M&M Mix: A Multimodal Multiview Transformer Ensemble

Xuehan Xiong, Anurag Arnab,
Arsha Nagrani, Cordelia Schmid

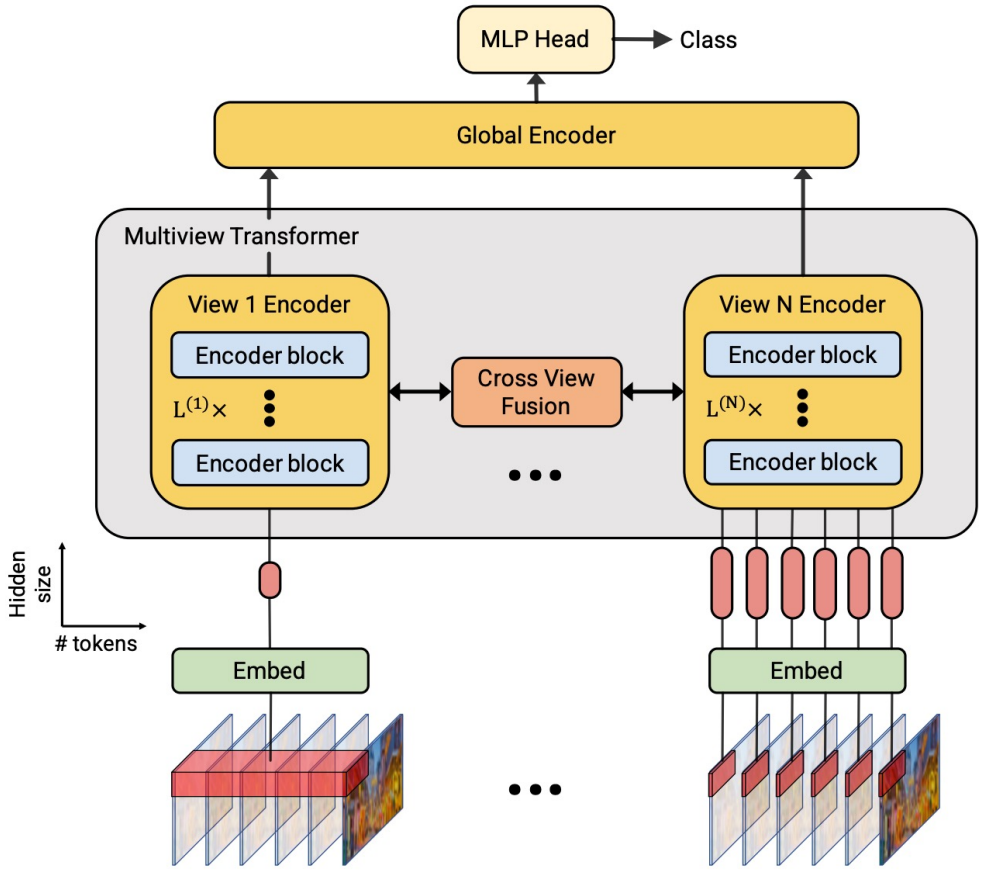
Google Research

Introduction

- Our entry extends our recent work, Multiview Transformers for Video Recognition (poster on Tuesday PM, 75b)
- Added audio and optical flow as additional modalities.

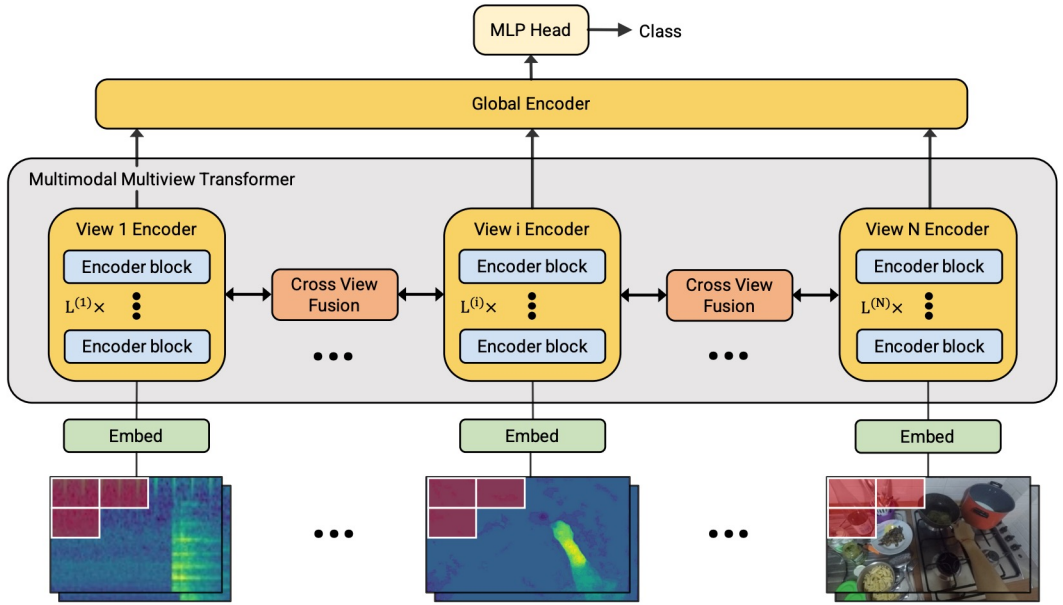
Multiview Transformers

- Model multiscale, temporal information
- Create different “views” of the input
- Process these views in parallel, with lateral connections between transformer layers.



Multimodal Multiview Transformers

- MTV: “Views” are constructed by tokenising the input with different tubelet sizes.
- M&M: “Views” are constructed by tokenizing different modalities.



Multiview Transformers

- Our naming convention example
- B/2 + S/4 + Ti/8
  - Three views
  - “Base” transformer with tubelet size of 16x2
  - “Small” transformer with tubelet size of 16x4
  - “Tiny” transformer with tubelet size of 16x8

Multimodal Multiview Transformers

- Our naming convention example
- B/2:**R** + S/4:**F** + Ti/8:**S**
  - R – RGB
  - F – Optical flow
  - S – Spectogram (audio)

Fusing different modalities

	Models	Top-1 Action	Top-1 Noun	Top-1 Verb
RGB only	B/2:R+S/4:R+Ti/8:R	52.7	66.1	71.2
Flow only	B/2:F+S/4:F+Ti/8:F	40.5	50.1	68.1

State-of-the-art single-model results.
MTV model from our CVPR paper.

Fusing different modalities

Models	Top-1 Action	Top-1 Noun	Top-1 Verb
B/2:R+S/4:R+Ti/8:R	52.7	66.1	71.2
B/2:F+S/4:F+Ti/8:F	40.5	50.1	68.1
B/2:R+S/4:F+Ti/8:R	53.4	**66.5**	71.9
B/2:R+S/4:S+Ti/8:R	53.2	66.3	**72.0**

RGB and flow

RGB and audio

Improvements from combining RGB and Audio, RGB and Flow.

Fusing different modalities

Models	Top-1 Action	Top-1 Noun	Top-1 Verb
B/2:R+S/4:R+Ti/8:R	52.7	66.1	71.2
B/2:F+S/4:F+Ti/8:F	40.5	50.1	68.1
B/2:R+S/4:F+Ti/8:R	53.4	**66.5**	71.9
B/2:R+S/4:S+Ti/8:R	53.2	66.3	**72.0**
B/2:R+S/4:S+Ti/8:F	**53.6**	66.3	**72.0**

Best model combines RGB, flow and audio.

Spatial resolution improves performance

- Using MTV B/2 + S/4 + Ti/8 with RGB input only.

Spatial resolution	Top-1 Action	Top-1 Noun	Top-1 Verb
224p	49.3	63.0	69.4
280p	50.5	63.9	69.9
432p	52.7	66.1	71.2

Pretraining datasets

- WTS is a dataset of videos scraped from YouTube by Stroud *et al*, arxiv 2020
- Labels are noisy. Clips follow the same format as Kinetics (10 seconds long, videos from YouTube)

Pretraining datasets	Top-1 Action	Top-1 Noun	Top-1 Verb
K400	46.7	60.5	67.8
K700	48.0	61.2	69.1
WTS	49.3	63.0	69.4

Single model results

- Our single model entry outperforms last year's winner (model ensemble), and existing state-of-the-art substantially.

Data split	Models	Top-1 Action	Top-1 Noun	Top-1 Verb
validation	MoViNet [18]	47.7	57.3	**72.2**
	MeMViT [28]	48.4	60.3	71.4
	Omnivore [12]	49.9	61.7	69.5
	M&M-B	**53.6**	**66.3**	72.0
test	2021 challenge winner [5]	48.7	59.2	**70.6**
	M&M-B	**49.6**	**63.7**	68.0

Final Results – Ensemble

- Ensemble separate models
- One ensemble for predicting verbs, another for predicting nouns.
- Ensembling improved results by X points, from Y to 52.8.

Model indices	Top-1 Action (val/test)	Top-1 Noun	Top-1 Verb
0,1,2,3,5,6,7,8,9,10	56.9/52.8	69.2/66.2	
4,5,6,7,8,9,10			75.0/70.9

Final Results – Ensemble

- Ensemble separate models
- One ensemble for predicting verbs, another for predicting nouns.
- Ensembling improved results by X points, from Y to 52.8.

Model indices	Model variants	Pretraining datasets	Resolution	Top-1 Action	Top-1 Noun	Top-1 Verb
0	B/2:R+S/4:R+Ti/8:F	WTS → K700	432p	53.4	66.4	71.8
1	B/2:R+S/4:F+Ti/8:R	WTS → K700	432p	53.4	66.5	71.9
2	L/2:R+B/4:F+S/8:F+Ti/16:R	WTS → K700	320p	53.0	66.7	71.1
3	L/2:R+B/4:R+S/8:R+Ti/16:R	WTS	352p	52.6	67.2	69.8
4	B/2:F+S/4:F+Ti/8:F	WTS → K700	432p	40.5	50.1	68.1
5	B/2:R+S/4:R+Ti/8:R (128×1)	WTS	304p	52.4	65.6	71.3
6	L/2:F+B/4:F+S/8:F+Ti/16:F	WTS → K700	352p	40.9	50.6	67.2
7	L/2:R+B/4:F+S/8:S+Ti/16:R	WTS	320p	53.6	67.0	71.7
8	B/2:R+S/4:S+Ti/8:F	WTS	432p	53.6	66.3	72.0
9	B/2:R+S/4:S+Ti/8:R	WTS	432p	53.2	66.3	72.0
10	B/2:R+S/4:R+Ti/8:S	WTS	432p	53.4	66.6	72.0

Conclusion

- Multimodal extension of our MTV architecture.
- High resolution, pretraining and ensembling further improved our results.

- Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Chen Sun, Cordelia Schmid. Multiview Transformers for Video Recognition, CVPR 2022.
- Code and models
- For more details, visit our poster session, Tuesday PM, 75b