SEJONG UNIVERSITY

PHAN QUANG THINH - ICL

PAPER REVIEW

T2VLAD: GLOBAL-LOCAL SEQUENCE ALIGNMENT FOR TEXT-VIDEO RETRIEVAL

CONTENT

I. OVERVIEW

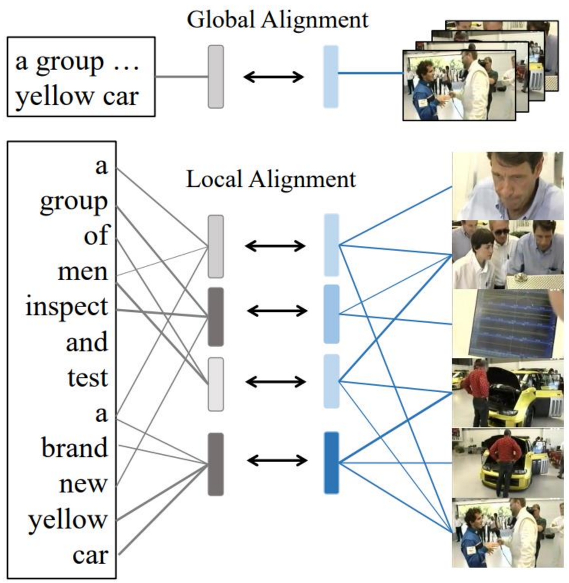
- X. Wang, L. Zhu, and Y. Yang, "T2VLAD: global-local sequence alignment for text-video retrieval,", In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.
- **"T2VLAD"** – **Text-to-Video VLAD**

VLAD: Vector of Locally Aggregated Descriptors (slide 6)

- **T2VLAD –** text-video retrieval model, which aligns text and video features in a global and local perspective:

- Local alignment: the fine-grained comparisons by computing the similarities between the local text-video features in sematic topics.

- Global alignment: encoding text and video content and comparing their similarities in the global perspective



X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

II. PROBLEM STATEMENT

- Most existing methods: encode the descriptions and video content to global representations and compare their similarities from a **global perspective;**
- Some other works: leveraged complex cross-modal matching operations to exploit the **local details** and align multiple semantic cues.
- **HGR** [1] *(CVPR 2020)* proposed a hierarchical graph reasoning model to capture both **global events and local actions** through local graph matching

=> require a **high computational cost** due to the expensive pairwise matching operation.

Global perspective

Local perspective

Global + Local perspective

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

III. RELATED MODELS (PAPERS)

Text representation: BERT (NAACL-HLT 2019)

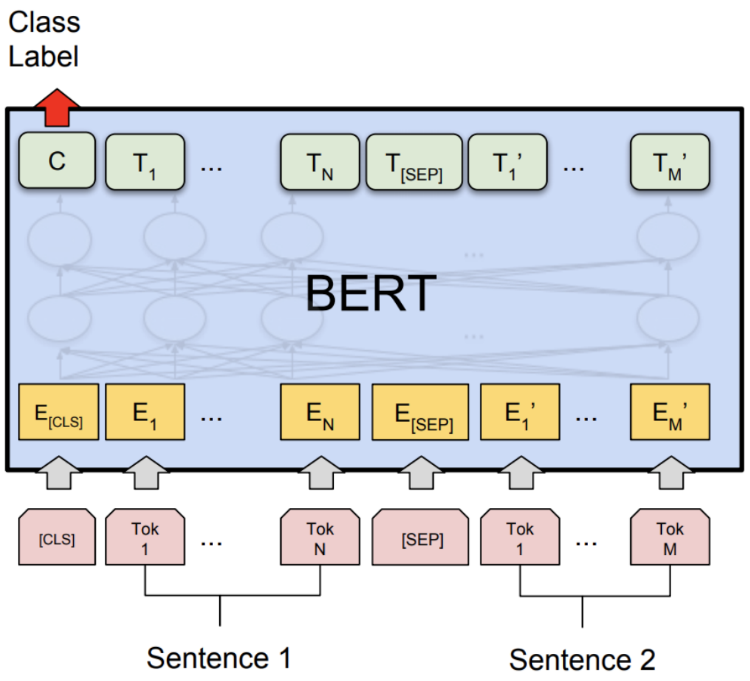
Video embedding in MME (2018)

Sematic topic alignment: NetVLAD (CVPR 2016)

1. BERT

- Bidirectional Encoder Representations from Transformers (BERT) [2] *(NAACL-HLT 2019)*
- BERT - *transformer-based* machine learning technique learns contextual relations between words (or sub-words) in a text.
- State of the art language model results in a wide variety of Natural Language Processing (NLP) tasks (question answering, text classification, ...)

=> Text representations



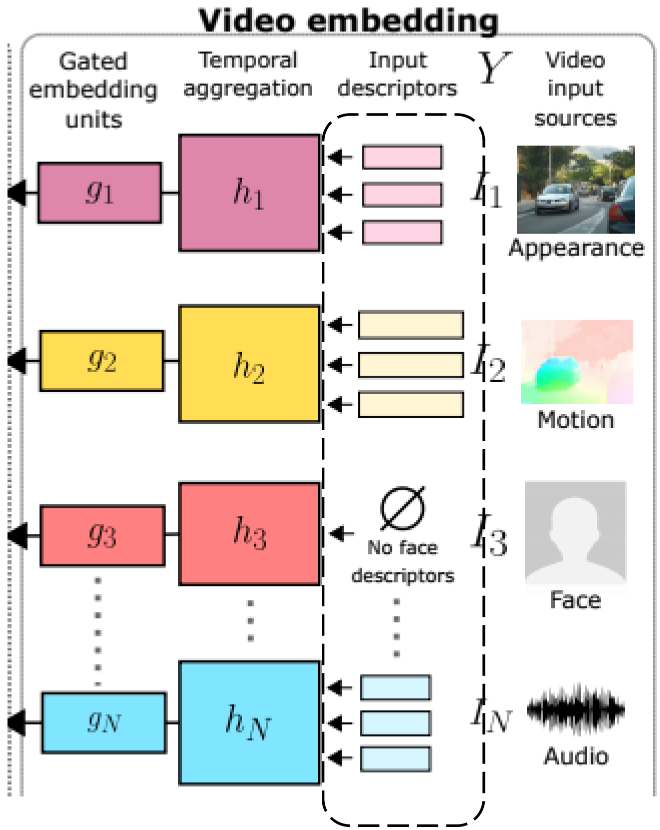
Example of BERT in language pair classification

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

2. MEE

- MEE [3] Mixture-of-Embedding-Experts: computes similarities between text and a varying number of video modalities
- "Expert" video feature extractor - an effective representation from different modalities inherent in video data (appearance, motion, audio,...)
- Video embedding in MME: take advantage of the rich and varied additional information present in videos: motion dynamics, speech and other background sounds.

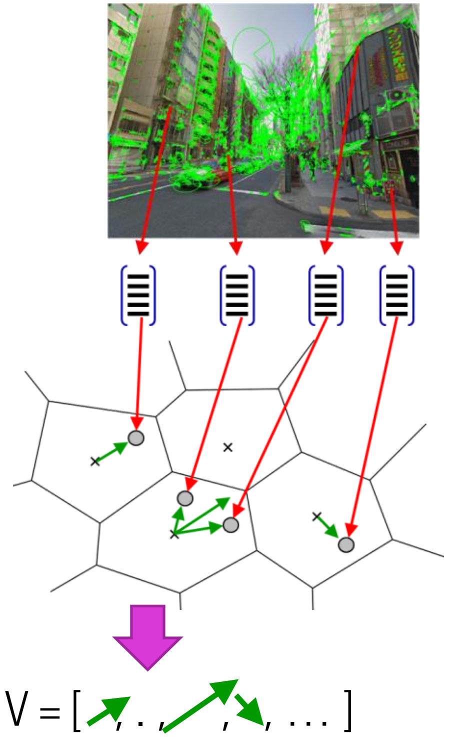
=> Use Video embedding in MME for Video representations



Expert Encoding

3. NETVLAD

- Vector of Locally Aggregated Descriptors (VLAD) [2] *(CVPR 2010)* is an image representation model that commonly used in image retrieval
  - Accumulate the residual of each descriptor with respect to its assigned cluster (K-mean clustering)
  - Store the sum of the differences of the descriptors assigned to the cluster and the centroid of the cluster
  - VLAD is pooling method, not a CNN architecture -> not trainable
- **NetVLAD** [3] *(CVPR 2016)* – a powerful image representation trainable end-to-end on the image retrieval - mimic VLAD in a CNN framework and design a trainable generalized VLAD layer.



X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

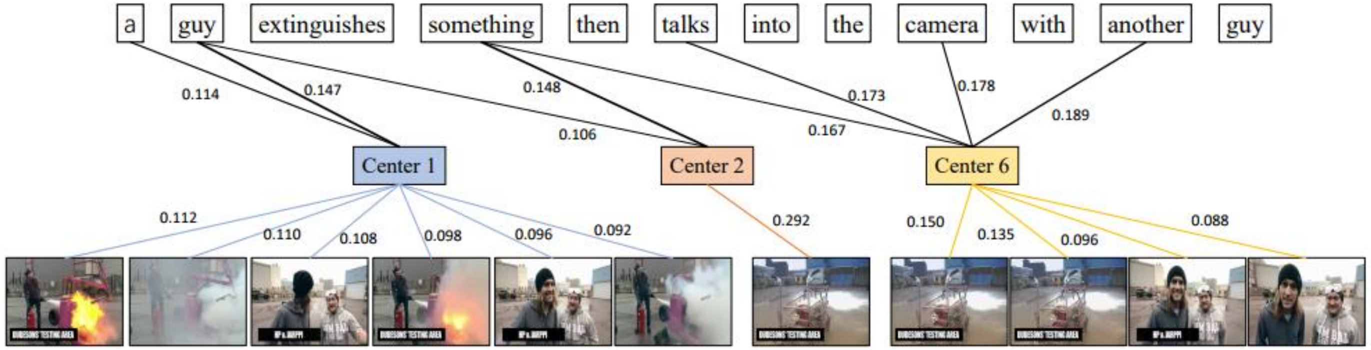
3. NETVLAD (CONT.)

=> **Sematic topic alignment** by NetVLAD: for **both text and video modalities** (can be readily utilized as latent semantic topics on cross-modal)

- NetVLAD operations can obtain an aggregated feature for each topic, where the **topic centers** are **shared** between the two modalities.

(a) Mobile phone query (b) Retrieved image of same place

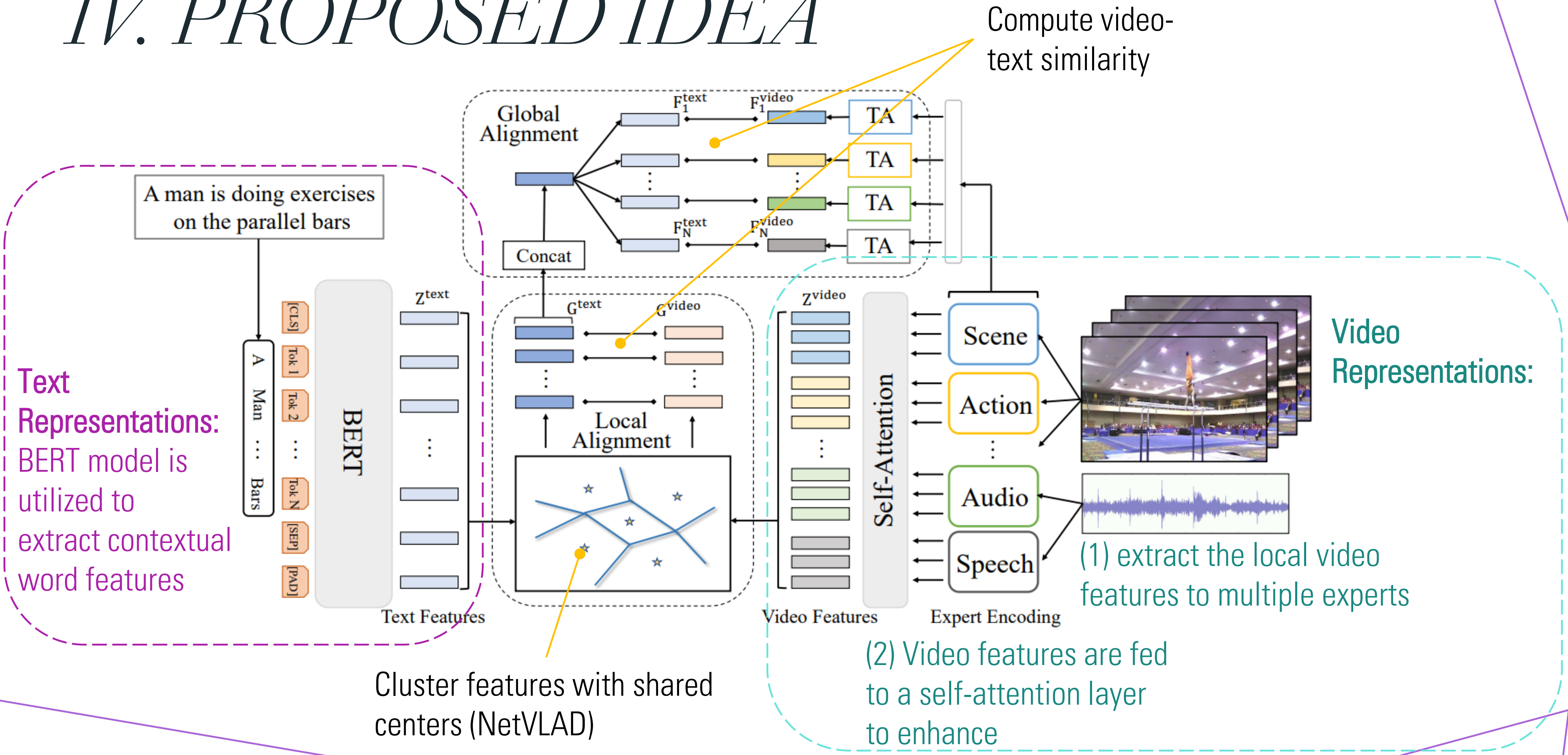
Example of NetVLAD
(in image retrieval task)



Example of shared topic center

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

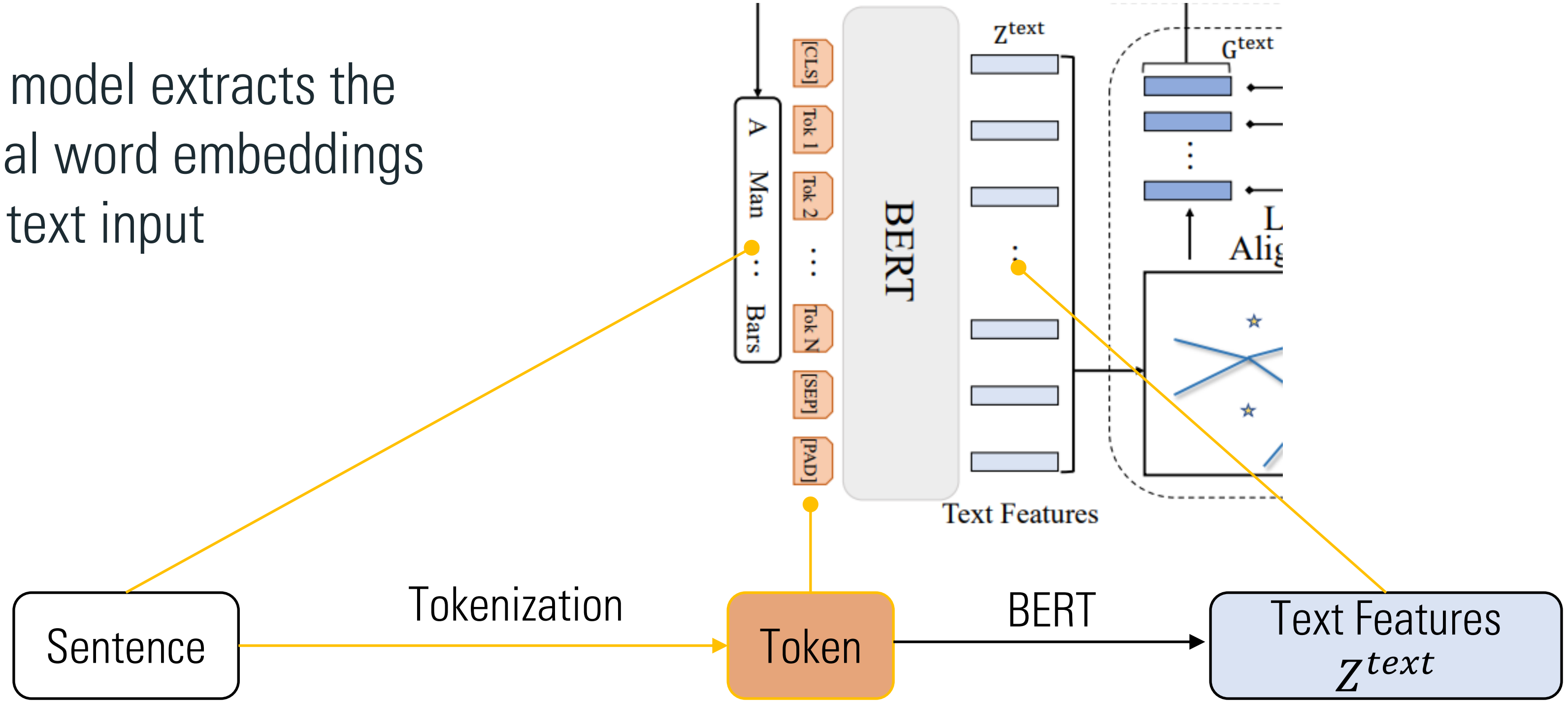
IV. PROPOSED IDEA



X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

1. TEXT REPRESENTATION

- BERT [2] model extracts the contextual word embeddings for each text input



[CLS]: class token, marked for begin a sequence
[SPS]: separator token, added among 2 sentences in a sequence

$Z^{text} = \{z_1^{text}, z_2^{text}, \ldots, z_B^{text}\}$
Where B is the sequence length

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

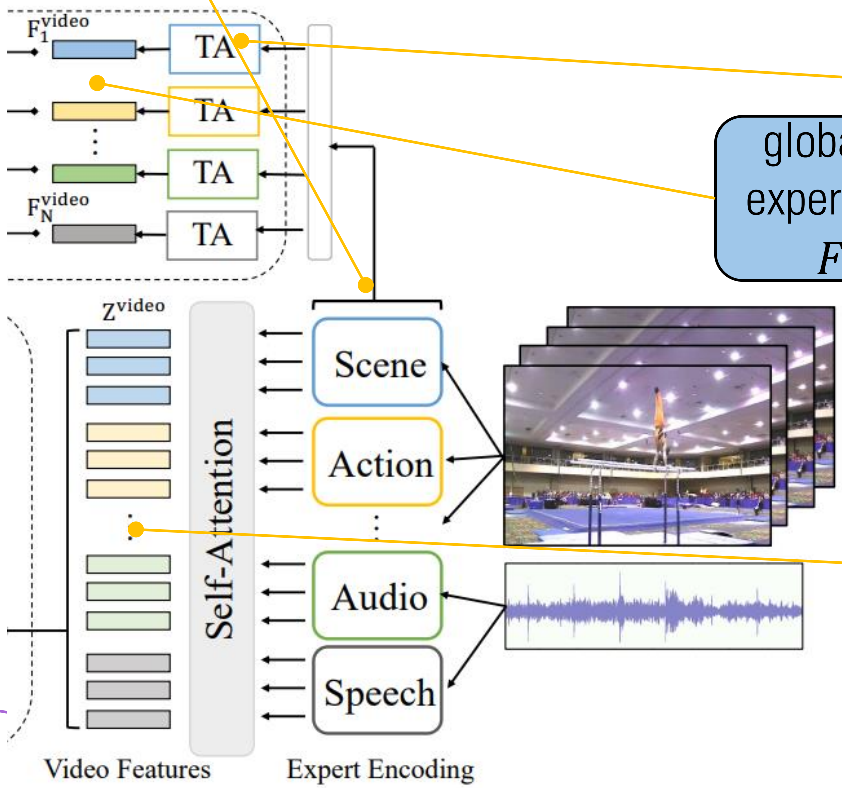
2. VIDEO REPRESENTATION

x_t is the t-th segment from a video

Segment-level video representation $E^n(x_t)$

Encode by expert E^n

Raw video



- Global expert features for global alignment:

global video expert feature F_i^{text}

Self-gating

Temporal-aggregated feature

Max-pooling

Segment-level video representation $E^n(x_t)$

- Fuse multi expert features for local alignment:

Video local features Z^{video}

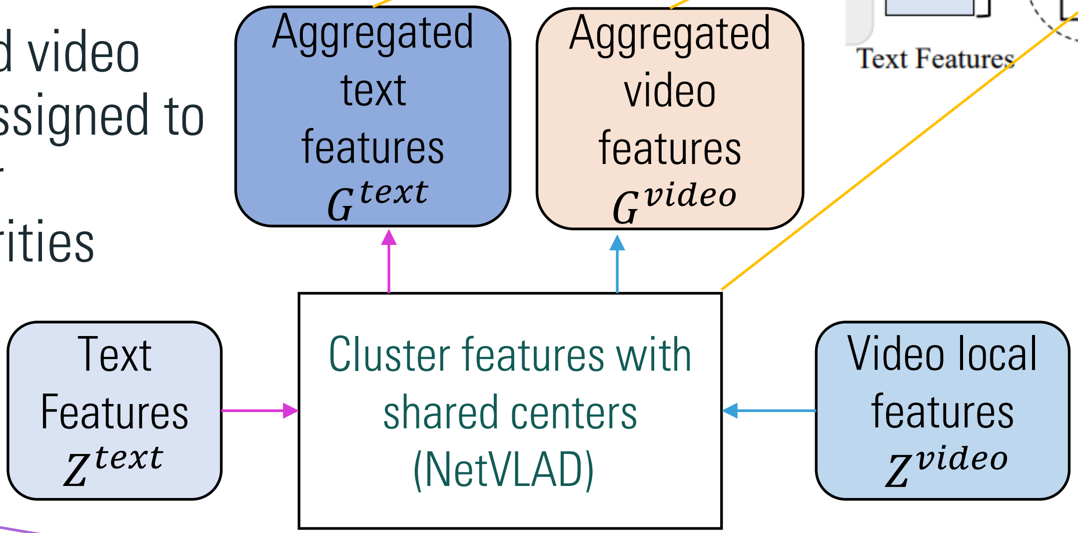
Self-attention layer

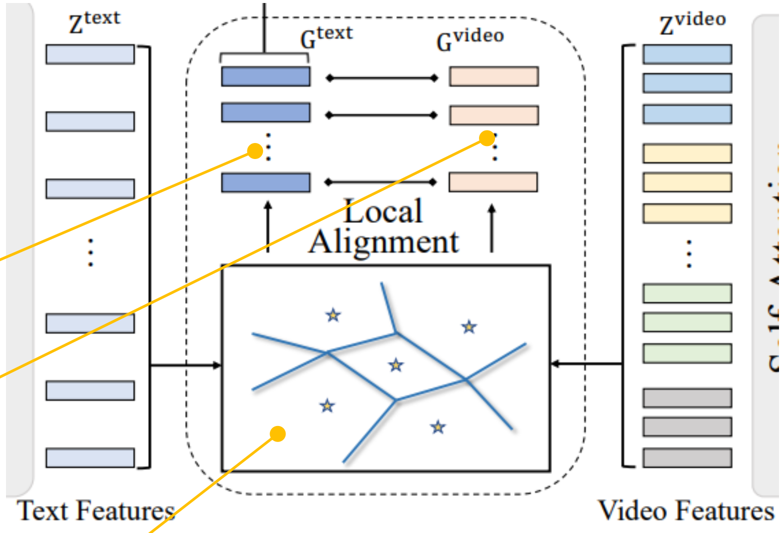
Segment-level video representation $E^n(x_t)$

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

3. LOCAL ALIGNMENT

- Use NetVLAD operations for both text and video modalities, to obtain an **aggregated feature** for each topic, where the topic centers are **shared** between the two modalities.
- The text features and video features are softly assigned to topics based on their corresponded similarities





X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

3. LOCAL ALIGNMENT (CONT.)

The aggregated features g_j^{video}, g_j^{text} are calculated by using the shared cluster centers c_j [3]

$$g_j^{video} = normalize(\sum_{i=1}^{M} \frac{\exp(z_i^{video} c_j^T + b_j)}{\sum_{k=1}^{K+1} \exp(z_i^{video} c_k^T + b_k)} (z_i^{video} - c_j'))$$

$$g_j^{text} = normalize(\sum_{i=1}^{B} \frac{\exp(z_i^{text} c_j^T + b_j)}{\sum_{k=1}^{K+1} \exp(z_i^{text} c_k^T + b_k)} (z_i^{text} - c_j'))$$

Where:

M: the number of features from all experts (video)

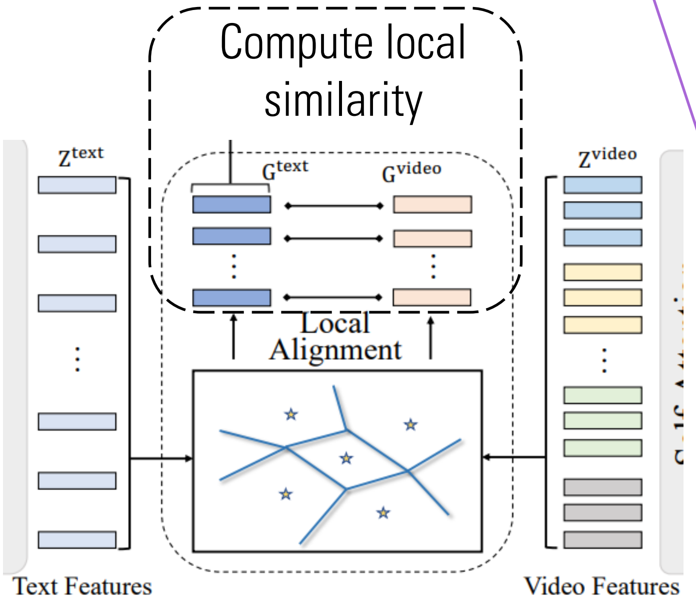
B: sequence length of text feature

$c_1, c_2, \dots c_{K+1}$ shared cluster center of 1 to K+1

b_j is a learnable bias term

z_i^{video}, z_i^{text} are local video feature / local word embedding

c_j' trainable weights



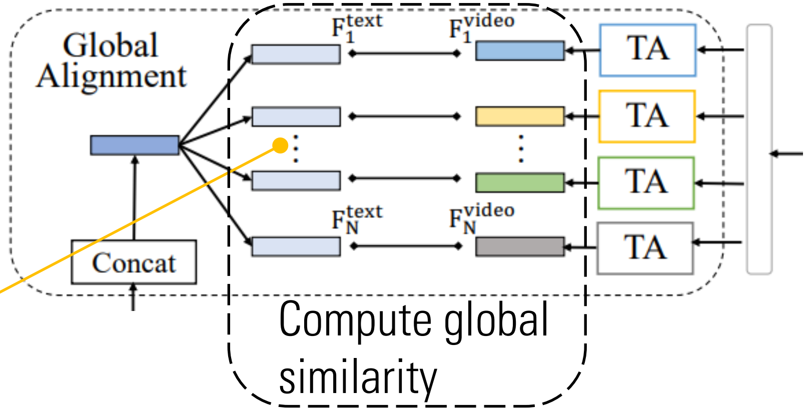
Local similarity = cosine distance between video feature $G^{video} = \{g_1^{video}, \dots, g_K^{video}\}$ and text feature $G^{text} = \{g_1^{text}, \dots, g_K^{text}\}$

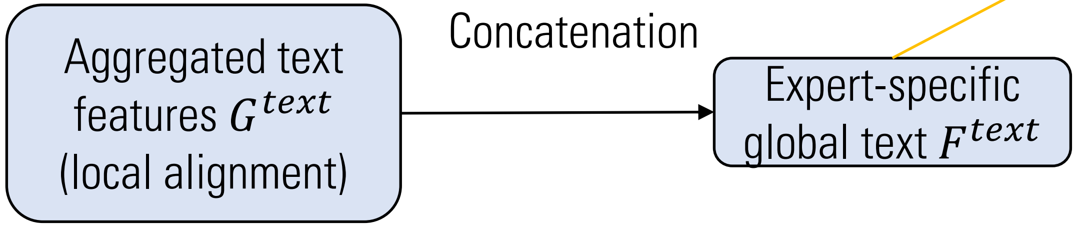
$$s_{local} = dist(G^{video}, G^{text})$$

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

4. GLOBAL ALIGNMENT

- 2 reasons of using global alignment:
  - More comprehensive and complementary to local features
  - Lacking auxiliary supervision of local alignment with trainable centers





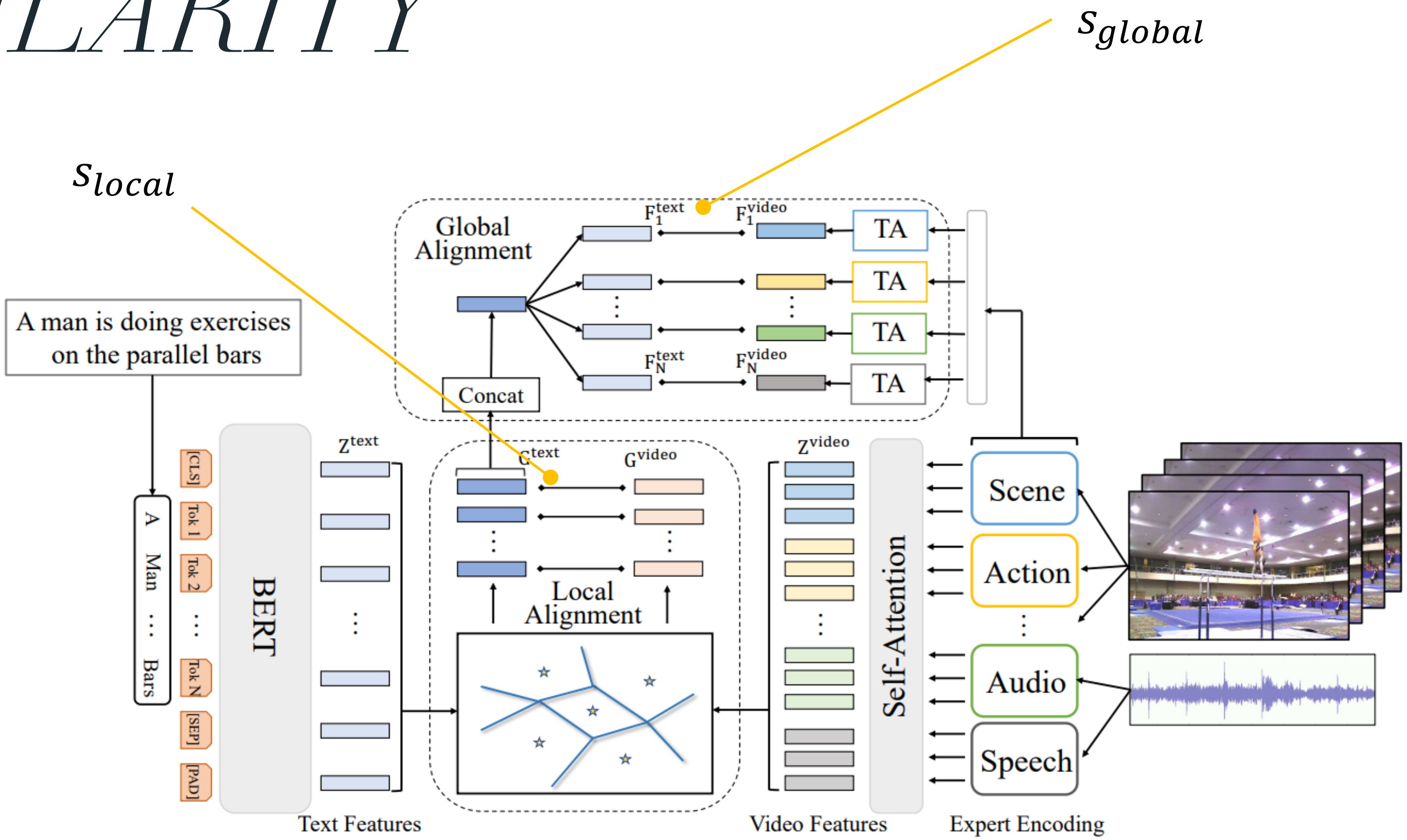
Global similarity: weighted sum of cosine distances between each global video expert feature F_i^{text} and corresponding text feature F_i^{video}

$$s_{global} = \sum_{i=1}^{N} w_i * dist(F_i^{text}, F_i^{video})$$

(w_i is the weight for the i-th expert, generated from G^{text} by a linear projection with a soft-max normalization)

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

SIMILARITY



Text-video similarity s:

$$s = \frac{1}{2}(s_{global} + s_{local})$$

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

V. RESULT

- T2VLAD outperforms **MMT** (Multi-modal Transformer for Video Retrieval) [5] *(ECCV 2020)* and other proposed methods with MRS-VTT dataset

Method	Split	Text → Video				Video → Text			
		R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓
JSFusion [37]	1k-A	10.2	31.2	43.2	13	-	-	-	-
HT [25]	1k-A	14.9	40.2	52.8	9	-	-	-	-
CE [22]	1k-A	20.9	48.8	62.4	6	20.6	50.3	64.0	5.3
MMT [9]	1k-A	24.6	54.0	67.1	4	24.4	56.0	67.8	4
MMT + HT pretrain [9]	1k-A	26.6	57.1	69.6	4	27.0	57.5	69.7	3.7
Our T2VLAD	1k-A	**29.5**	**59.0**	**70.1**	**4**	**31.8**	**60.0**	**71.1**	**3**
MEE [24]	1k-B	13.6	37.9	51.0	10	-	-	-	-
JPose [31]	1k-B	14.3	38.1	53.0	9	16.4	41.3	54.4	8.7
MEE-COCO [24]	1k-B	14.2	39.2	53.8	9	-	-	-	-
CE [22]	1k-B	18.2	46.0	60.7	7	18.0	46.0	60.3	6.5
MMT [9]	1k-B	20.3	49.1	63.9	6	21.1	49.4	63.2	6
Our T2VLAD	1k-B	**26.1**	**54.7**	**68.1**	**4**	**26.7**	**56.1**	**70.4**	**4**

- T2VLAD outperforms MMT and other proposed methods with ActivityNet dataset

Method	Text → Video				Video → Text			
	R@1 ↑	R@5 ↑	R@50 ↑	MdR ↓	R@1 ↑	R@5 ↑	R@50 ↑	MdR ↓
FSE [39]	18.2	44.8	89.1	7	16.7	43.1	88.4	7
CE [22]	18.2	47.7	91.4	6	17.7	46.6	90.9	6
HSE [39]	20.5	49.3	-	-	18.7	48.1	-	-
MMT [9]	22.7	54.2	93.2	5	22.9	54.8	93.1	4.3
Ours	**23.7**	**55.5**	**93.5**	**4**	**24.1**	**56.6**	**94.1**	**4**

Table 2. The comparisons with the state-of-the-art methods on the ActivityNet Captions dataset.

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

V. RESULT (CONT.)

- Comparison with papers of CVPR/CVPRW 2021

(Text -> Video Retrieval with **ActivityNet** dataset)

Method	Model	R@1	R@5	R@50	MdR
M. Dzabraev, el al. "MDMMT: Multidomain Multimodal Transformer for Video Retrieval," *IEEE/CVF CVPRW, Jun. 2021*, pp. 3354-3363.	MDMMT	17.7	41.6	-	8.3
L. Jie, et al, "Less is more: ClipBERT for video-and-language learning via sparse sampling," In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 7331-7341.	ClipBERT	21.3	49.0	-	6.0
X. Wang, el al. "T2VLAD: global-local sequence alignment for text-video retrieval," In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.	T2VLAD	**23.7**	**55.5**	93.5	**4.0**

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

VI. ABLATION STUDY

- The effectiveness of the the global-local alignment
  - Global-local alignment proves local alignment (**global feature is complementary to the local information**)

Method	Text → Video				Video → Text			
	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓
Ours w/o Global Alignment	24.3	51.5	63.4	5	26.6	52.9	62.6	5
Ours w/o Local Alignment	22.2	49.9	64.6	6	24.0	51.7	65.6	5
Full model	**29.5**	**59.0**	**70.1**	**4**	**31.8**	**60.0**	**71.1**	**3**

Table 4. The ablation studies on the MSRVTT [35] dataset to investigate the effectiveness of global-local alignment.

- The effectiveness of collaborative VLAD
  - **The sharing centers (shared VLAD) outperform separated VLAD**

Method	Text → Video				Video → Text			
	R@1↑	R@5↑	R@10↑	MdR↓	R@1↑	R@5↑	R@10↑	MdR↓
Ours w/ only text VLAD	27.4	57.3	68.2	4	27.5	57.4	69.7	4
Ours w/ two separate VLAD	28.6	58.1	70.4	4	30.4	60.7	72.1	3
Ours w/ two shared VLAD	**29.5**	**59.0**	70.1	**4**	**31.8**	60.0	71.1	**3**

Table 5. The ablation studies on the MSRVTT [35] dataset to investigate the effectiveness of the VLAD encoding.

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

VII. CONCLUSION

- Contributions:
    - Propose **automatically learn** text-and-video semantic topics
    - Re-emphasize the importance of **local semantic alignment between texts and videos** for better cross-modal retrieval
    - T2VLAD – exploit **shared centers** (NetVLAD) to reduce the sematic gap between texts and videos
    - T2VLAD outperforms recently proposed method
- Local semantic alignment between texts and videos is critical for high-performance
- Future work: obtain better global video features with end-to-end optimization.

X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

REFERENCES

[1] S. Chen, Y. Zhao, Q. Jin, and Q. Wu. “Fine-grained video-text retrieval with hierarchical graph reasoning.” In CVPR, 2020.

[2] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova,“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” In *NAACL-HLT*, 2019.

[3] A. Miech, I. Laptev, and J. Sivic. “Learning a text-video embedding from incomplete and heterogeneous data.” arXiv preprint arXiv:1804.02516, 2018

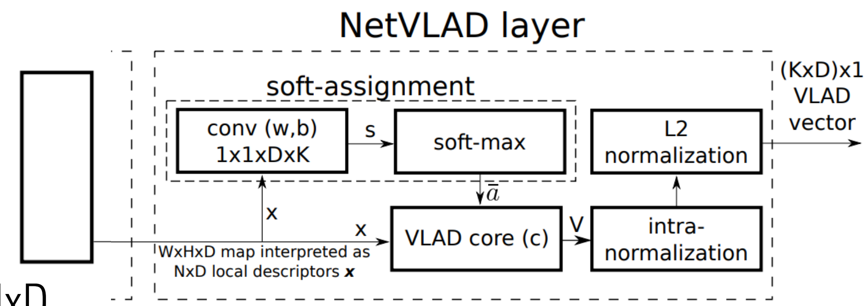
[4] H. Jegou, M. Douze, C, and P. Perez. “Aggregating local descriptors into a compact image representation.” In CVPR, 2010.

[5] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. “NetVLAD: Cnn architecture for weakly supervised place recognition.” In CVPR, 2016.

[6] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, “Multi-modal Transformer for Video Retrieval,” In *ECCV, 2020.*

APPENDIX

- Calculation in NetVLAD [3]:



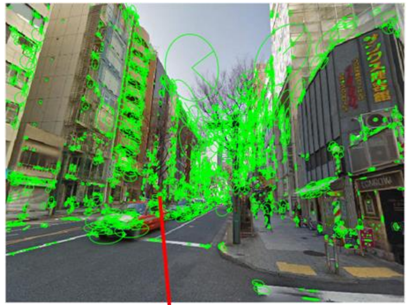
WxHxD
feature map
of last conv.
layer

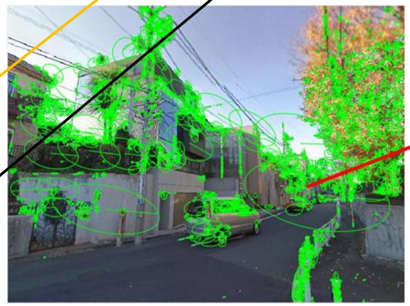
Convolution with filters
w_k and biases b_k

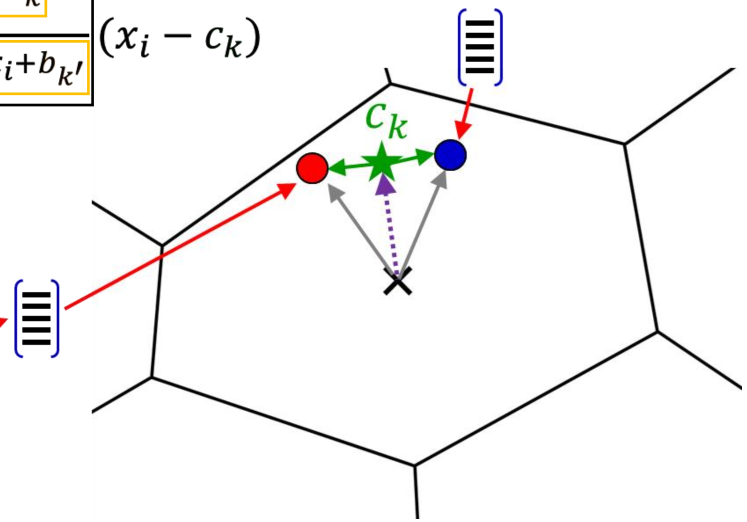
Soft-max -> $\bar{a}$

Decouple assignment (w_k b_k) from anchor point c_k

$$V(:,k) = \sum_{i=1}^{N} \frac{\boxed{e^{w_k^T x_i + b_k}}}{\sum_{k'} \boxed{e^{w_{k'}^T x_i + b_{k'}}}} (x_i - c_k)$$







X. Wang, L. Zhu, and Y. Yang, "T2LVAD: global-local sequence alignment for text-video retrieval,". In *Proc. IEEE/CVF CVPR*, Jun. 2021, pp. 5079-5088.

https://www.di.ens.fr/willow/research/netvlad/

THANK YOU!