Everything at Once – Multi-modal Fusion Transformer for Video Retrieval

Nina Shvetsova [1] Brian Chen [2] Andrew Rouditchenko[3] Samuel Thomas[4,5]
Brian Kingsbury[4,5] Rogerio Feris[4,5] David Harwath[6] James Glass[3] Hilde Kuehne[1,5]
[1]Goethe University Frankfurt, [2]Columbia University, [3]MIT CSAIL [4]IBM Research AI, [5]MIT-IBM Watson AI Lab, [6] UT Austin

Abstract

We present a novel approach for learning multi-modal representation from unlabeled video data. In particular, we propose: 1) a multi-modal, modality agnostic fusion transformer that learns to exchange information between multiple modalities, such as video, audio, and text, and integrates them into a joint multi-modal representation; 2) a new combinatorial loss to train the system on everything at once, single modalities as well as any combination of modalities. The proposed approach is evaluated on four challenging benchmark datasets and obtains state-of-the-art results in zero-shot video retrieval and step action localization. Our code for this work is also available.[1]

1. Introduction & Related Work

Information of co-occurrence of inputs from different modalities can be leveraged to learn meaningful representation of its content [1–3, 5–7, 11, 12, 15, 16, 21, 22]. Recently Miech *et al*. [16] used contrastive learning to train a multi-modal text-video embedding space from a large-scaled HowTo100M dataset of instructional videos in a self-supervised fashion where text description is obtained by an automated speech recognition system. Current methods [1, 4, 9, 10, 14–16, 21, 22] learn modality-specific encodings by projecting inputs to a common space and comparing representations of different modalities with each other by pairwise contrastive losses. Approaches that create different embedding space for different modality combinations [2], or learn a fused representation of several modalities (such as video-audio [11, 17, 19]), or train modality-agnostic projection [1] have also been studied. However, we believe that so far, cross-modal information has not been fully utilized during training, and none of these methods allows to obtain a joint representation of any given number of input modalities.

Our work aims to fill this gap and thus presents an approach that leverages self-attention for multi-modal learning to process any number of modalities jointly allowing modalities to attend to each other. As shown in Figure 1, input tokens from one or more modalities are passed through a modality-agnostic fusion transformer attending relevant features for the combined input. The model is trained with a novel combinatorial loss that considers contrastive loss between all possible and available modality combinations. As a result, our model can fuse any combination of input modalities and project it into a common embedding space incorporating cross-modality information and enabling such tasks as cross-modal retrieval and action localization. The proposed method allows us to improve performance on four challenging benchmark datasets.

[1]https://github.com/ninatu/everything_at_once

2. Method

Problem Statement. Our goal is to learn a projection function of single modalities or a set of modalities into the joint embedding space in a way that semantically similar inputs would be close to each other. We consider three modalities: video v, audio a, and text t, but the proposed method can be easily extended to more modalities. More formally, given a set of text-video-audio triplets $\{(t_i, v_i, a_i)\}_{i=1}^N$ of N video clips we are learning a projection $f(\cdot,\cdot,\cdot)$ that takes up to three modalities: v, a, and t, and produces $d-$dimensional embedding representation of the input.

Token Creation. As illustrated in Figure 1, our architecture starts from token extraction using modality-specific backbones, projection and normalization layers.

Multi-modal Fusion Transformer. To learn a projection f that can fuse information from multiple modalities to enhance the joint representation, we propose a multi-modal, modality agnostic transformer, where the keys, queries, and values of the input tokens are computed independently from the modality. We adopt a regular transformer blocks [23]; but note, the difference compared to other methods is not in the architecture itself, but in the way it is trained to fuse any combination of input modalities. We train the system with a combinatorial input. Namely, we apply it to joint sets of input tokens from all possible combinations of modalities: singles - a, v, t, and pairs - (a, v), (a, t), (t, v), allowing tokens from one modality to attend tokens from other modalities. Therefore, we apply it six times to obtain six representations, such as the combination (v, a) will result in

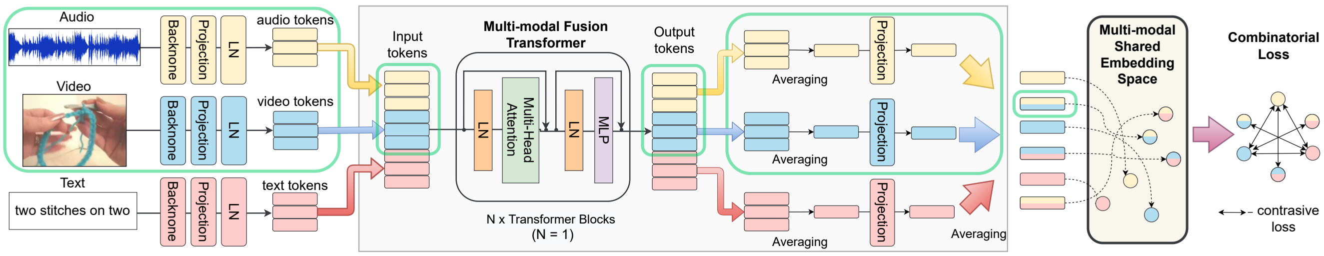

Figure 1. The proposed method. During training, we apply the model six times to obtain six embeddings corresponding to text, video, audio, text-video, text-audio, and video-audio modalities to compute the combinatorial loss, we exemplary consider the audio-video pair marked with green rectangles here. LN – normalization layer [8]

a fused representation of va.

Projection to Shared Embedding Space. As an example of one of six embeddings, we consider creating the final representation for va. Since modalities, even enhanced with other modalities, are still very different, we divide output tokens into groups based on modality (v and a in the considered case) and average them. Then we project embeddings into the shared embedding space by the modality-specific projections and average embeddings for v and a to get a final representation of va.

Combinatorial Loss. Unlike other methods [1, 2, 10, 22] that apply contrastive losses only between single-modalities, we force tokens to exchange information between modalities by enabling contrastive losses with fused modalities as well using our *combinatorial loss*: $L = \lambda_{t\_v} L_{t\_v} + \lambda_{v\_a} L_{v\_a} + \lambda_{t\_a} L_{t\_a} + \lambda_{t\_va} L_{t\_va} + \lambda_{v\_ta} L_{v\_ta} + \lambda_{a\_tv} L_{a\_tv}$, where λ_{x_y} denotes a weighting coefficient and L_{x_y} denotes contrastive loss between (x, y). For L_{x_y}, we use NCE [18] with temperature τ and batch size B:

$$L_{x\_y} = -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\exp(x_i^\top y_i/\tau)}{\sum_{j=1}^{B}\exp(x_i{}^\top y_j/\tau)}\right) - \\ -\frac{1}{B}\sum_{i=1}^{B} \log\left(\frac{\exp(x_i^\top y_i/\tau)}{\sum_{j=1}^{B}\exp(x_j{}^\top y_i/\tau)}\right), \quad (1)$$

that pushes embeddings x_i and y_i (for modalities x and y) of the same clip together and pushes them apart to other examples in a minibatch.

3. Experimental Evaluation

To ensure comparability, we follow the setup of most previous works [2, 4, 9, 10, 16, 22] wherever possible (8-sec training clips, backbones, gating projections, etc.). For the sake of space, we excluded comparison with methods that use much stronger backbones.

Tasks & Datasets Following previous works [4, 10, 16, 22] we train our model on the HowTo100M dataset [16] and evaluate it in zero-shot text-to-video retrieval on MSR-VTT [24] and YouCook2 [25] datasets and zero-shot step action localization on CrossTask [27] and Mining YouTube [13] datasets

Method	Visual Backbone	YouCook2 R@10↑	YouCook2 MedR↓	MSR-VTT R@10↑	MSR-VTT MedR↓
		$t \to v$			
ActBERT [26]	Res3D+F.R-CNN	38.0	19	33.1	36
Support Set [20]	R152 + R(2+1)D-34	-	-	31.1	31
HT100M [16]	R152 + RX101	24.8	46	29.6	38
NoiseEstim. [4]	R152 + RX101	-	-	30.4	36
Ours	R152 + RX101	38.9	19	**35.3**	**25**
		$t \to va$			
MMT [11]	7 experts	-	-	-	66
AVLNet [22]	R152+RX101	44.3	16	27.4	47
MCN [10]	R152+RX101	45.2	-	33.8	-
Ours	R152+RX101	**51.3**	**10**	31.8	30

Table 1. Zero-shot text-to-video retrieval on YouCook2/MSR-VTT.

Method	Tr. Mod.	Tr. BB v	Visual Backbone	Recall↑ CrossTask	Recall↑ MYT
CrossTask [27]	tv		R152 + I3D	31.6	-
HT100M [16]	tv		R152 + RX101	33.6	15.0
MIL-NCE [15]	tv	✓	I3D	36.4	-
MCN [10]	tva		R152 + RX101	35.1	18.1
Ours	tva		R152 + RX101	**39.3**	**19.4**

Table 2. Zero-shot action localization. Tr Mod=Training Modalities, Tr BB v= Trainable Backbone for video modality.

(we follow the inference procedure in [27]). We use fused va representation for video.

Results. In zero-shot text-to-video retrieval (Table 1), our method achieves state-of-the-art results over all baselines on YouCook2, particularly, significantly outperforming the AVLnet [22] and MCN [10] that also train with three modalities and use the same backbones. For MSR-VTT however, a fusion of video and audio modalities is not so beneficial and best performance is reached when considering only text to video retrieval and leaving out audio information. We attribute this behaviour to the domain shift as audio of the HowTo100M mainly contains speech and text as a transcription of speech, while in MSR-VTT audio can be much less related to the textual description. In zero-shot step action localization (Table 2) the proposed approach clearly outperforms the directly comparable MCN approach on both datasets, as well as HT100M [16] and MIL-NCE [15] with a trainable I3D backbone [15] and a fully supervised CrossTask [27].

4. Conclusion

In this work, we proposed the multi-modal, modality agnostic transformer that learns to fuse information from multiple modalities and integrates it into a joint multi-modal representation. We showed that training the system with the combinatorial loss on any possible combinations of modalities allows the fusion transformer to learn a strong multi-modal embedding space and achieve state-of-the-art results in zero-shot video retrieval and zero-shot step action localization.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021. 1, 2

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *NeurIPS*, 2020. 1, 2

[3] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 1

[4] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*, 2020. 1, 2

[5] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 1

[6] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 1

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 1

[8] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2

[9] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1, 2

[10] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. *ICCV*, 2021. 1, 2

[11] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 1, 2

[12] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018. 1

[13] Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. Mining youtube - a dataset for learning fine-grained action concepts from webly supervised video data. *arXiv preprint arXiv:1906.01012*, 2019. 2

[14] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 1

[15] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 1, 2

[16] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 1, 2

[17] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[18] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[19] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 1

[20] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1

[22] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. Avlnet: Learning audio-visual language representations from instructional videos. In *Interspeech*, 2021. 1, 2

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 1

[24] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2

[25] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2

[26] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 2

[27] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 2