Spoken Moments: Learning Joint Audio-Visual Representations from Video Descriptions

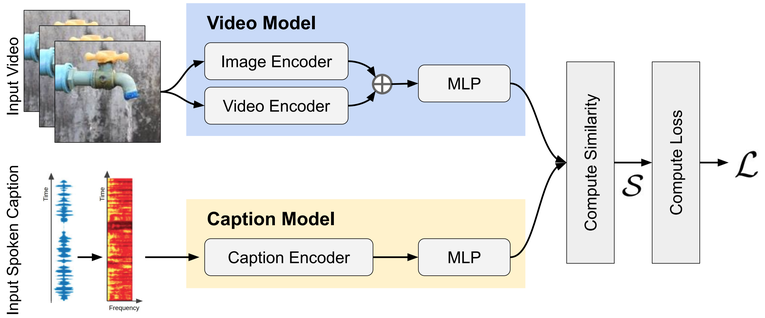
Mathew Monfort[1]*, SouYoung Jin[1]*, Alexander Liu[1], David Harwath[2], Rogerio Feris[3], James Glass[1], Aude Oliva[1]
[1]MIT CSAIL, [2]UT Austin, [3]IBM Research * equal contribution

The Spoken Moments Dataset

- We propose a new video caption dataset of **500k videos** (sourced from the Multi-Moments dataset [1]) each with a **unique spoken caption** describing what is happening in the video. We also generate **text captions** using an automatic speech recognition system to feed them to language models.
- Comparison of our dataset to existing video caption datasets.

Dataset	Clips	Videos	Captions	Words	Vocab	Domain
TACoS	7,206	127	18,227	146,771	28,292	Cooking
YouCook II	15,400	2,000	15,400	121,418	2,583	Cooking
MSVD	1,970	1,970	70,028	607,339	13,010	General
Charades	10,000	10,000	27,800	645,636	32,804	General
MPII-MD	68,337	94	68,375	653,467	24,549	General
MSR-VTT	10,000	7,180	200,000	1,856,523	29,316	General
ActivityNet Captions	100,000	20,000	100,000	1,348,000	15,564	General
VideoStory	123,000	20,000	123,000	1,633,226	-	General
Epic-Kitchens	76,885	633	76,885	227,974	1,737	Cooking
Vatex-en	41,300	41,330	413,000	4,994,768	44,103	General
Spoken Moments	**515,912**	**459,742**	**515,912**	**5,618,064**	**50,570**	**General**

Learning Audio-Visual Representations



- To learn from the large set of spoken captions in the Spoken Moments dataset (S-MiT) we adopt a cross-modal architecture [2, 3, 4].
- We propose a novel approach to contrastive loss where we compute the margin based on the difference between the similarity of the positive pair and the set of negative pairs in each batch (Eq. 2). We call this **Adaptive Mean Margin (AMM)**. We replace M in Masked margin softmax (Eq. 1) with our proposed AMM, where α=0.5 is a dampening parameter to weight the strength of the margin.

$$\mathcal{L}_{xy} = -\frac{1}{B}\sum_{i=1}^{B} log \frac{e^{\mathcal{S}(x_i,y_i)-M}}{e^{\mathcal{S}(x_i,y_i)-M} + \sum_{j=1}^{B} I_{i\neq j} e^{\mathcal{S}(x_i,y_j)}} \quad (1)$$

$$M_{xy} = \alpha\big(\mathcal{S}(x_i,y_i) - \frac{1}{B-1}\sum_{j=1}^{B} I_{i\neq j}\mathcal{S}(x_i,y_j)\big) \quad (2)$$

Video/Caption Retrieval

- The proposed AMM loss function **consistently achieves the highest performance** in multiple spoken/language caption model architectures.
- Loss function comparisons on S-MiT where models are trained
  - with ResNet-50 architecture for spoken captions.

Loss	Caption to Video				Video to Caption			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Noise contrastive estimation (NCE) [5]	32.7	60.8	70.6	45.6	33.1	59.4	69.6	45.5
Semi-hard negative mining (SHN) [6]	33.9	60.1	70.9	45.8	34.0	60.6	70.1	46.0
Masked margin softmax (MMS) [7]	37.2	65.4	75.1	50.0	37.8	64.6	74.2	50.1
Adaptive Mean Margin (AMM)	**39.5**	**65.7**	**75.5**	**51.6**	**40.1**	**66.3**	**74.5**	**52.0**

- **Cross Dataset Evaluation:** The S-MiT model shows it **generalizes strongly to the other datasets** even beating the MSR-VTT model on its own test set.

Trained On	Evaluated On: Vatex				ActivityNet				MSR-VTT				S-MiT			
	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP	R@1	R@5	R@10	mAP
Vatex	**45.9**	**79.7**	**87.5**	**60.7**	15.6	39.4	51.7	27.1	22.6	49.8	63.2	35.6	13.1	33.0	45.8	23.5
ActivityNet	25.0	56.0	68.4	39.1	**19.1**	**48.1**	**61.2**	**32.5**	15.1	37.1	50.4	26.4	9.8	28.7	40.6	19.7
MSR-VTT	21.0	51.3	64.8	35.1	9.9	28.3	39.7	19.6	29.1	64.2	**77.9**	44.8	14.6	39.3	53.4	26.9
S-MiT	42.7	75.4	84.2	57.1	17.6	41.6	53.8	29.2	**33.1**	**64.8**	77.4	**47.6**	**38.4**	**68.5**	**78.7**	**52.1**

[1] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, Aude Oliva, Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding. In arXiv preprint arXiv:1911.00232, 2020
[2] 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 Int. Conf. Comput. Vis., pages 2630–2640, 2019.
[3] David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. Int. J. Comput. Vis., (128):620–641, 2020.
[4] Andrew Rouditchenko, Angie Boggust, David Harwath, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. Avlnet: Learning audio-visual language representations from instructional videos. In arXiv:2006.09199, 2020.
[5] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics,pages 297–304. JMLR Workshop and Conference Proceedings, 2010.
[6] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In IEEE Conf. Comput. Vis. Pattern Recog., June 2015.
[7] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large scale representation learning from visually grounded untranscribed speech. In Conference on Computational Natural Language Learning, pages 55–65, Nov. 2019.

http://moments.csail.mit.edu/spoken.html