shown that Centered Kernel Alignment (CKA) is a appropriate measure for representation similarity because CKA is robust to the lack of data.

Also, there have been studies that use these similarity measures between neural networks directly or indirectly in deep representation learning. By applying CCA, SVCCA, and CKA, Maheswaranathan et al. (2019) have discovered that the geometry of Recurrent Neural Network (RNN) architecture varies by task while the underlying scaffold is universal. On multilingual machine translation task, Kudugunta et al. (2019) have leveraged SVCCA across languages to show that there are shared representations among language representations. Bau et al. (2018) have applied SVCCA to identify meaningful directions in machine translation and concluded that the top-few directions of SVCCA similarity indicates a key representation.

Unlikely, we propose a method that directly optimizes CKA between multi-modal representation structures to be maximized. The robustness in CKA enables our method in the way that CKA is reliable even in a minibatch where the number of data is small.

2.2. Video Question Answering

Video-and-Language learning requires fine-grained interaction with information from multiple modalities. To study the fusion of visual modality and text modality, Image QA task which takes a single image input with a question in natural language has attracted the attention of many researchers (Li et al., 2020b; Zhang et al., 2021; Chen et al., 2020a). However, unlike single image processing, video information is made up of a large number of image frames in a sequence, which is much larger and includes additional temporal information.

To date, the *de facto* way to solve the Video QA task is to fuse and learn both modality information using cross-modal attention after processing the video input and text input respectively. The video processing part has been developed based on existing video analysis schemes, such as recurrent networks of frame functions (Kim et al., 2017) or 3D convolution operators for action recognition (Tran et al., 2018). Video representation is then fused via a co-attention module with textual input as query (Jang et al., 2017; Ye et al., 2017), a hierarchical attention (Liang et al., 2019; Zhao et al., 2018), or a memory networks module (Wang et al., 2019; Kim et al., 2019). These methods have applied their novel methods on how to fuse two modality information well, but they all merely combine multi-modality information without considering the differences in modality characteristics.

We observe that there is a significant difference in characteristics between the two modalities which may aggravate the cross modality attention. Thus, our method increases the similarity between multi-modality representation structures to enhance the fusion more effective. We validate our proposed method for synthetic dataset first, and then apply it to a real-world VideoQA dataset which has significant differences in characteristics between the two modalities.

3. Aligning Multi-modal Representations

With a new use of CKA as a learning objective, we propose a novel Modality Alignment method that directly maximizes CKA to align representation structures between various modalities.

3.1. Centered Kernel Alignment Review

As a tool to measure similarity between two deep representations, Centerend Kernel Alignment (CKA) has been proposed (Cristianini et al., 2002; Cortes et al., 2012). Recently, Kornblith et al. (2019) bring CKA back to the surface, addressing that CKA can aid in gaining a deep understanding of internal neural network architectures.

CKA is obtained by normalizing Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2005). For a pair of neural network representations $X_i = (x_{i1}, x_{i2}..., x_{iN})^T$ and $X_j = (x_{j1}, x_{j2}..., x_{jN})^T$, we define two matrices $K_{ikl} = \kappa(x_{ik}, x_{il})$ and $K_{jkl} = \kappa(x_{jk}, x_{jl})$ where κ is kernel function and N is a number of sampled data from each representation. Then, HSIC between two representations is computed as follows:

$$\text{HSIC}(K_i, K_j) = \frac{1}{(N-1)^2} tr(K_i C K_j C), \tag{1}$$

where C is a centering matrix $C = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ ($\mathbf{1}$ is a vector of ones and N is the number of sampled data). For linear kernels (e.g. $\kappa(x, y) = x^T y$), HSIC computes the squared Frobenius norm of the cross-covariance:

$$\left\|\text{cov}(X_i^T, X_j^T)\right\|_F^2 = \frac{1}{(n-1)^2} tr(X_i X_i^T X_j X_j^T). \tag{2}$$

Thus, HSIC can be interpreted as the similarity between the inter-example similarity structures. Normalizing HSIC results CKA as follows:

$$\text{CKA}(K_i, K_j) = \frac{\text{HSIC}(K_i, K_j)}{\sqrt{\text{HSIC}(K_i, K_i)\text{HSIC}(K_j, K_j)}}. \tag{3}$$

The normalizing process makes the output value of CKA between 0 and 1 where $\text{CKA}(X, Y) = 0$ implies independence. Also, this process makes CKA invariant to isotropic scaling.

3.2. Why CKA is proper to modality alignment

CKA exhibits desirable properties for not only measuring similarity between two deep representations but also training the alignment of inter-example similarity structures with gradient descent. We list three properties that enable our methodology for the modality alignment.

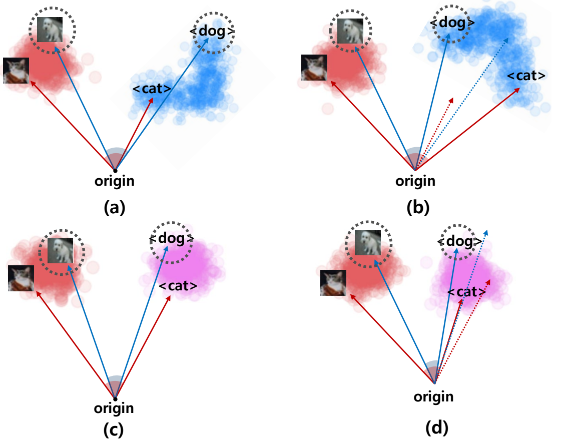

Figure 1: Main concept of Modality Alignment. *(a)*: During training cross modal attention module with a given mini-batch (inside the dotted circle), the model is trained to increase the attention score based on cosine similarity between the vector of "dog" and the correlated video frame vector. *(b)*: After a training step, the model is updated to narrow the gap. However, because the inter-example similarity structures are differently formed, there is potential harm to examples outside of the mini-batch; the cosine similarity between the "cat" vector and the correlated video vector decreases. *(c)* and *(d)*: Our method keep the inter-example similarity structures to be close to each other, significantly reducing such adverse effects.

- **Invariance to orthogonal transformations:** Kornblith et al. (2019) especially pointed out that CKA is invariant to orthogonal transformations of deep representation, i.e. $\text{CKA}(X, Y) = \text{CKA}(XU, YV)$ for any orthonormal matrices U and V. Because neural networks are randomly initialized and trained by gradient descent with random mini-batches, there is a high probability that neurons are permuted even in the same networks. Therefore, invariance to orthogonal transformations, which includes permutations, is one of the essential characteristics required for the similarity indexes.

  Although other similarities such as CCA (Hardoon et al., 2004) or SVCCA (Raghu et al., 2017) are invariant to affine transformations, Kornblith et al. (2019) spotted the limitation of invariance to affine transformations that it requires more examples than the size of dimension to robustly measure the similarity between representations. This limitation makes CCA and SVCCA unsuitable for training objective where the number of examples in a mini-batch is usually smaller than the size of dimension. However, unlike CCA or SVCCA, CKA shows robustness even with a small number of data (e.g. in a mini-batch).

- **Invariance to isotropic scaling:** CKA is also invariant to isotropic scaling, i.e. $\mathrm{CKA}(X, Y) = \mathrm{CKA}(\alpha X, \beta Y)$ for any $\alpha, \beta \in \mathbb{R}^+$. Invariance to isotropic scaling implies that CKA value remains the same even if each representation is scaled respectively, which often happens in neural networks training. Kornblith et al. (2019) also mentioned that invariance to non-isotropic scaling is not a desired property because a similarity index that is invariant to both orthogonal transformations and non-isotropic scaling is invariant to any invertible linear transformation, which lacks the robustness.

- **Trainabilty via gradient descent methods:** As CKA is calculated with fully differentiable operations such as dot-product, CKA itself is also differentiable with respect to the parameters of neural networks. That said, CKA itself can be used as a training objective for gradient descent algorithms. Yun et al. (2021) reported that CKA between representations of different layers in a model can be minimized or maximized via stochastic gradient descent. We exploit the trainability of CKA in order to align each representation in different modalities.

Using above three properties of CKA, we set CKA between video representation and text representation as an auxiliary training objective to be maximized. Note that maximizing CKA does not assure two representations to be overlapped. Nevertheless, it urges the inter-example structures of the two representations to be similar. For example, maximizing CKA between video representation and text representation makes the cosine similarity between a word "dog" and a word "cat" close to the similarity between a video frame with a dog and a video frame with a cat.

Meanwhile, the cross modality attention module computes its attention score based on cosine similarity between two vectors. In other words, the cosine similarity between a frame-level video embedding vector and a word-level text embedding vector becomes higher as the cross modality attention module is optimized, if there is semantic correlation between the video frame and the word. Maximizing CKA can enhance the training of the attention module since the inter-example structures of two different modalities are kept aligned.

Figure 1 further depicts the main concept of our Modality Alignment method. The objective of a cross attention module is to narrow the gap between two vectors with semantic correlation. Suppose fitting a cross attention module with a mini-batch which includes a word "dog" and a video frame with the dog (*(a)* of figure 1). After one training step, the parameters of networks are updated to narrow the angle between "dog" and the frame with the dog (*(b)* of figure 1). However, the cosine similarity between a word "cat" and a video frame with the cat decreases after the training step due to the difference of inter-example similarity structures. On the other side, with our Modality Alignment, such adverse effects are significantly reduced because the inter-example structures are also trained to be similar (*(c)* and *(d)* of figure 1). Hence, maximizing CKA between multi-modal representations can boost the training of cross modality attention, resulting fast convergence and higher performance.

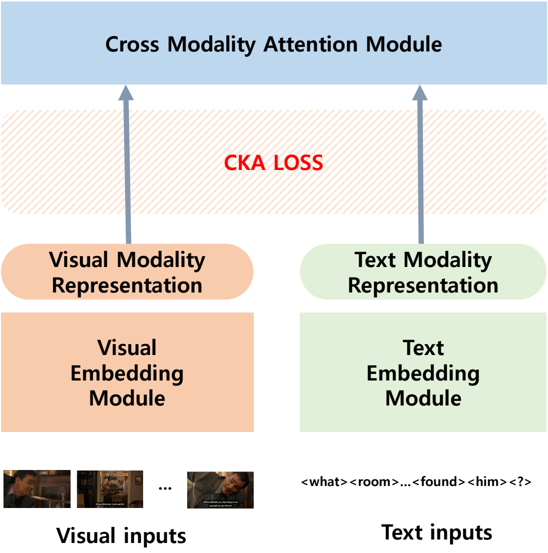
Cross Modality Attention Module
CKA LOSS
Visual Modality Representation
Text Modality Representation
Visual Embedding Module
Text Embedding Module
...
<what><room>...<found><him><?>
Visual inputs
Text inputs

Figure 2: Our proposed method. The input of each modality is embedded into the representation vector through each encoder module. CKA between representation vectors with different modalities is directly maximized to align the inter-example structure of each representation.

3.3. Our Proposed Method

Our proposed Modality Alignment method computes CKA between the output representation of the video embedding module and that of the text embedding module in each mini-batch and directly maximizes it as an auxiliary objective.

In Video-and-Language learning, a model usually consists of a video embedding module, a text embedding module, and a video-text fusion module. Let $V = [v_1, ..., v_L]$ be a sequence of video frames v_i and $T = [t_1, ..., t_M]$ be a sequence of word tokens t_j. A video embedding module f_{vid} encodes the sequence of video frames into the video embedding representation: $f_{vid}(V) = X$, where $X \in \mathbb{R}^{L \times d}$ is a sequence of embedded video representation vectors with dimension size of d. Similarly, a text embedding module f_{text} encodes the sequence of tokens into the text embedding representation: $f_{text}(T) = Y$, where $Y \in \mathbb{R}^{M \times d}$ is a sequence of text representation vectors. We randomly sample N-many representation vectors from both video representation tensor X and text representation tensor Y in order to match the number of examples. Finally, the modality alignment s between two representations is measured with equation (3).

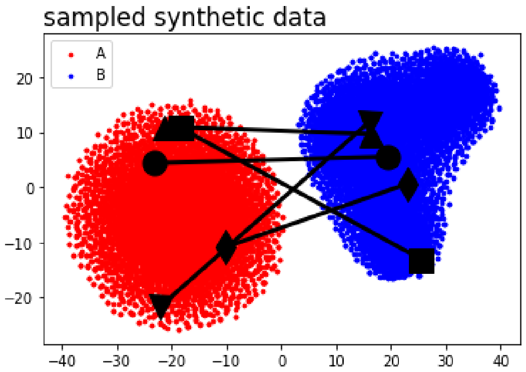

Figure 3: t-SNE visualization of our synthetic data distribution. We sampled two groups from different distributions and set one-to-one alignments to emulate hard attention.

We directly maximize the CKA as an auxiliary objective of the original loss to align two representations. Specifically, with a scaling hyperparameter λ_{cka}, we construct the final loss objective for minimizing by subtracting CKA loss term $\mathcal{L}_{cka}$ to the original loss term $\mathcal{L}_{orig}$ as follows:

$$\mathcal{L}_{final} = \mathcal{L}_{orig} - \lambda_{cka} * \mathcal{L}_{cka}. \quad (4)$$

Thus, our method can be applied to any model that handles multi-modality with cross attention module based on cosine similarity. We search the appropriate value of λ_{cka} by grid-searching in each experiment.

One can interpret our Modality Alignment method as a new variant of contrastive learning since our method takes account of relationships between data examples within a mini-batch. Our method has a novel strength in the respect that it can optimize the whole representation structures at every training step because of the robustness of CKA, while most of contrastive learning methods maximize the gap between irrelevant examples only in a mini-batch (Pan et al., 2021; Schroff et al., 2015).

4. Experiments

We conduct three experiments to thoroughly verify our Modality Alignment method. Firstly, we show that the auxiliary CKA loss term boosts cosine similarity learning with a synthetic dataset. Secondly, we empirically examine that our method improves the cross modal attention on image captioning task with qualitative examples. Lastly, we manage the real-world experiment on Video QA task in which our method outperforms conventional baselines. All three experiments demonstrate that our Modality Alignment method enhances the cross modal attention module, consequently resulting higher performance of the multi-modal model. All experimental codes will be publicly available.

4.1. Cosine Similarity Learning with CKA

We empirically verify that optimizing CKA is helpful for cosine similarity learning. The attention mechanism learns cosine similarity between two corresponding source representation and target representation to be increased during training. However, because there are no ground truth attention weights in most real-world datasets, directly evaluating the performance of cross attention module is difficult. In order to verify that our method improves the performance of cross attention module, we conduct an experiment with a synthetic dataset in which a model is trained to maximize cosine similarity with one-to-one correspondence.

4.1.1. Experiment settings

We make a synthetic dataset which simulates two different modalities with completely different characteristics as following three steps.

- We sample 10,000 class ‘A’ examples from a multivariate normal distribution with dimension size of 64.
- We also sample 10,000 class ‘B’ examples from a intricately designed mixture of multivariate normal distribution with the same dimension size.
- To simulate ground truth hard attention, we randomly make one-to-one correspondences between each example of ‘A’ and ‘B’.
- The goal is to train two encoders for both ‘A’ and ‘B’ in the way that maximizes cosine similarity between two corresponding embedded vectors.

The main criterion for evaluation is the averaged cosine similarity between all corresponding examples of class ‘A’ and class ‘B’. Figure 3 describes the t-SNE visualization of our synthetic dataset.

Then, we build two neural networks models to substitute for embedding modules. Each neural networks takes samples of each class as input respectively and encodes them into output vectors. We regard the output of each networks as two different representations of different modalities. Both encoders have the same architecture but do not share the weights. Each encoder has three fully-connected layers with the hidden size of 32, each layer followed by the ReLU activation and Batch Normalization. The mini-batch size is set to 512 and λ_{CKA} value is 0.1. We train the model with ADAM optimizer with initial learning late of 0.001, $\beta_1 = 0.9$, and $\beta_2 = 0.999$.

We test three methods for comparison; (a) directly maximize only the averaged cosine similarity, (b) directly maximize CKA only, and (c) our Modality Alignment method that optimize both the criterion and CKA loss $\mathcal{L}_{cka}$. In the experiment with our method, we observe that pre-training CKA alone for few steps before optimizing the final loss $\mathcal{L}_{final}$ as a warm-up increases the performance. All experiments with our method in this paper are also performed this warm-up.

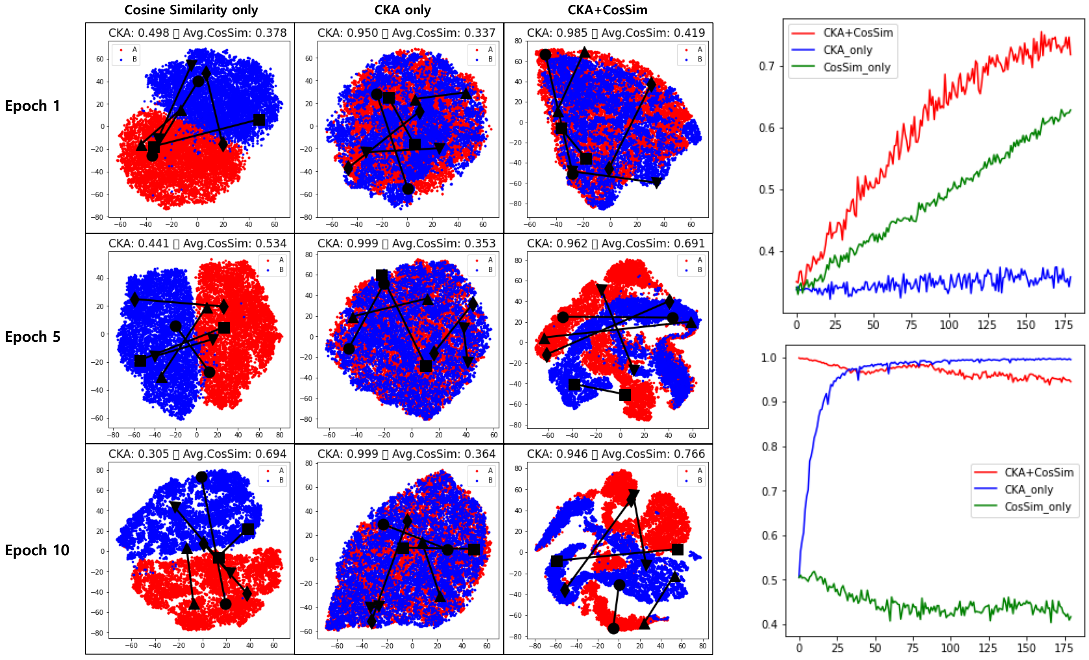

Figure 4: Cosine similarity learning results on the synthetic dataset. *(Left)* t-SNE visualizations of both encoded representations after 1, 5, and 10 epochs. *(Right top)* Training curve of the averaged cosine similarity over training steps. *(Right bottom)* Training curve of CKA between two representations over training steps.

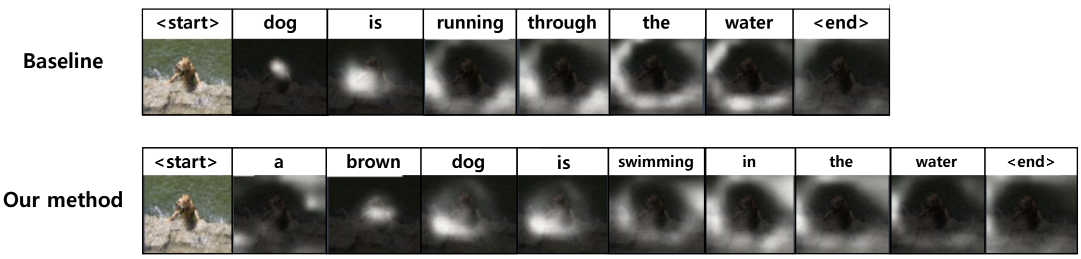

Figure 5: Attention map visualizations from Image captioning models. Our method produces correct captions with much richer expressions. Aligned modality using CKA enhances the cross-modal attention to capture wider meaningful area (e.g. *dog*), resulting in performance improvements.

4.1.2. Experiments Results

We summarize the results in Figure 4. The t-SNE visualizations of encoded representations over epochs reveals an interesting effect of our method (Left of Figure 4). Comparing the first column (trained with only cosine similarity loss) and the second column (trained with only CKA loss), only maximizing cosine similarity like existing multi-modal models does not make two representations similar as shown as CKA value drops from 0.498 to 0.305. In contrast, training with only CKA loss makes two encoders learn the inter-example structures extremely well. Also, it even increases the average cosine similarity slightly implying that there is indeed a correlation between inter-example structure similarity and cosine similarity. Finally, our method which optimizes both CKA loss and cosine similarity outperforms the conventional methods (*CosSim_only*), showing that our method can boost the training of cross attention module (Right top of Figure 4).

4.2. Multi-modal Scenario: Image Captioning

Secondly, we verify that CKA is effective in cross-modal attention by helping cosine similarity learning in multi-modal settings. We construct an image captioning experiment to verify the improvement of the cross model attention through an attention map. The baseline model used is the *Show, attend and tell* model (Xu et al., 2015) with the Flickr8k dataset. We conduct a qualitative analysis through the attention map to analyze that our proposed method improves the cross-modal attention. As shown in Figure 5, a baseline model that has not learned with CKA generates false phrases like *running through the water*. Otherwise, in the case of our model, it creates a caption that is correct and rich in expressiveness. The captioning performance measured in BLEU-4 with the test data also increases from 0.1560 to 0.1617. These results show that our method directly helps cross-modal attention, leading to improve the

Model	CKA(Vid$_{emb}$,QA$_{emb}$)	CKA(Sub$_{emb}$,QA$_{emb}$)	CKA(Cpt$_{emb}$,QA$_{emb}$)
	Multi-modality	Uni-modality(Text)	Uni-modality(Text)
TVQA$_{abc}$	0.3907	0.8798	-
TVQA$_{abc}$ + CKA	**0.7815**	0.8528	-
STAGE	0.2694	0.8999	-
STAGE + Caption	0.3998	0.8625	0.8741
STAGE + Caption + CKA	**0.6708**	0.8878	0.9215

Table 1: CKA between various modalities. In the case of uni-modality, the CKA value is initailly high, which means that the similarity between the representations is high, but the case of multi-modality is not. However, after CKA learning through our method, multi-modality also shows a high CKA value, increasing the similarity between the representations.

Model	QA Accuracy (%)
TVQA$_{abc}$	67.70
TVQA$_{abc}$ + CKA	**69.38**
STAGE (video only)	52.75
STAGE (sub only)	67.99
STAGE	70.31
STAGE + CKA	72.89
STAGE + CKA + Caption	**73.88**

Table 2: VideoQA results evaluated with QA accuracy.

multi-modal model's performance. We put additional details and results of the image captioning experiment in Appendix.

4.3. Real-World Scenario: Video QA

Lastly, we verify our Modality Alignment method in Video Question Answering tasks as real-world scenarios. Video QA is one of the most challenging among multi-modal tasks because there exhibits a great deal of differences between the video and text modalities, causing severe text bias problem. With two standard benchmarks, following experiments demonstrate that our Modality Alignment method also improves conventional models even in video QA tasks as our method closes the gap between two modalities.

4.3.1. Datasets and baseline models

We evaluate our approach on two benchmarks: TVQA (Lei et al., 2018) and TVQA+ (Lei et al., 2020a). TVQA is a large-scale video question answering dataset based on six popular TV shows: *The Big Bang Theory, How I Met Your Mother, Friends, Grey's Anatomy, House, Castle*. As a baseline model, we utilize TVQA$_{abc}$ which is proposed together with the TVQA benchmark. We apply CKA loss between the video embedding representation and the QA embedding representation of TVQA$_{abc}$ to apply our Modality Alignment.

TVQA+ is a subset of TVQA that only uses The Big Bang Theory clips yet contains additional bounding box annotation for visual region feature. The training, validation, and test-public set consist of 23,545, 3,017, and 2,821 questions, respectively. We utilize STAGE as a baseline, a model proposed in TVQA+ benchmark paper. In STAGE model, the input images are encoded with pretrained Faster R-CNN as a visual embedding module and the input texts are encoded with pretrained BERT encoder as a text embedding module. We compute and maximize CKA between video representaion and text representation before the cross-modal attention layer. We apply grid-searching method to find the value of λ_{CKA}. Additional details of the dataset and the models are provided in Appendix.

4.3.2. Experiments Results

We report the experimental results evaluated with QA accuracy in Table 2 and corresponding CKA values in Table 1.

In experiments on TVQA dataset, QA accuracy of the baseline (TVQA$_{abc}$) is 67.70%, while our Modality Alignment method increases the accuracy up to 69.38%. CKA value between video embedding representation and QA representation is also increased significantly from 0.3907 to 0.7815. We suppose that the trained alignment between two different modalities leads to the final performance improvement.

Similarly in experiments on TVQA+ dataset, comparing STAGE and STAGE+CKA in Table 2 shows a significant accuracy improvement from 70.31 to 72.89 with our Modality Alignment method. CKA value also shows a large increase from 0.2694 to 0.6708 in Table 1, indicating the inter-example similarity structures of image representation and text representation are well trained to be similar. Through these results, we conclude that training the representational alignment between multiple modalities improves a Video QA model by enhancing the cross attention module.

In addition to our Modality Alignment method, we exploit the generated caption in order to reduce the text bias by a video captioning model (Lei et al., 2020b). In Table 2, STAGE (video only) indicates the result of the model using only video features without subtitle information, and STAGE (sub) is the result of vice versa. The significant accuracy drop in STAGE (video only) implies that STAGE model is biased toward text modality as known as the text bias problem (Cadene et al., 2019). We generate additional captions with Multi Modal Transformer (MMT) model (Lei et al., 2020b).

The generated captions are passed to the model as additional text inputs. With our aligning method plus the generated captions, we achieve the best result as shown at the bottom of Table 2 showing an additional 0.99 accuracy improvement finally resulting 73.88 accuracy. We also examine the impact of our Modality Alignment method on embedding representation similarity of multiple modalities. As shown in Table 1, both CKA(Cpt$_{emb}$,QA$_{emb}$) and CKA(Sub$_{emb}$,QA$_{emb}$) are high because the subtitles, the generated captions and the QA pairs have the same text modality. However, CKA(Vid$_{emb}$,QA$_{emb}$) values are low without our method, indicating the different characteristics between two modalities. Applying our Modality Alignment, CKA(Vid$_{emb}$,QA$_{emb}$) values become high as CKA of a single modality. Thus, our aligning method closes the gap between the video modality and the text modality. In a nutshell, all three experiments verify that learning the representational alignment with CKA fits two different representations to have similar structures, enhances the cross attention module, and eventually leads to the performance improvement in Video-and-Language learning. In addition, our method can be easily applied to not only Video QA models but also any models for multi-modal tasks.

5. Conclusion

In multi-modal tasks such as Video QA, there is a difference in characteristics between the two modalities, which reduces the effectiveness of cross-modal attention. To address this, we propose Modality Alignment method that optimizes the similarity between two embedding representation structures of two different modalities. Specifically, we maximize the similarity between representations by directly exploiting CKA as a training objective. In our experiments, we verify that our Modality Alignment method boosts cosine similarity learning in a synthetic environment, which is the basis of the attention method, and further improves the performance of multi-modal models for real word tasks. In the future, we will test the proposed method on various multi-modal learning tasks including Video-and-Language learning in order to confirm that it improves state-of-the-art modality alignment strategies.

6. Acknowledgements

This work was supported by Samsung Electronics Co., Ltd. This work was also supported by the Automation and Systems Research Institute (ASRI), Seoul National University.

7. Bibliographical References

Bau, A., Belinkov, Y., Sajjad, H., Durrani, N., Dalvi, F., and Glass, J. (2018). Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

Cadene, R., Dancette, C., Cord, M., Parikh, D., et al. (2019). Rubi: Reducing unimodal biases for visual question answering. *Advances in neural information processing systems*, 32:841–852.

Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., and Zhuang, Y. (2020a). Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10800–10809.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020b). Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Cortes, C., Mohri, M., and Rostamizadeh, A. (2012). Algorithms for learning kernels based on centered alignment. *The Journal of Machine Learning Research*, 13(1):795–828.

Cristianini, N., Shawe-Taylor, J., Elisseeff, A., and Kandola, J. S. (2002). On kernel-target alignment. In *Advances in Neural Information Processing Systems*, pages 367–373.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664.

Jang, Y., Song, Y., Yu, Y., Kim, Y., and Kim, G. (2017). Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Kim, K.-M., Heo, M.-O., Choi, S.-H., and Zhang, B.-T. (2017). Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022.

Kim, J., Ma, M., Kim, K., Kim, S., and Yoo, C. D. (2019). Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8337–8346.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. (2019). Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR.

Kudugunta, S., Bapna, A., Caswell, I., and Firat, O. (2019). Investigating multilingual nmt representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575.

Lei, J., Yu, L., Bansal, M., and Berg, T. (2018). Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379.

Lei, J., Yu, L., Berg, T., and Bansal, M. (2020a). Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225.

Lei, J., Yu, L., Berg, T. L., and Bansal, M. (2020b). Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Li, Y., Yosinski, J., Clune, J., Lipson, H., Hopcroft, J. E., et al. (2015). Convergent learning: Do different neural networks learn the same representations? In *FE@ NIPS*, pages 196–212.

Li, L., Chen, Y.-C., Cheng, Y., Gan, Z., Yu, L., and Liu, J. (2020a). Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065.

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al. (2020b). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Liang, J., Jiang, L., Cao, L., Kalantidis, Y., Li, L.-J., and Hauptmann, A. G. (2019). Focal visual-text attention for memex question answering. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1893–1908.

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Maheswaranathan, N., Williams, A. H., Golub, M. D., Ganguli, S., and Sussillo, D. (2019). Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in neural information processing systems*, 2019:15629.

Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Morcos, A., Raghu, M., and Bengio, S. (2018). Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems*, 31.

Pan, X., Wang, M., Wu, L., and Li, L. (2021). Contrastive learning for many-to-many multilingual neural machine translation. *arXiv preprint arXiv:2105.09501*.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. (2017). Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783. IEEE.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., and Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, A., Luu, A. T., Foo, C.-S., Zhu, H., Tay, Y., and Chandrasekhar, V. (2019). Holistic multi-modal memory network for movie question answering. *IEEE Transactions on Image Processing*, 29:489–499.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Yang, A., Miech, A., Sivic, J., Laptev, I., and Schmid, C. (2021). Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697.

Ye, Y., Zhao, Z., Li, Y., Chen, L., Xiao, J., and Zhuang, Y. (2017). Video question answering via attribute-augmented attention network learning. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 829–832.

Ye, L., Rochan, M., Liu, Z., and Wang, Y. (2019). Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10502–10511.

Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. (2019). Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Yun, H., Kang, T., and Jung, K. (2021). Analyzing and controlling inter-head diversity in multi-head attention. *Applied Sciences*, 11(4):1548.

Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J. S., Cao, J., Farhadi, A., and Choi, Y. (2021). Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34.

Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., and Gao, J. (2021). Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Zhao, Z., Jiang, X., Cai, D., Xiao, J., He, X., and Pu, S. (2018). Multi-turn video question answering via multi-stream hierarchical attention context network. In *IJCAI*, volume 2018, page 27th.

8. Appendix

8.1. Additional details of image captioning experiments

For the image captioning experiment at Section 4.2, we use *Show, attend and tell* model (Xu et al., 2015) as baseline model with the Flickr8k dataset. The *Show, attend and tell* model is a representative example of utilizing cross modal attention as an encoder-decoder structure. Encoder uses the pre-trained RESNET-101 model, encodes to 14x14 pixels with 2048 dimensions, and then enters it as decoder input with the golden caption labels. Decoder uses the LSTM model with teacher forcing. In this case, we apply CKA loss between the text embedding processed by the decoder and the image embedding, which is an output of the encoder.

In the case of an experiment on verifying the boosting effect of cross modal attention through attention map, both the baseline model and our model use the model in epoch 5, and the measured BLUE-4 score is the result of the experiment in the test set. The model compared to the baseline model through the attention map used in the main text uses a λ_{CKA} of 0.05.

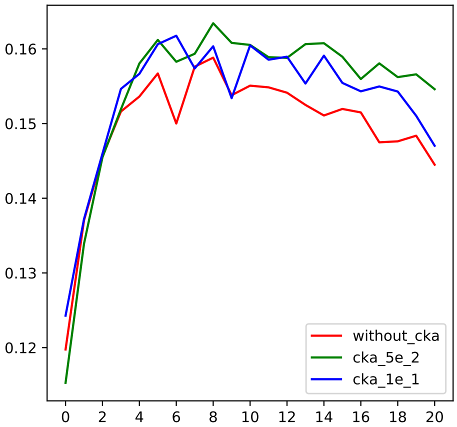

Figure 6: BLEU-4 on validation set for the image captioning experiment.

8.2. Image Captioning performance on validation set

We verify the boosting effects of cross modal attention through attention map in the main text of our paper. Moreover, as similar to the result of experiment 4.1, the model using our method in the validation set learns faster and performs better. The validation performance can be seen in figure 6. The figure 6 shows the BLEU-4 score according to the epoch for the validation set. We set the λ_{CKA} as 0.05 and 0.1, and our models show faster and better learning performance in the validation set than baseline.

8.3. Additional details of Video QA experiments

For TVQA$_{abc}$, the pretrained Faster R-CNN and LSTM are used as visual embedding modules, and word embedding with LSTM are used as text embedding modules. And then CKA loss is configured before context matching information where cross modal attention takes place as in figure 7. The details of implementations are the same as the baseline model, λ_{CKA} value is 0.2, and the batch size is 32.

In the case of STAGE, conv encoder with the pretrained Faster R-CNN is used as visual embedding modules, and conv encoder with the BERT embedding is used as text embedding modules. Thereafter, CKA loss is configured before the two presentations cross-modal attention is performed. The details of implementations are the same as the baseline model, λ_{CKA} value is 0.1, and the batch size is 4. We used a dense video captioning model MMT (Lei et al., 2020b) to solve text bias of the baseline model, to create captions from video information and use them as additional information. It can be seen that capture information is added as input stream in the right part of the figure 8.

In both cases, linear kernel is used for computing CKA, based on the finding of a study by Kornblith et al. (2019) that there is no significant difference from other kernels such as RBF kernel.

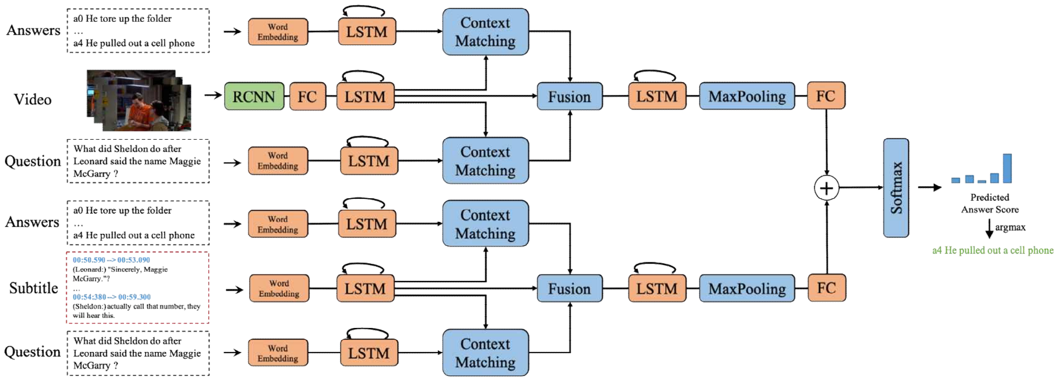
Answers
a0 He tore up the folder
...
a4 He pulled out a cell phone
Word Embedding
LSTM
Context Matching
Video
RCNN
FC
Fusion
MaxPooling
Question
What did Sheldon do after Leonard said the name Maggie McGarry ?
Subtitle
00:50.590 --> 00:53.090
(Leonard:) "Sincerely, Maggie McGarry."?
...
00:54:380 --> 00:59.300
(Sheldon:)actually call that number, they will hear this.
Softmax
Predicted Answer Score
argmax
a4 He pulled out a cell phone

Figure 7: TVQA$_{abc}$ model. We utilize auxiliary CKA Loss between both modalities before the context matching module.

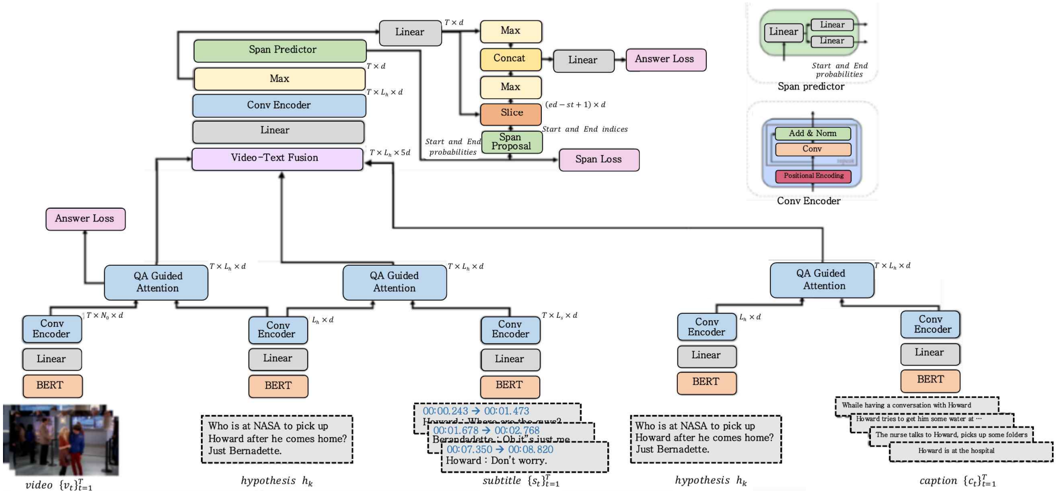

Figure 8: STAGE model with dense video captions. We utilize auxiliary CKA Loss between both modalities before the QA Guided Attention. We used a dense video captioning model MMT to solve text bias of the baseline model, to create captions from video information and use them as additional information.