Machine Intelligence lab
SEOUL NATIONAL UNIVERSITY

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

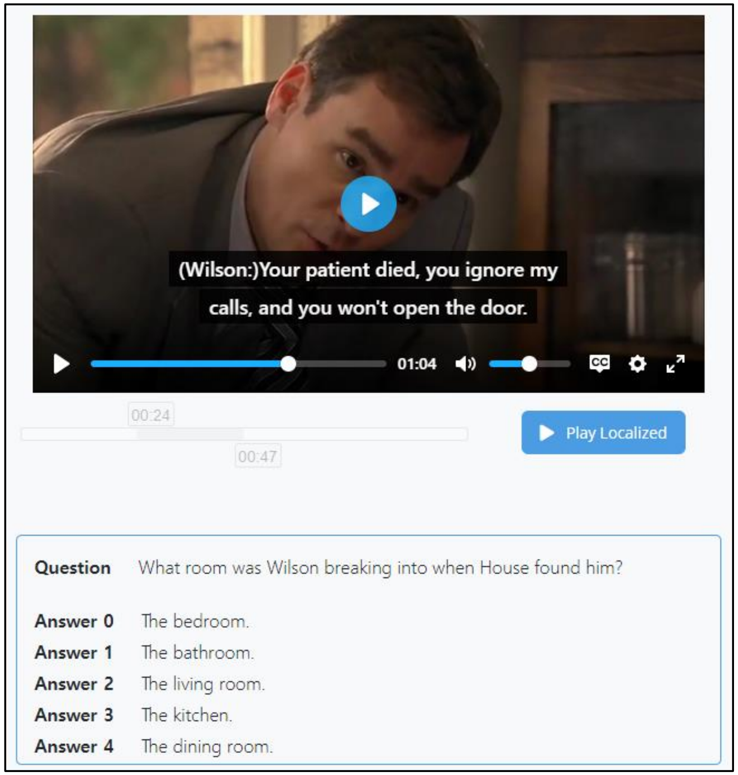
HYEONGU YUN
YONGIL KIM
KYOMIN JUNG

SEOUL NATIONAL UNIVERSITY
Department of Electrical and Computer Engineering

Multi-modality Tasks: Video-QA

- Video-Question Answering:
  - Given dataset {Video clip, Description, Query, Answer candidates},
  - Choose the most appropriate answer among the candidates.

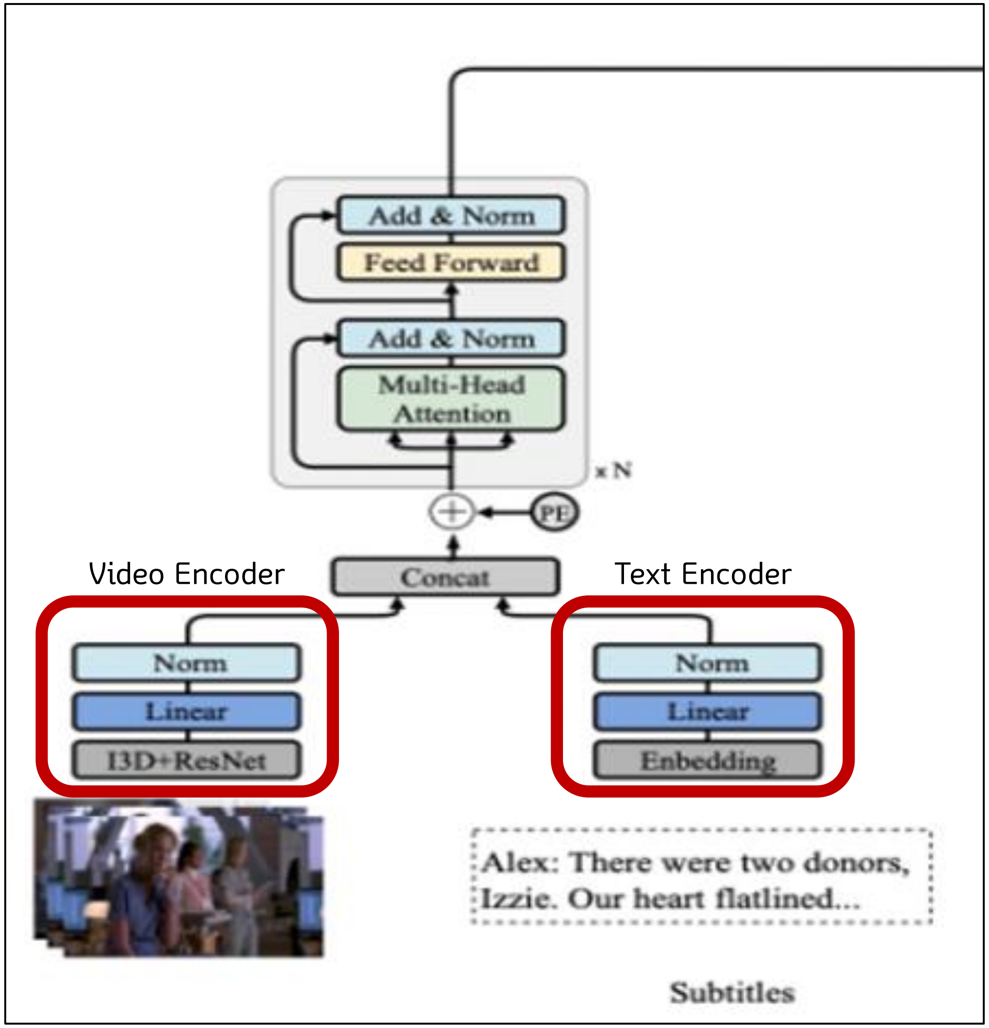
- Common Benchmark: TVQA*



*Lei et al., Tvqa: Localized, compositional video question answering., EMNLP 2018

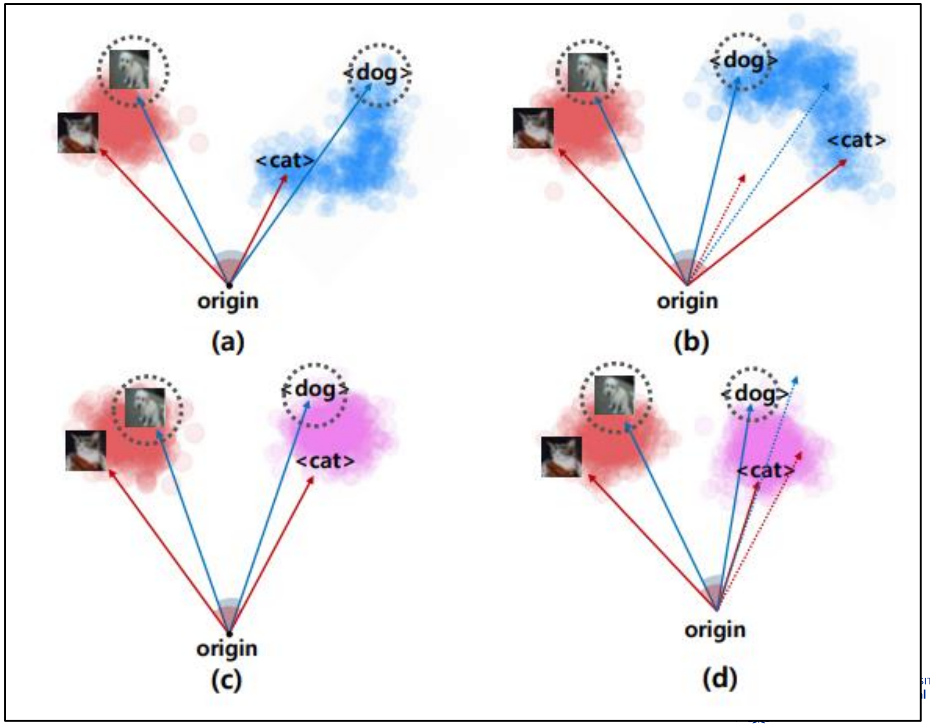
Multi-modal Transformers

- Common structures to handle Video-and-Language modalities.



Attention Mechanism for Multi-modality

- Potential hazard of the previous cross-modality attention:
  - There is a possible difference between the “structure” of Video representation vectors and the “structure” of Text representation vectors.
  - This difference may cause a side-effect of the attention mechanism, which is based on the cosine similarity.



Centered Kernel Alignment

- Centered Kernel Alignment (CKA)*:
  - A similarity measure between deep neural networks.
  - A method to compare the inter-example similarity structures.

- Pros. of CKA:
  - Robustness: CKA can measure similarity between two representational spaces with a small amount of data.
  - ➔ We can apply CKA in a mini-batch.
  - Differentiability: CKA is computed by simple differentiable equations.
  - ➔ We can optimize CKA by common frameworks with gradient descent.

*Kornblith et al., Similarity of neural network representations revisited., ICML 2019

Centered Kernel Alignment

- Centered Kernel Alignment (CKA)*:
  - **Similarity between the inter-example similarity structures.**

$$\langle \mathrm{vec}(XX^{\mathrm{T}}), \mathrm{vec}(YY^{\mathrm{T}}) \rangle = \mathrm{tr}(XX^{\mathrm{T}}YY^{\mathrm{T}}) = ||Y^{\mathrm{T}}X||_{\mathrm{F}}^2.$$

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

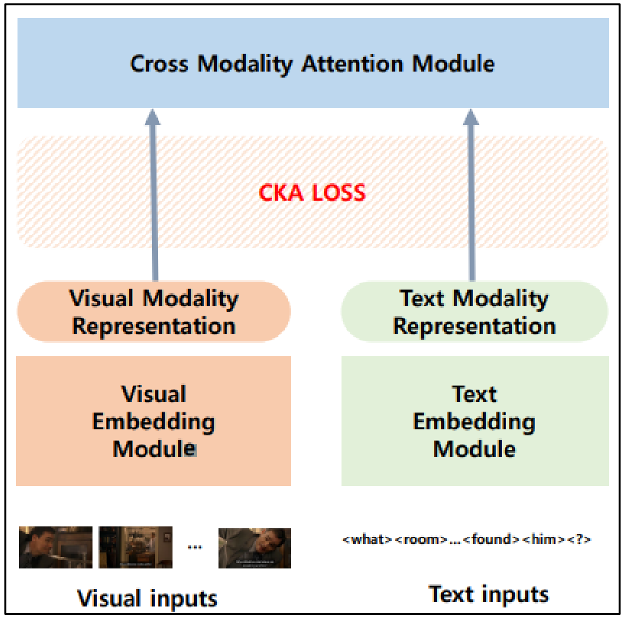
$$\mathrm{HSIC}(K_i, K_j) = \frac{1}{(N-1)^2} tr(K_i C K_j C),$$

$$\mathrm{CKA}(K_i, K_j) = \frac{\mathrm{HSIC}(K_i, K_j)}{\sqrt{\mathrm{HSIC}(K_i, K_i)\mathrm{HSIC}(K_j, K_j)}}.$$

*Kornblith et al., Similarity of neural network representations revisited., ICML 2019

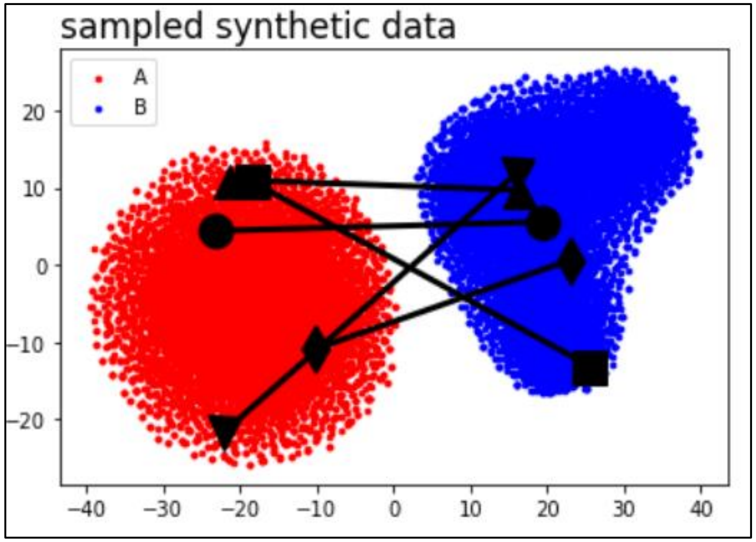
Our Modality Alignment method

- Add *CKA_loss* to the final loss (like a regularization term)
  - Let f_vid be a video encoder and f_text be a text encoder;
  - Then, with a sequence of video frames V=[v_1,⋯,v_L] and a sequence of tokens T=[t_1,⋯,t_M], calculate $\mathcal{L}_{CKA} = CKA(f_{vid}(V), f_{text}(T))$.
  - Add $-\lambda_{cka} * \mathcal{L}_{CKA}$ to the final loss.



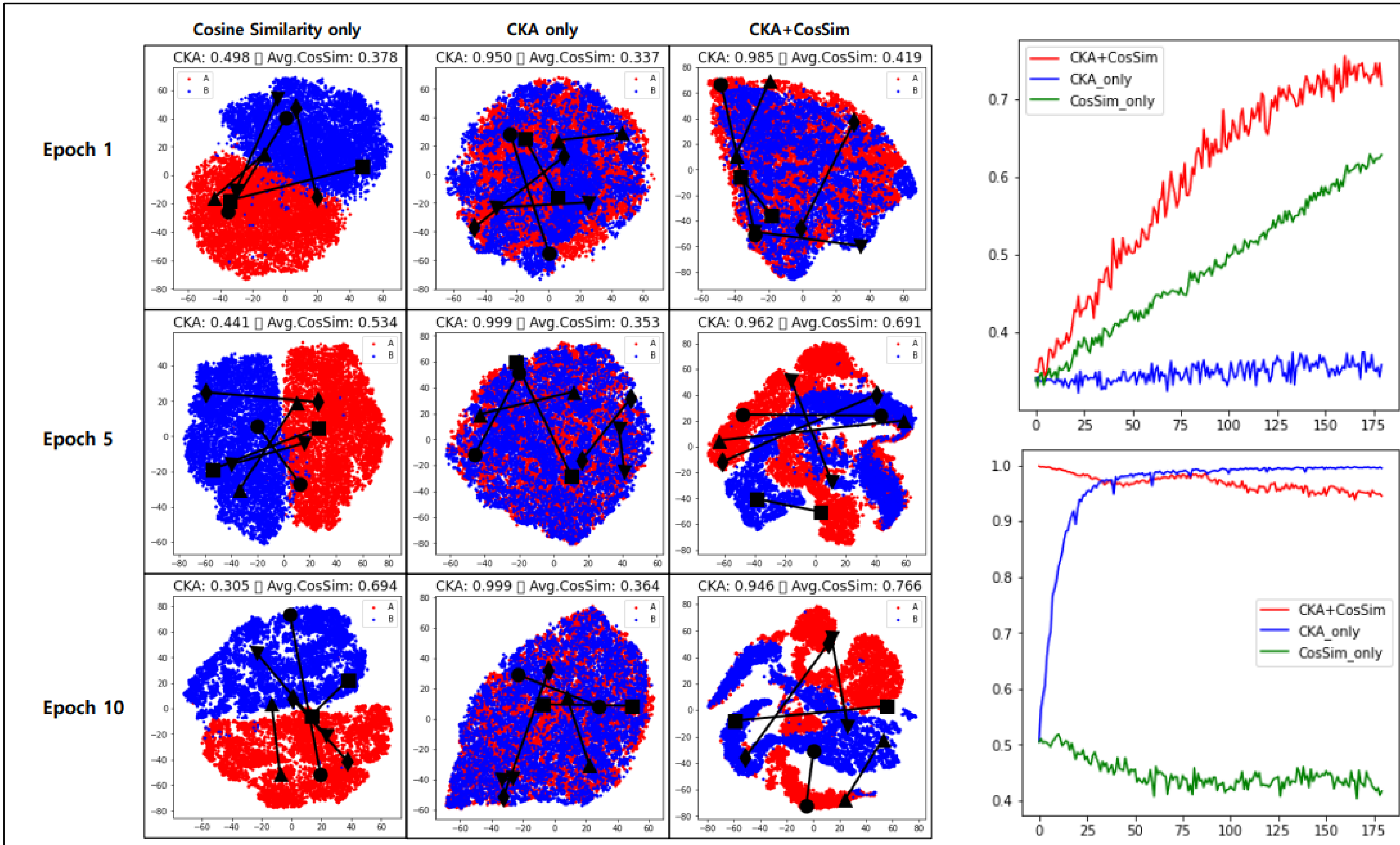
Applying CKA with Gradient Ascent: Synthetic dataset

- Synthetic dataset (maximizing cosine similarity):
  - Examples of class 'A' are sampled from a multivariate normal dist.
  - Examples of class 'B' are sampled from a mixture of multivariate normal dist.
  - To simulate the "attention" golden-truth, we randomly assign 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 the cosine similarity.



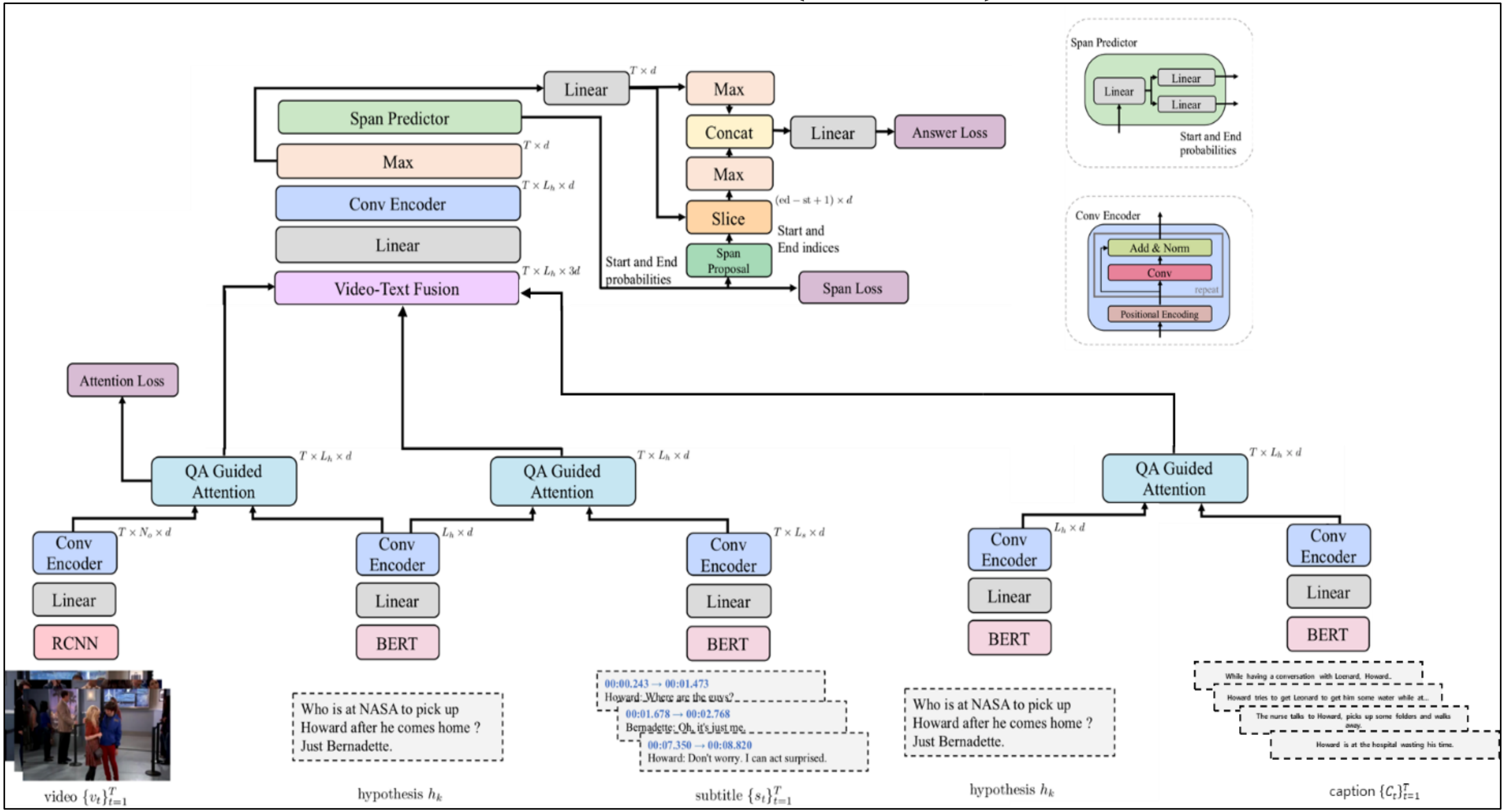
Applying CKA with Gradient Ascent: Synthetic dataset

- Synthetic dataset (maximizing cosine similarity):



Applying CKA with Gradient Ascent: Real-world dataset (TVQA+)

•A baseline structure for Video-QA (STAGE*):



*Lei et al., Tvqa: Localized, compositional video question answering., EMNLP 2018

Applying CKA with Gradient Ascent: Real-world dataset (TVQA+)

- Are there really the differences between Video representation and Text representation? *Yes.*

Model	CKA(Vid_{emb},QA_{emb}) Multi-modality	CKA(Sub_{emb},QA_{emb}) Uni-modality(Text)	CKA(Cpt_{emb},QA_{emb}) 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

- Does our CKA optimization improve the final accuracy? *Yes.*

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.

Summary

- We show that CKA can align two embedding representations from different modalities.
- We demonstrate that our Modality Alignment improves the performance in multi-modal tasks.

Thank You!

Hyeongu Yun
youaredead@snu.ac.kr

Yongil Kim
miles94@snu.ac.kr

Kyomin Jung
kjung@snu.ac.kr