Second International Workshop on Holistic Video Understanding

MDMMT: Multi-domain Multimodal Transformer for Video Retrieval

https://github.com/papermsucode/mdmmt

M. Dzabraev, M. Kalashnikov, S. Komkov, A. Petiushko

Lomonosov Moscow State University, Huawei Moscow Research Center

25th of June, 2021

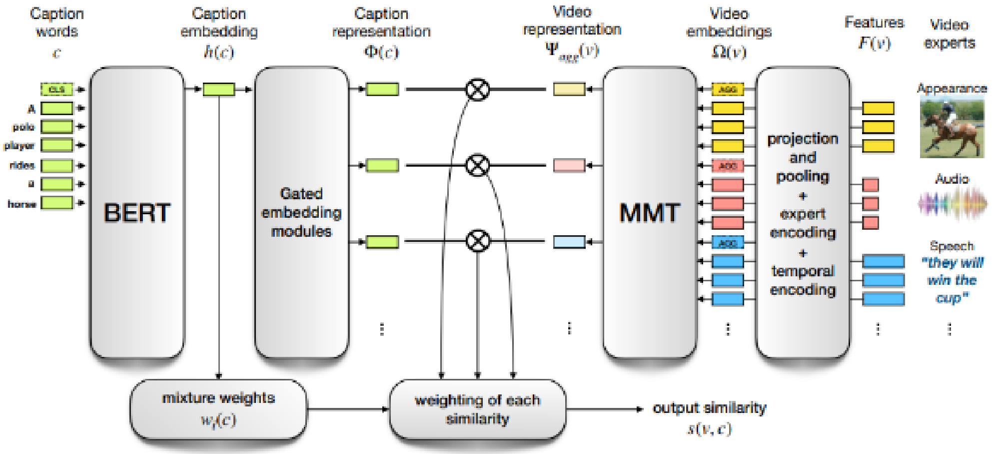
MDMMT: in brief

Main points of our work:

- We are solving the task of **text to video retrieval**: searching **video segments** using **textual queries**
- Our model is designed for **general usage** (our goal is to **decrease the bias** for any specific test dataset)
- Our model **extracts and fuses** information from **different modalities**: video, static images, sound

Result: we managed to create the **single model** that shows **state-of-the-art** performance on **different benchmarks** such as [*LSMDC*] and [*MSRVTT*]

Video retrieval task



Video retrieval task:

- ***G*** - gallery of videos, ***q*** - textual query
- ***q*** is a natural language description of video we'd like to find in ***G***
- **Task**: to find the most relevant videos from ***G*** for the query ***q***

Input text query

girl smiling and kissing

Text2Video engine

Output video

Prior art:

- Our work is based on [*MMT*] architecture
- MMT =
  - **pre-trained "experts"** to extract sequence of embeddings from **different modalities** +
  - **aggregation** by **transformer** encoder into the single video embedding $\boldsymbol{e_v}$ +
  - [***BERT***] to obtain the single **textual** embedding $\boldsymbol{e_t}$ +
  - cosine similarity **sim(**$\boldsymbol{e_v}$**,** $\boldsymbol{e_t}$**)** which is treated as the **score** between **text** and **video**

Video retrieval: datasets and metrics

Text-to-video retrieval **datasets** (MSRVTT, LSMDC, etc):

- Consist of pairs ***(video segment, textual description)***

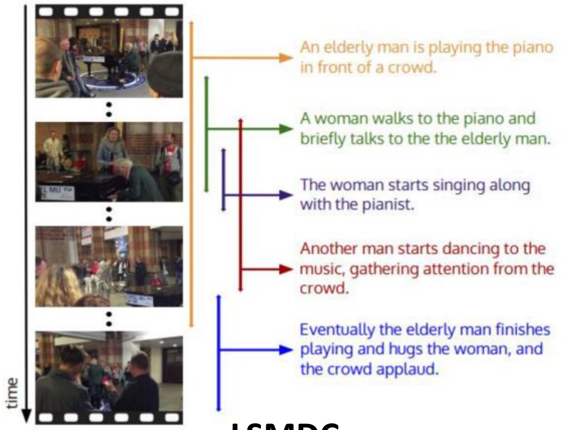
Evaluation metric:

- Most common for this task is **R@5** (Recall@top-5)
  - Test: set of queries $\boldsymbol{Q=\{q_1,..,q_n\}}$ and videos $\boldsymbol{G=\{g_1,...,g_n\}}$
  - $\boldsymbol{q_k}$ describes $\boldsymbol{g_k}$
  - Init R@5 := 0
  - If $\boldsymbol{score(q_k, g_k)}$ in largest top-5 scores $\boldsymbol{score(q_k, g_j)}$, then
    - R@5 += $1/|Q|$
  - Repeat for each $\boldsymbol{q_k}$

MSR-VTT

1. A black and white horse runs around.
2. A horse galloping through an open field.
3. A horse is running around in green lush grass.
4. There is a horse running on the grassland.
5. A horse is riding in the grass.

ActivityNet



LSMDC

AD: Abby gets in the basket.

Script: After a moment a frazzled Abby pops up in his place.

Mike leans over and sees how high they are.

Mike looks down to see – they are now fifteen feet above the ground.

Abby clasps her hands around his face and kisses him passionately. For the first time in her life, she stops thinking and grabs Mike and kisses the hell out of him.

MDMMT: motivation

Train data	MSR-VTT	ActivityNet	LSMDC
MSR-VTT	**29.0**	13.4	12.9
ActivityNet	14.7	**30.9**	10.4
LSMDC	8.8	7.2	**24.7**

- **Previously**:
  - Solutions are mostly **dataset-specific**, e.g.:
    - Such solutions work well either for MSR-VTT or for LSMDC, but **not for both**
    - Such solutions are **not applicable for dataset-unaware scenarios**

- Our **goal** is to create the **single model** which:
  - **Works well** on MSR-VTT, LSMDC and **other video retrieval datasets at the same time**

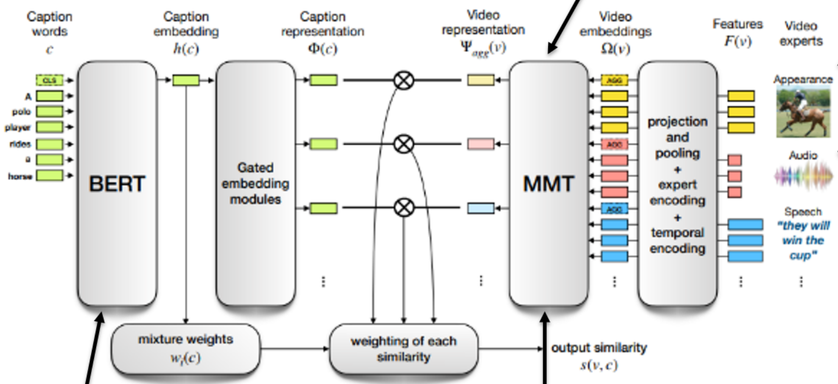
MDMMT: approach

Our **main improvements** over [*MMT*] are:

- We use **stronger pre-trained** feature extractors:
  - [*irCSN152 IG65M*] for **video** stream
  - [*CLIP*] for processing independent **frames** from video
  - [*VGGish*] for raw **audio** stream
- We use **significantly more data** for training:
  - MSR-VTT, [*ActivityNet*], LSMDC, [*TwitterVines*], [*YouCook2*], [*MSVD*], [*TGIF*], [*Something-to-SomethingV2*] at once

We **increase number of heads** to allow model learn single modal heads and cross-modal heads. Additionally we **increase depth** of transformer because larger train database has more knowledge

SotA on **Kinetics700 benchmark** (sort of ImageNet for Video) is **not the best** pre-trained experts. We tested many pre-trained networks for different tasks and **found the best one**



Original **MMT** uses **7 modalities**: *motion, appearance, audio, ASR, FaceID, OCR, scene*. It is difficult to scale such models to real life.
We use the **only motion** or **motion + appearance + audio**

Despite we significantly enlarge database we still observe **overfitting**. We **enlarge dropout** for text model and video model

MDMMT: results

Train data \ Test	MSRVTT	ActivityNet	LSMDC
MSR-VTT	29.0	13.4	12.9
ActivityNet	14.7	30.9	10.4
LSMDC	8.8	7.2	24.7
Above+other	**34.5**	**32.4**	**27.4**

Use only MSR-VTT for training: **poor** generalization

Use only ActivityNet for training: **poor** generalization

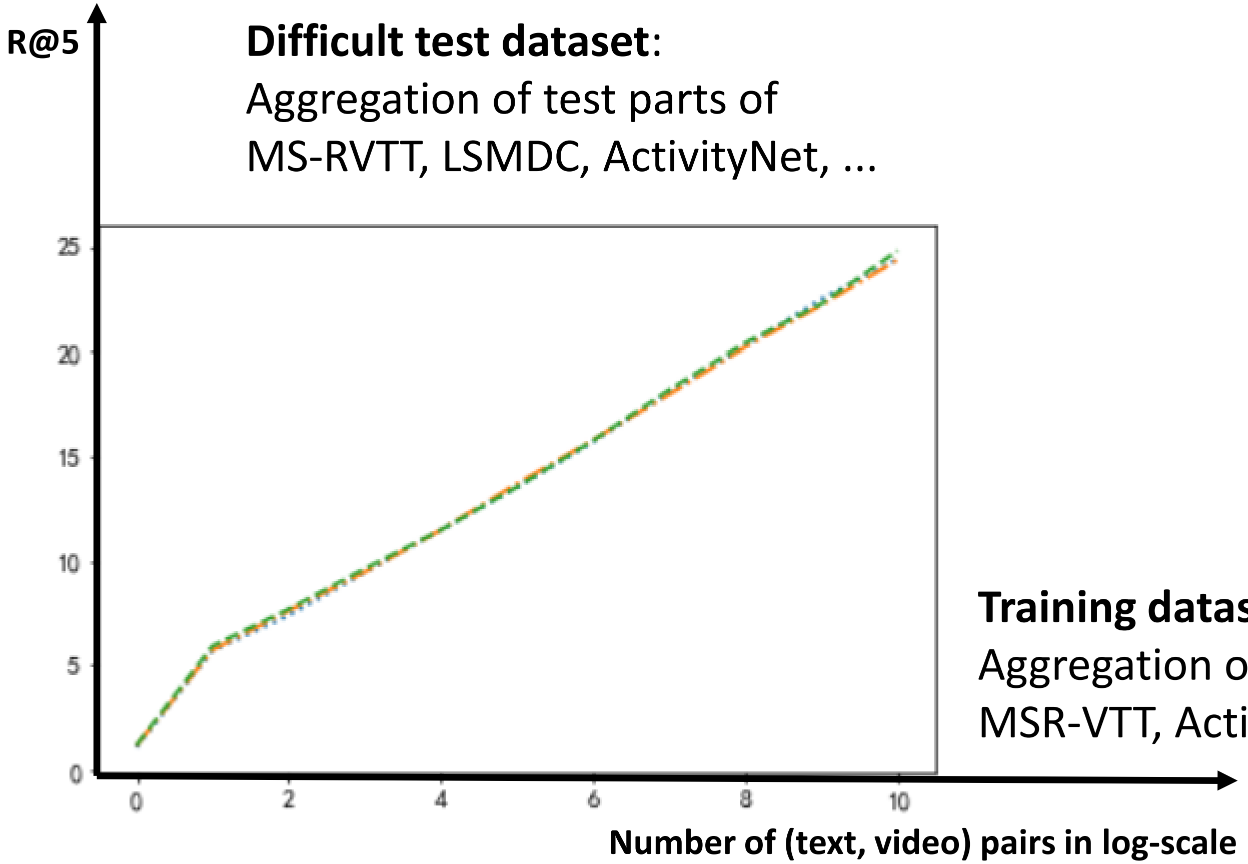
Use only LSMDC for training: **poor** generalization

Use 8 datasets for training: **good** generalization

Method	Train data	MSR-VTT 1kA	LSMDC
MMT 7mod	MSR-VTT 1kA	57.1	-
	LSMDC	-	29.9
MDMMT 3mod	MSR-VTT 1kA + LSMDC + other	**69.0**	**38.5**
CLIP	WIT	44.3	23.7
[Clip4Clip]	MSR-VTT 1kA + WIT	71.5	-
	LSMDC + WIT	-	41.8

MDMMT: log-linear generalization

Difficult test dataset:
Aggregation of test parts of
MS-RVTT, LSMDC, ActivityNet, ...



Training dataset:
Aggregation of training parts of
MSR-VTT, ActivityNet, LSMDC, ...

This figure shows:

- To create the video retrieval system for **general usage** we still have a room to **significantly increase** the training dataset

MDMMT: summary and conclusion

- We presented **MDMMT** which is improvement of original [*MMT*], where we use to the *best* to our knowledge pre-trained *feature extractors* and significantly *more data* for training
- Our solution achieves **state-of-the-art** result on MSR-VTT and LSMDC benchmarks using a **single model without fine-tuning**
- There is still a **big room of improvement** because we observe **log-linear R@5 growing** depending on training dataset size
- **Our plans** are:
  - **enlarge training database**: collect more image/video-captioning datasets
  - use **end-2-end training** (like [*CLIP4CLIP*]) using large training dataset

References

[**MMT**] Gabeur, Valentin, et al. "Multi-modal transformer for video retrieval." European Conference on Computer Vision (ECCV). Vol. 5. 2020.

[**CLIP4CLIP**] Luo, Huaishao, et al. "CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval." arXiv preprint arXiv:2104.08860 (2021).

[**CLIP**] Radford, Alec, et al. "Learning transferable visual models from natural language supervision." *arXiv preprint arXiv:2103.00020* (2021).

[**irCSN152, IG65M**] Ghadiyaram, Deepti, Du Tran, and Dhruv Mahajan. "Large-scale weakly-supervised pre-training for video action recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[**VGGish**] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017.

[**BERT**] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[**MSRVTT**] Xu, Jun, et al. "Msr-vtt: A large video description dataset for bridging video and language." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[**ActivityNet**] Krishna, Ranjay, et al. "Dense-captioning events in videos." *Proceedings of the IEEE international conference on computer vision*. 2017.

[**LSMDC**] Rohrbach, Anna, et al. "Movie description." International Journal of Computer Vision 123.1 (2017): 94-120.

[**TwitterVines**] Awad, George, et al. "TRECVID 2020: A comprehensive campaign for evaluating video retrieval tasks across multiple application domains." *arXiv preprint arXiv:2104.13473* (2021).

[**YouCook2**] Zhou, Luowei, Nathan Louis, and Jason J. Corso. "Weakly-supervised video object grounding from text by loss weighting and object interaction." arXiv preprint arXiv:1805.02834 (2018).

[**MSVD**] Chen, David, and William B. Dolan. "Collecting highly parallel data for paraphrase evaluation." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011.

[**TGIF**] Li, Yuncheng, et al. "TGIF: A new dataset and benchmark on animated GIF description." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[**Something-to-SomethingV2**] Goyal, Raghav, et al. "The" something something" video database for learning and evaluating visual common sense." Proceedings of the IEEE International Conference on Computer Vision. 2017.

Thank you!