Google Research

Multimodal Learning for Videos

Arsha Nagrani

\+ Work by many others

The artist brain is the sensory brain: sight and sound, smell and taste, touch. These are the elements of magic.
— Julia Cameron

Brief Intro

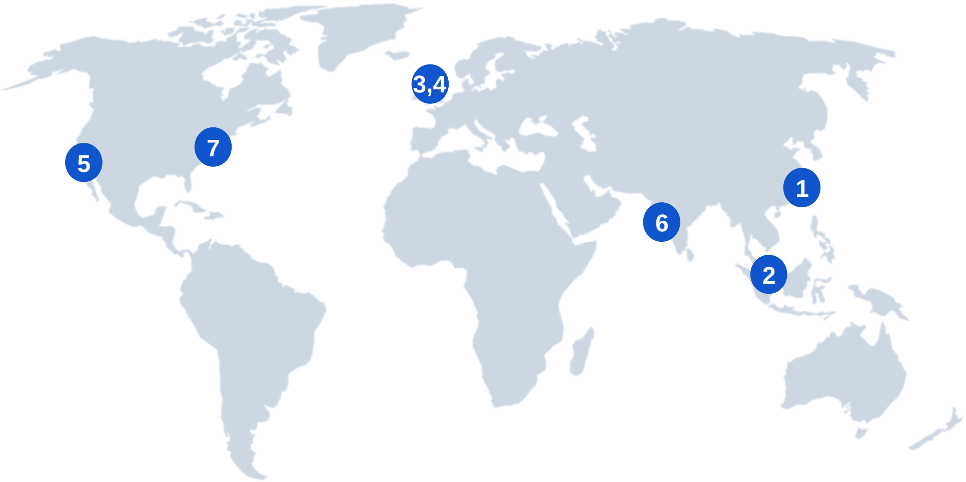
3 **2012-16:** **University of Cambridge**
UG in Computer Eng

4 **2016-20:** **University of Oxford**
DPhil, Computer Vision
Andrew Zisserman

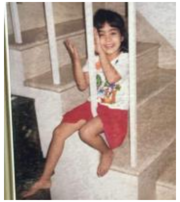
5 **2019:** **Google Research**
Intern,
Cordelia Schmid, Chen Sun

6 **2020:** **Wadhwani AI**
Visiting Researcher, non-profit

7 **2020-now:** **Google Research**
Research Scientist, Perception



2



Singapore

1

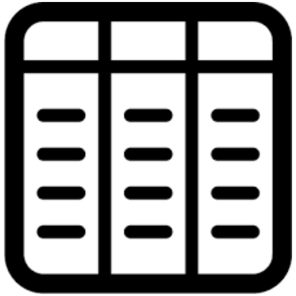
Hong Kong

Overview

- ❖ Why do we need multimodal learning?
  - ➢ Both a human and machine perspective
- ❖ How can we use it?
  - ➢ Cross-modal learning
  - ➢ Multimodal fusion
  - ➢ MBT (NeurIPS 2021), VideoCC (ArXiv 2022), AVATAR (Interspeech 2022)

What is multimodal learning?

Learning with more than one input data type



Numerical

Textual Data

Images/
Videos

Speech/
Sound

Why do we need it?

Look at a useful biological prototype - Humans

- We perceive the world with multiple sensory systems— vision, audition, touch, smell, proprioception, balance

(1) Degeneracy in neural structure

- ➢ A system functions even with the loss of one component
- ➢ Eg. Spatial properties are developed even in the blind, using touch, echolocation with tongue clicks and cane taps

Why do we need it?

Look at a useful biological prototype - Humans

- We perceive the world with multiple sensory systems— vision, audition, touch, smell, proprioception, balance

(2) Sensory systems can educate each other

- ➢ Learn to associate multiple representations - time locked and correlated
- ➢ Children spend hours gazing at their own hands, touching and feeling objects (Yuan et al 2019)

Transparency is difficult to learn

- 8 month old infants often struggle to retrieve from transparent boxes
- Infants who play with the objects physically were able to retrieve objects better

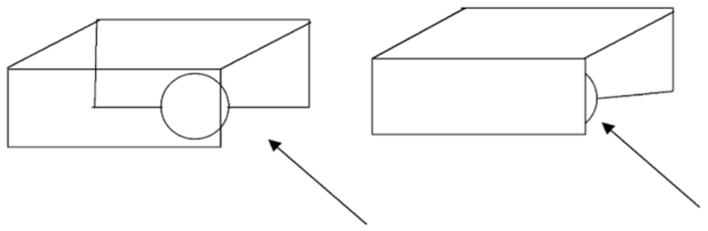


Figure 2. A toy (ball) hidden under a transparent box and an opaque box in the Diamond task. The opening is indicated by the arrow.

Titzer, Thelen, and Smith et al

Why do we need it?
Look at a useful biological prototype - Humans

- We perceive the world with multiple sensory systems— vision, audition, touch, smell, proprioception, balance

(3) Fusion of multiple senses helps with robustness

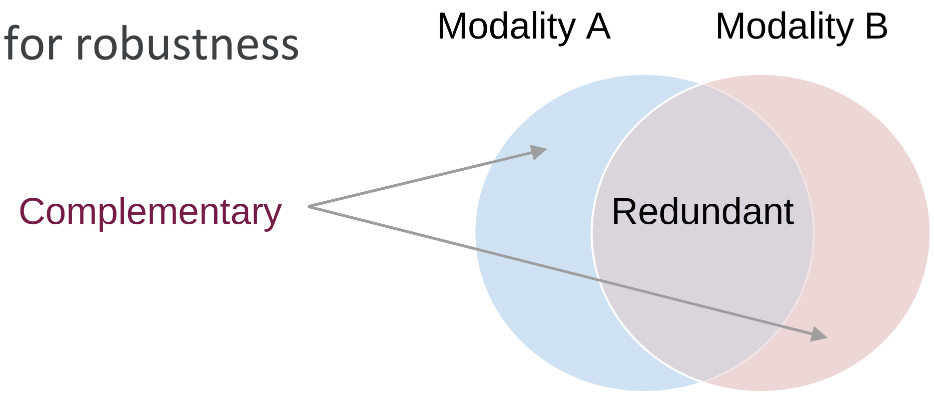
- Use multiple signals to come to a conclusion
- What we see affects what we hear and vice-versa, eg. the McGurk Effect

Machine learning perspective

- **Robustness**: Content on the web is inherently multimodal (captions, text, titles, descriptions). Why limit ourselves to use only one?
- **Self-supervision**: Use redundancy to learn with fewer labels
- **Applications**: Some applications are inherently multimodal
  - Video captioning Video-> Text
  - Automatic Speech recognition Audio -> Text

How can we use it? Some examples

1. **Cross-modal supervision:** Use one modality to help learn in another
    a. *Labelling data manually is tough*
        i. *expensive and subjective, hours of human time*
    b. *Use knowledge in one modality to inform another modality*
    c. *This can give us a source of 'free supervision'*
    d. *Exploits 'redundancy'*
2. **Fusion:** Combine multiple modalities for robustness
    a. *Exploits 'complementarity'*

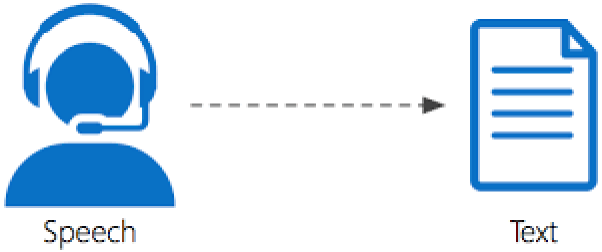


How can we use it? Dive into some recent papers

Audio + RGB Fusion: Combine multiple modalities for robustness

a. New transformer fusion architecture for video classification (NeurIPS 2021)
b. Learning audio-visual modalities from image captions (ArXiv 2022)
c. Audio-visual fusion for ASR (Interspeech 2022)

Action recognition, video retrieval



ASR

Google Research

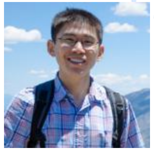
MBT: Attention Bottlenecks for Multimodal Fusion

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, Chen Sun

NeurIPS 2021





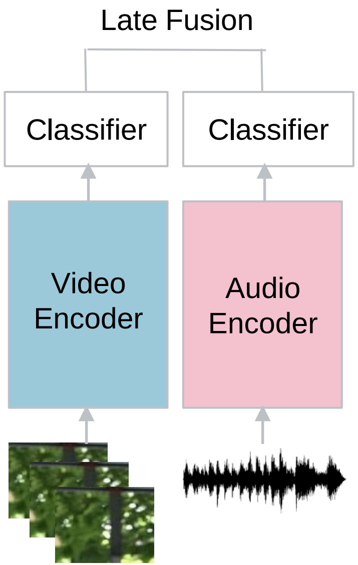
Multimodal Fusion

- Video is inherently multimodal - audio, vision, text etc
- Uni-modal inputs can be missing, corrupted, occluded, or have various levels of background noise
- Multimodal Fusion allows robustness, and disambiguation
- We want a single multimodal model that is:
  - Robust
  - Efficient and Scalable
  - Variable Length Inputs

Independent Communities

Multimodal Inputs

- ❖ Heterogeneity of inputs (RGB frames, audio spectrograms)
- ❖ Specialised architectures
- ❖ Different datasets and evaluation benchmarks



MBT: Attention Bottlenecks for Multimodal Fusion

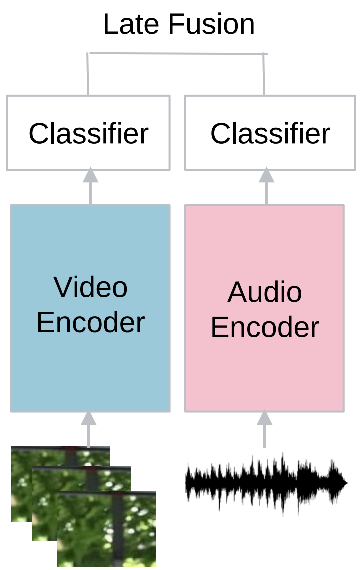
Independent Communities - Late Fusion

Multimodal Inputs

- ❖ Heterogeneity of inputs (RGB frames, audio spectrograms)
- ❖ Specialised architectures
- ❖ Different datasets and evaluation benchmarks

“The Dominant Paradigm”

- ❖ Different encoders
- ❖ Output scores or representations are fused right at the end
- ❖ This is in contrast to human perception (early or mid fusion)



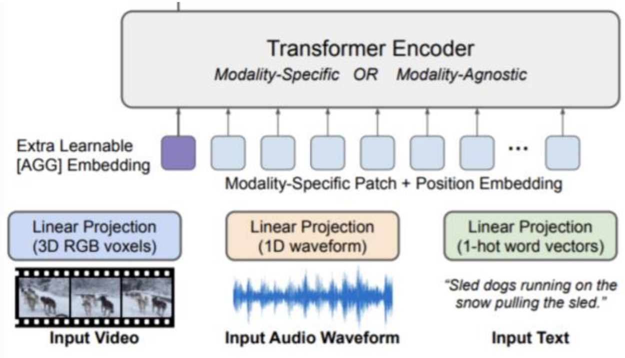
AUDIO-VISUAL SCENE ANALYSIS

Evidence for a “very-early” integration process in audio-visual speech perception

Jean-Luc Schwartz, Frédéric Berthommier, Christophe Savariaux

Advantages of Transformers

- Great for modelling context
  - Each token can have access to all other tokens in the sequence
- A generic architecture:
  - Operates on any inputs that can be tokenized! “Universal Perceptual Models”
- Parallelizable
- Empirically shown to perform excellently at scale



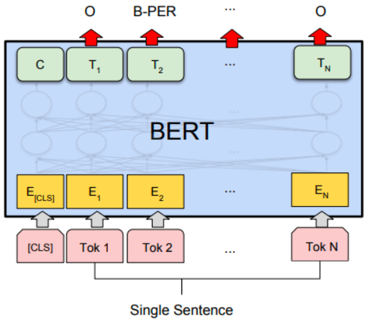
Akbari, H., Yuan, L., Qian, R., Chuang, W. H., Chang, S. F., Cui, Y., & Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*

Google Research

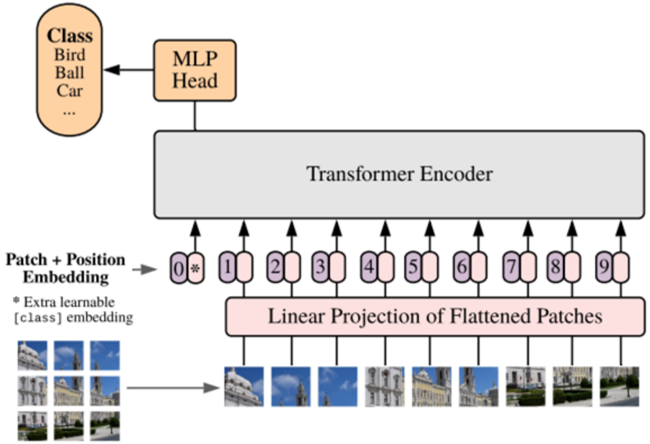
Transformers for early fusion?

- Transformers have had great success on different modalities individually
- Operate on tokens (and any modality can be tokenized)
- SOTA for Text (BERT), Images (ViT), Videos (ViViT, Timesformer), Audio (AST)

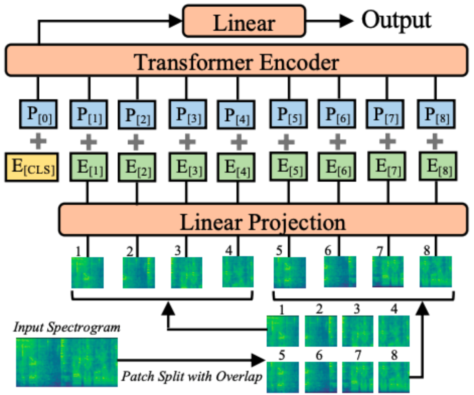
BERT



ViT

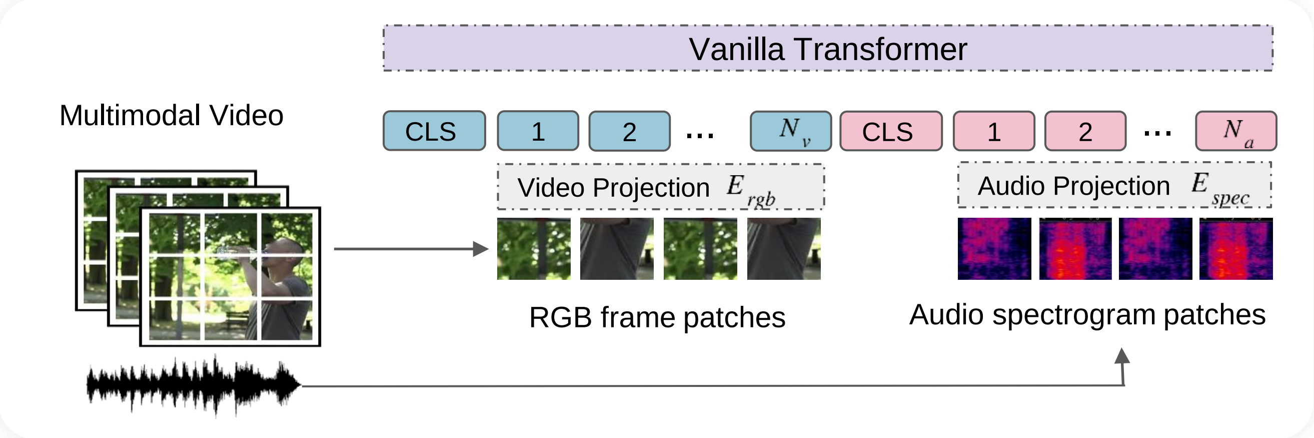


AST



A Vanilla Multimodal Transformer

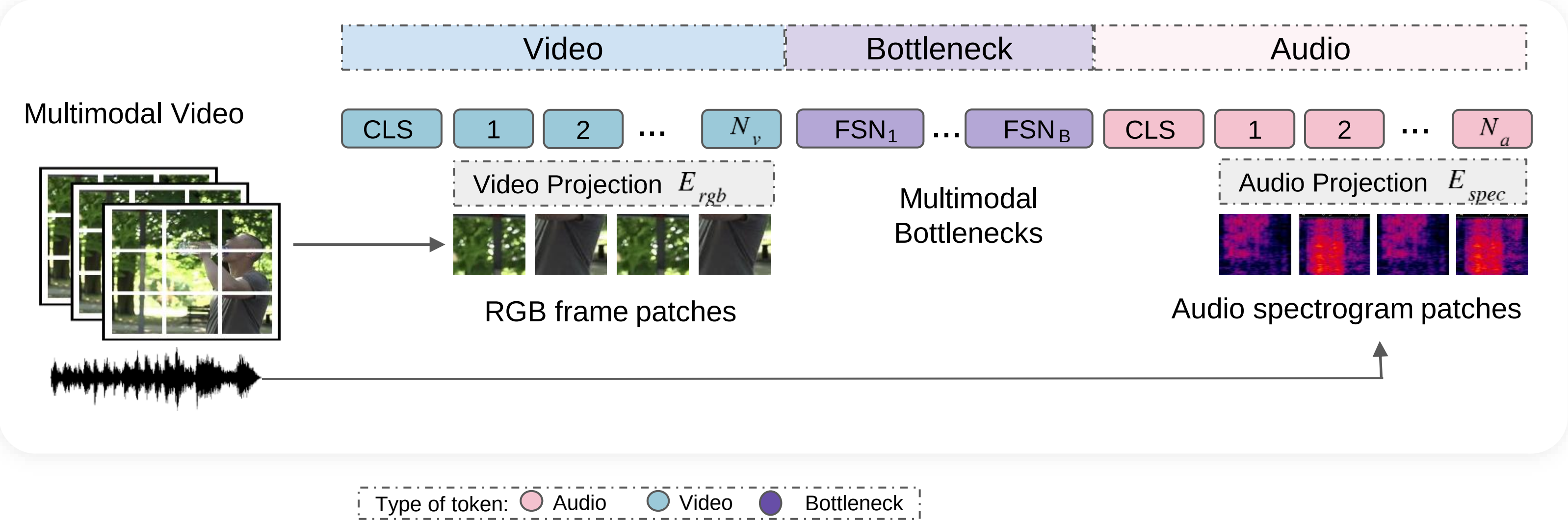
- ❖ Tokenize RGB frame and spectrogram patches
- ❖ Universal Perception model - feed all tokens to a transformer
- ❖ Pairwise self-attention between all tokens (early fusion)



- ❖ *scales quadratically with sequence length*
- ❖ *video has a lot of redundancy*

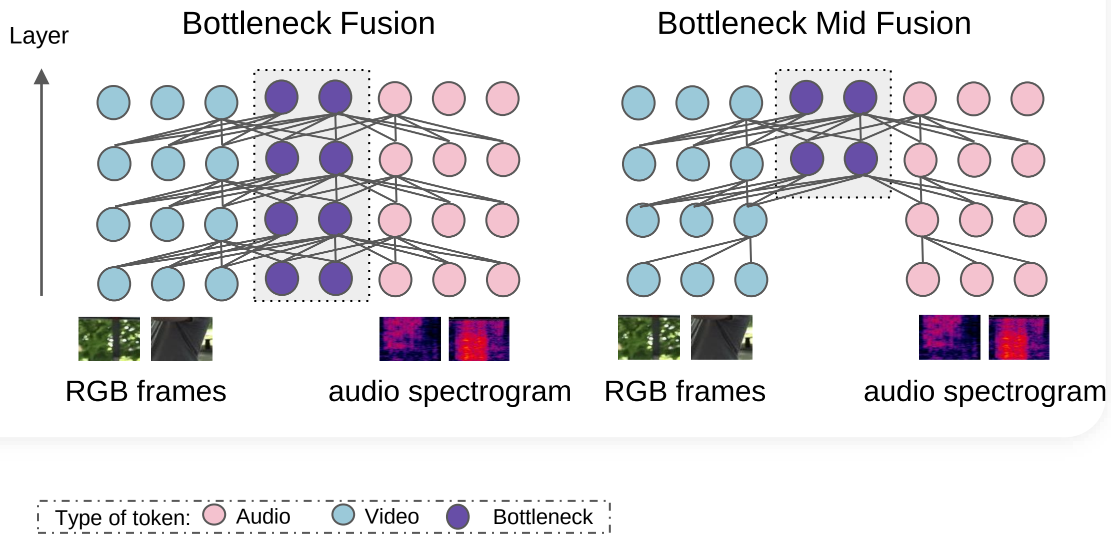
Multimodal Bottleneck Transformers (MBT)

- Introduce a small number of bottleneck tokens (B=4)
- Full pairwise self attention within a modality
- Attention between the visual tokens and the bottleneck tokens
- Attention between the audio tokens and the bottleneck tokens



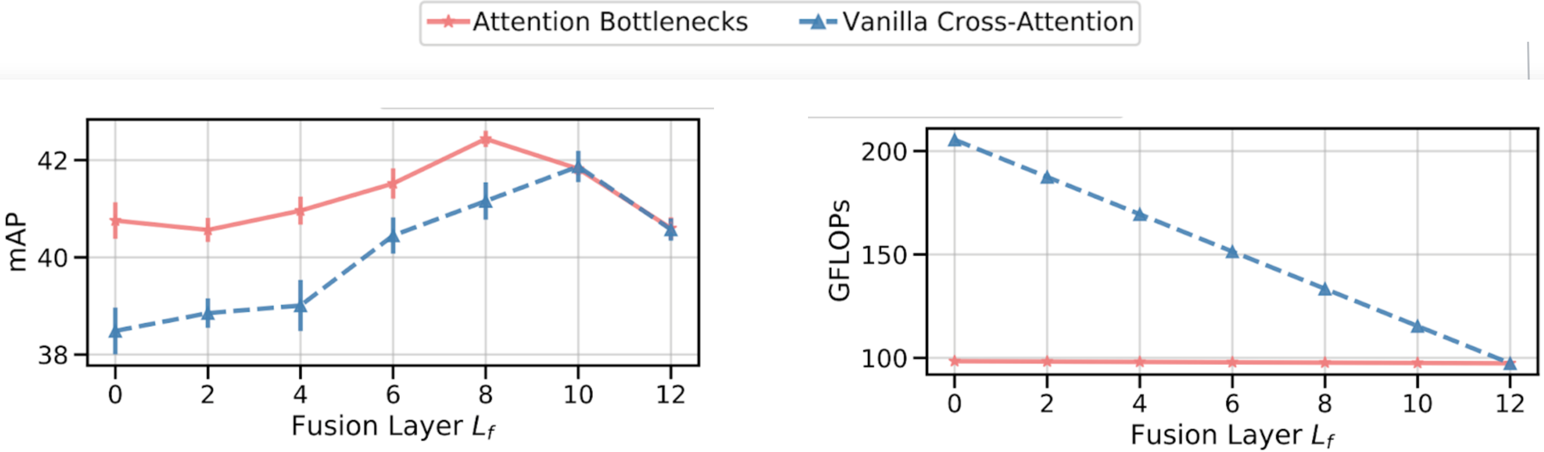
Do all layers need to be cross-modal?

- Restrict cross-modal information to later layers (mid-fusion)
- The layer we introduce cross-modal interactions is called the “fusion layer”
- Allows early layers to “specialise” to unimodal patterns



Improved performance and efficiency

- Mid Fusion outperforms early and late fusion on most datasets
- Using bottlenecks improves performance and reduces computational cost



Google Research

Results on 6 video classification datasets

Apply our model to two different video classification tasks

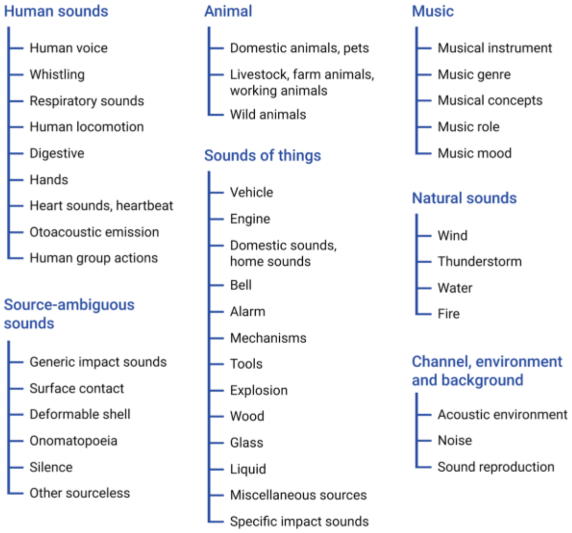
Action Recognition

Kinetics
Moments in Time

Epic Kitchens

Sound Event Classification

Audioset
VGGSound
Kinetics-Sounds



MBT: Attention Bottlenecks for Multimodal Fusion

State of the art performance

Model	Training Set	A only	V only	AV Fusion
GBlend [58]	MiniAS	29.1	22.1	37.8
GBlend [58]	FullAS-2M	32.4	18.8	41.8
Attn Audio-Visual [19]	FullAS-2M	38.4	25.7	46.2
Perceiver [29]	FullAS-2M	38.4	25.8	44.2
MBT	MiniAS	31.3	27.7	43.9
MBT	AS-500K	**44.3**	**32.3**	**52.1**

Table 1: **Comparison to the state of the art on AudioSet [22].** We report mean average precision (mAP). For audio-visual fusion, our method outperforms others that use the entire AudioSet training set (almost 2M samples), while we train on only 500K.

Model	Modalities	Verb	Noun	Action
Damen et al. [13]	A	42.1	21.5	14.8
AudioSlowFast [34]†	A	46.5	22.78	15.4
TSN [57]	V, F	60.2	46.0	33.2
TRN [63]	V, F	65.9	45.4	35.3
TBN [33]	A, V, F	66.0	47.2	36.7
TSM [42]	V, F	**67.9**	49.0	38.3
SlowFast [20]	V	65.6	50.0	38.5
MBT	A	44.3	22.4	13.0
MBT	V	62.0	56.4	40.7
MBT	A, V	64.8	**58.0**	**43.4**

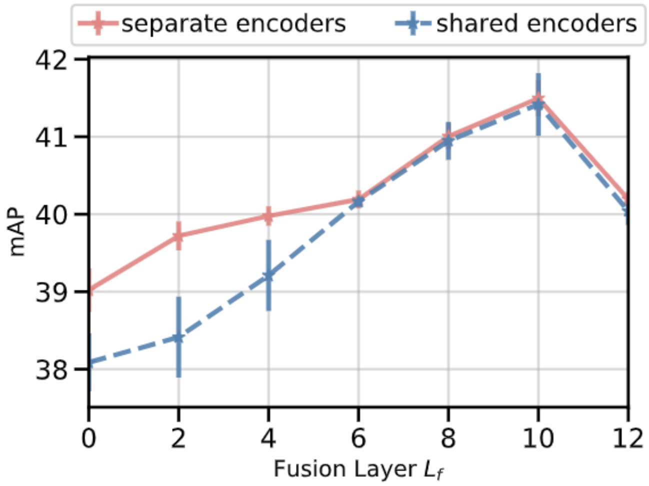
Table 2: **Comparison to the state of the art on Epic Kitchens 100 [13].** Modalities (Mods) are **A:** Audio, **V:** Visual, **F:** Optical flow.

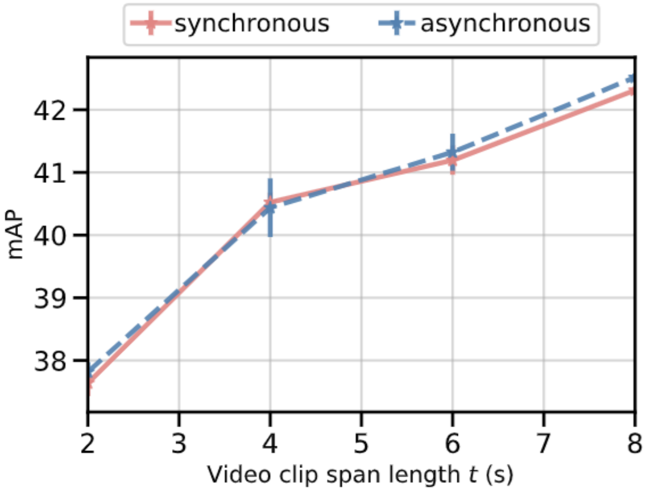
Model	Modalities	Top-1 Acc	Top-5 Acc
Chen et al‡ [11]	A	48.8	76.5
AudioSlowFast‡ [34]	A	50.1	77.9
MBT	A	52.3	78.1
MBT	V	51.2	72.6
MBT	A,V	**64.1**	**85.6**

Table 3: **Comparison to the state of the art on VGGSound [11].** Modalities are **A:** Audio, **V:** Visual, **F:** Optical flow. † Uses pretraining on VGGSound. ‡ We calculate metrics on our test set for a fair comparison using the scores provided by the authors.

Ablations

- ❖ For earlier fusion, separate weights for each modality is beneficial
- ❖ Asynchronous sampling provides a slight boost

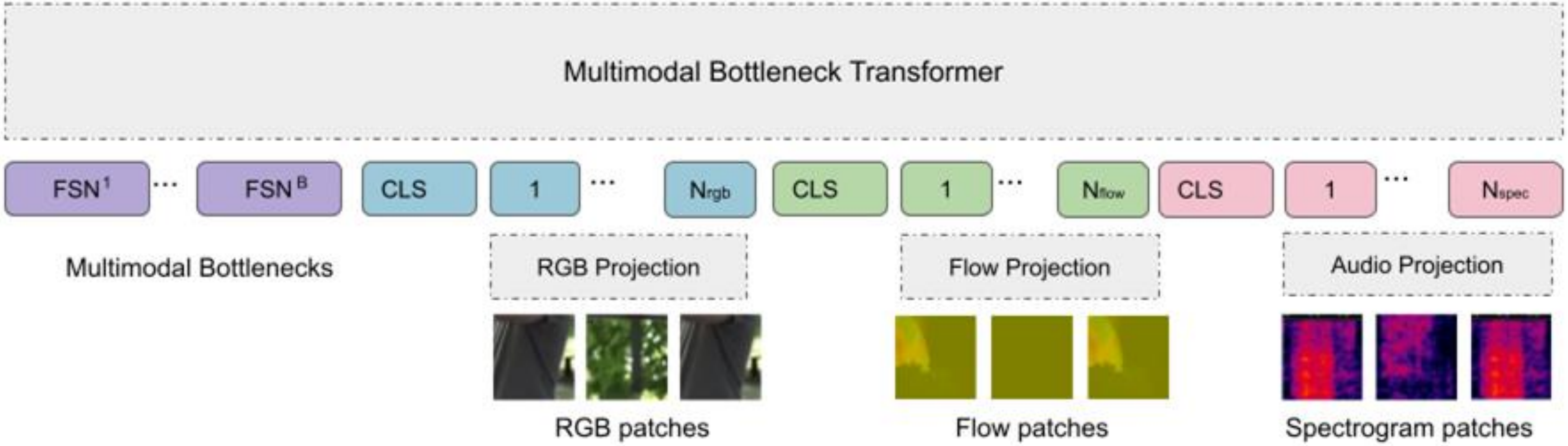




Google Research

More modalities?

- Our framework is general
- Can we used for any modality that can be tokenized
- Also can be used with any number of modalities
- So far we have added optical flow and are working on adding text

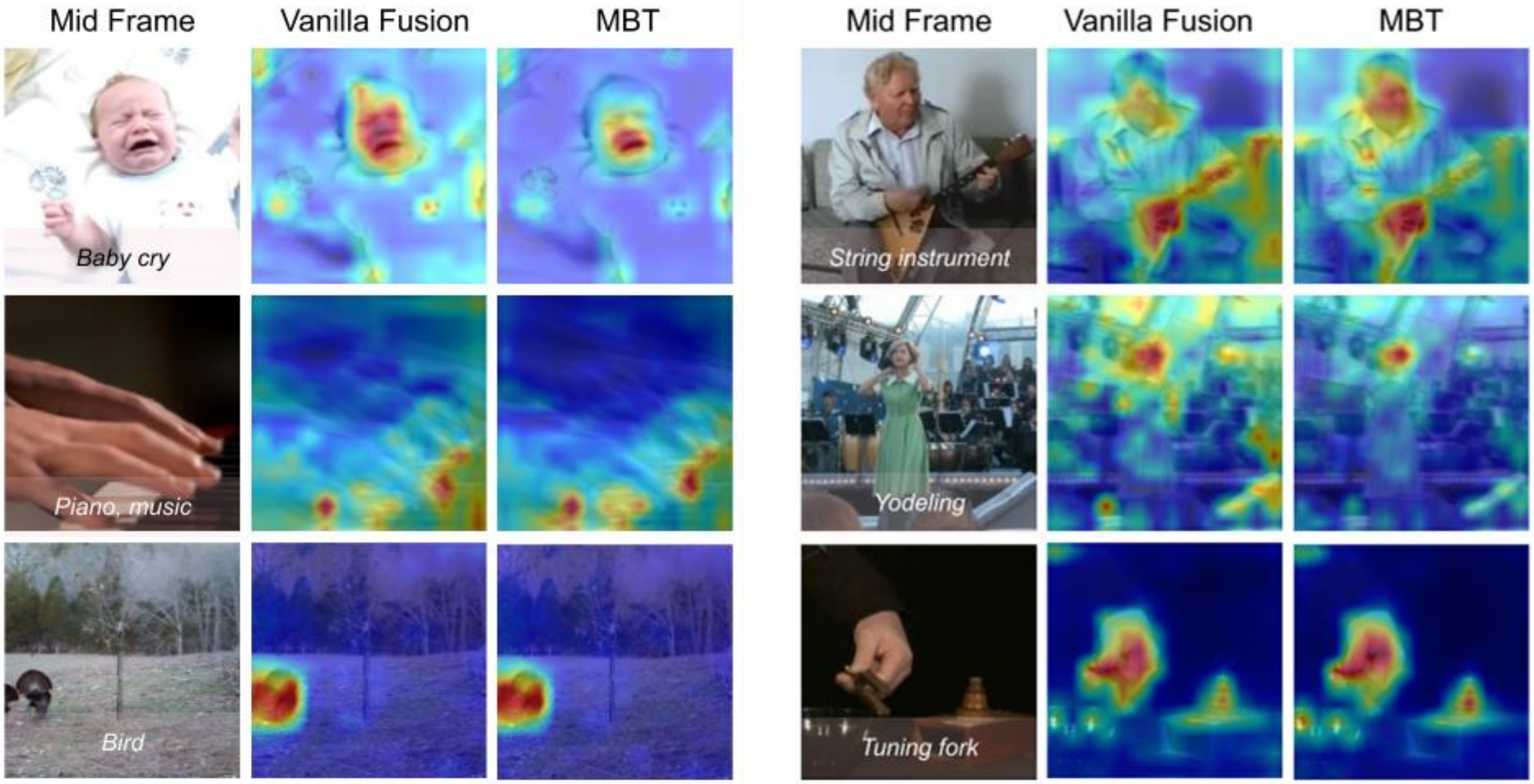


MBT: Attention Bottlenecks for Multimodal Fusion

Google Research

Attention Heatmaps

Focus on smaller regions, sound sources (mouth, fingertips)



Google Research

Conclusion

- Single transformer model for Multimodal Fusion
- Resources:
  - ArXiv, Webpage, Google AI blog
- Models are developed in JAX and FLAX.
  - We use the scenic codebase, code has been open-sourced and models released

Google Research

VideoCC: Learning Audio-Video Modalities from Image Captions

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, Cordelia Schmid,

Under submission, 2022

Why is paired video and text data so valuable?

- Natural language descriptions can be as detailed or as coarse as we like, no need to define a fixed label space
- Applications
  - Video captioning, video retrieval, videoQA etc
- From an AI perspective
  - Natural language (communicate), videos (perceive)
  - Bridge the gap between human communication and perception

"Person throws a pitch during a game against university"

Existing datasets

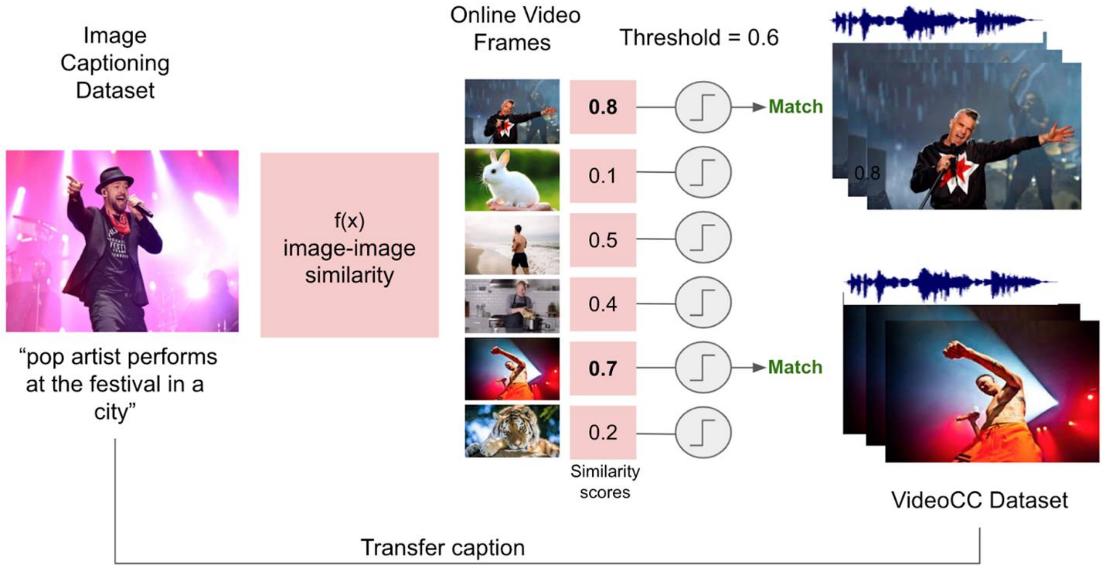
	Video - Text	*Audio - Text*
Manually Labelled Expensive, time-consuming, => small	ActivityNet-captions, MSR-VTT, MSVD, YouCook2, etc SpokenMiT	AudioCaps, CLOTHO
Semi-automatic/automatic Weak, noisy => require millions of samples to get good performance => text is not really a 'caption'	HowTo100M, WebVideoText, Instagram Hashtags,	*None*

Image captioning datasets, however, such as Conceptual Captions are large (millions), and relatively clean

Google Research

Transfer image captions to video and audio modalities

- Start with a seed image-captioning dataset
- Find frames in videos with high similarity scores to the seed image.
- Extract short video clips around the matching frames and transfer the caption



Google Research

Transfer image captions to video and audio modalities

- Use the Conceptual Captions 3M dataset as the image captioning seed dataset
- Image features extracted for YouTube frames at 1fps

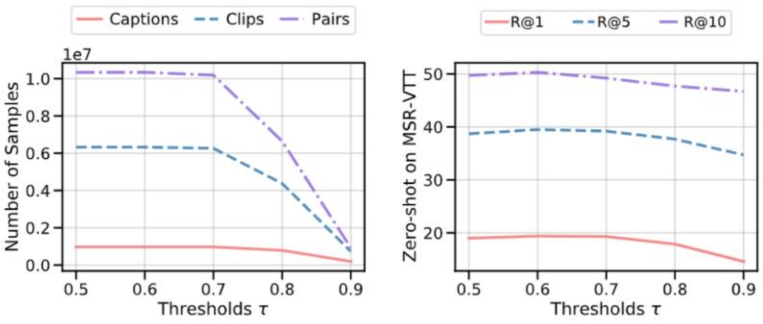


Figure 4. **Effect of match threshold τ** on mining statistics (left) and zero-shot performance on MSR-VTT (right). Increasing the threshold beyond 0.6 decreases the size of the dataset, which leads to a corresponding performance drop on zero-shot retrieval. We use an optimal match threshold of 0.6.

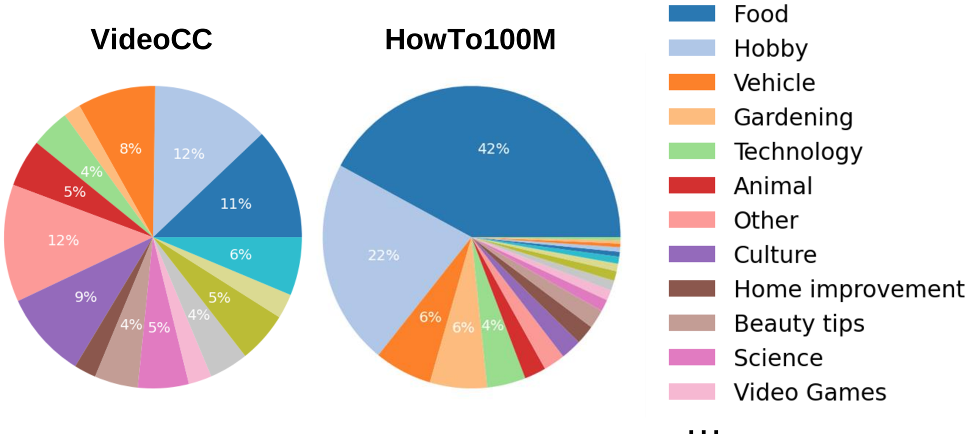
$t(s)$	3	5	10	20	30
MSR-VTT (ZS)	16.4	17.1	18.9	18.8	18.8

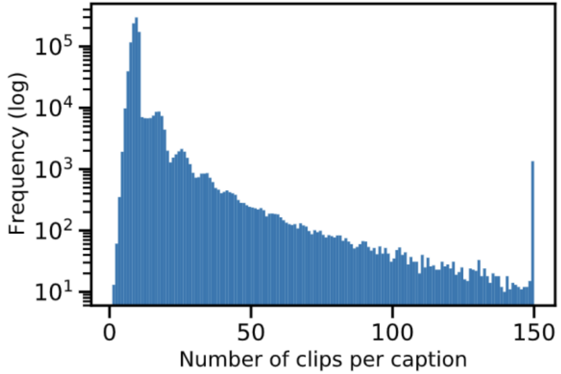
Table 8. **Temporal Span t of the mined clips.** We report zero-shot R@1 performance on the MSR-VTT dataset.

Google Research

VideoCC3M - Properties

- **VideoCC3M:**
  - **CC3M seed datasets. 10.3M** pairs, **6.3M** videos, **970K** unique captions
- **Alignment:** Highly likely that at least one frame is aligned to the caption
- **Less Specialised:** Multiple captions per video clip and multiple videos per caption
- **Diversity:** More balanced and diverse than HowTo100M
- **Multimodal:** Both audio and video (unlike WebVid-2M)
- **Fairness:** Filtered for fairness

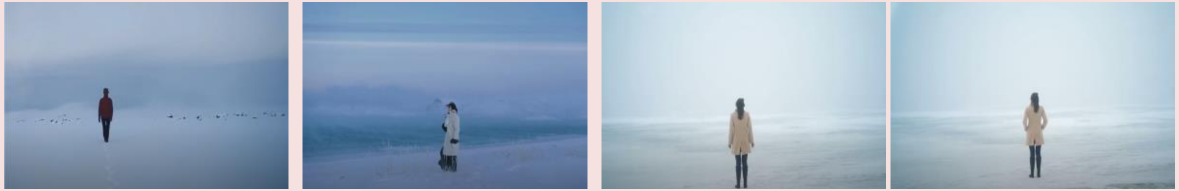




VideoCC3M - Examples

Caption	Seed Image	Mined Videos
“Person throws a pitch during a game against university”		
“Rap artist perform onstage during day at festival”		
“Sea anemone in a dark blue water of aquarium”		
“And this is a statue”		

VideoCC3M - Examples

Caption	Seed Image	Mined Videos
“A black bear in a forest”		
“Palm trees on a white background in cartoon style		
“A lonely figure stands on an endless icy field”		
“Welcome dinner table decor”		

Level of noise in the data

- Manual Study of 100 samples: 91/100 are relevant
    - 9 not relevant, 31 somewhat relevant, 60 highly relevant

Caption	Seed Image	Mined Videos		
“The robot playing electric guitar.”				
“Cricket player embraces cricket player on scoring the winning runs during the international cricket match				
“”The view of a red car blurred through broken glass				

Results - Video Retrieval

SOTA

Training on VideoCC3M outperforms training on HowTo100M with 20x less data

Pretraining Data	Modality	# Caps	R@1	R@5	R@10
Finetuned					
-	V	-	30.2	60.7	71.1
HowTo100M [49]	V	130M	33.1	62.3	72.3
VideoCC3M	V	970K	35.0	63.1	75.1
VideoCC3M	A+V	**970K**	**35.8**	**65.1**	**76.9**
Zero-shot					
HowTo100M [49]	V	130M	8.6	16.9	25.8
VideoCC3M	V	970K	18.9	37.5	47.1
VideoCC3M	A+V	**970K**	**19.4**	**39.5**	**50.3**

Table 2. **Effect of pretraining data on text-video retrieval for the MSR-VTT dataset. # Caps:** Number of unique captions. Training on VideoCC3M provides much better performance than Howto100M, with a fraction of the dataset size (VideoCC3M has only 970K captions and 6.3M clips compared to the 130M clips in HowTo100M) . The performance boost is particularly large for the zero-shot setting.

Method	Visual-Text PT	# Caps	R@1	R@5	R@10
Finetuned					
HERO [41]	HowTo100M	136M	16.8	43.4	57.7
NoiseEst. [5]	HowTo100M	136M	17.4	41.6	53.6
CE [44]†	-		20.9	48.8	62.4
UniVL [45]	HowTo100M	136M	21.2	49.6	63.1
ClipBERT [39]	Coco, VisGen	5.6M	22.0	46.8	59.9
AVLnet [61]	HowTo100M	136M	27.1	55.6	66.6
MMT [25]†	HowTo100M	136M	26.6	57.1	69.6
T2VLAD [73]†	-		29.5	59.0	70.1
Support Set [54]	HowTo100M	136M	30.1	58.5	69.3
VideoCLIP [74]	HowTo100M	136M	30.9	55.4	66.8
FIT [9]	CC3M	3M	25.5	54.5	66.1
FIT [9]	Multiple‡	6.1M	32.5	61.5	71.2
Ours	VideoCC3M	**970K**	**35.8**	**65.1**	**76.9**
Zero-shot					
MIL-NCE [49]	HowTo100M	136M	7.5	21.2	29.6
SupportSet [54]	HowTo100M	136M	8.7	23.0	31.1
VideoCLIP [74]	HowTo100M	136M	10.4	22.2	30.0
FIT [9]	WebVid2M*	2.5M	15.4	33.6	44.1
Ours	VideoCC3M	**970K**	**19.4**	**39.5**	**50.3**

Table 3. **Comparison to state-of-the-art results on MSR-VTT 1k-A split for text-to-video retrieval. Visual-Text PT:** Visual-text pretraining data. **# Caps:** Number of unique captions used during pretraining. † These works use numerous experts, including Object, Motion, Face, Scene, Speech, OCR and Sound classification features. ‡ Pretrained on WebVid-2M, CC3M and COCO datasets. *Numbers obtained from the authors.

Results - Video Captioning

- First results for zero-shot video captioning
- Outperforms HowTo100M by a large margin

Method	PT	Modality	B-4	C	M
Zero-shot					
Ours	HowTo100M	V	7.5	0.5	8.23
Ours	VideoCC3M	V	**13.23**	**8.24**	**11.34**

Table 4. **Results on the MSR-VTT dataset for video captioning.** Zero-shot results are obtained without any annotated video-text data. Modalities: **V:** RGB frames. **T:** ASR in videos.

GT:	a man is discussing the parts in an engine compartment in a vehicle	clouds are moving in the sky	this is about sports players making big plays during the game
HowTo100M:	So I'm going to go ahead and remove this	It's a great place to live and it's a great place to work.	I don't know if you can see that but there's a little bit of a gap in the middle of the field.
VideoCC3M:	the engine bay of an automobile model	clouds moving in the blue sky	american football player scores a touchdown against sports team

Google Research

Results - Audio Retrieval

- No audio supervision used at all
- Pre-training on HowTo100M gives poor zero-shot performance (speech)
- Pre-training on VideoCC3M gives a boost for both fine-tuning and zero-shot
- State of the art results on both AudioCaps and CLOTHO

“baby on a white blanket”

“person performs live with blues artist at festival”

“mirror image in a stream”

Model	Pretraining	Modality	R@1	R@10
SOTA [52]†	-	A	24.3	72.1
Ours	-	A	32.0	82.3
Ours	HowTo100M	A	33.7	83.2
Ours	VideoCC3M	A	35.5	84.5
Ours (ZS)	HowTo100M	A	1.4	6.5
Ours (ZS)	VideoCC3M	A	8.7	37.7
SOTA [52]†	-	A+V	28.1	79.0
Ours	-	A+V	41.4	85.3
Ours	VideoCC3M	A+V	**43.2**	**88.9**
Ours (ZS)	VideoCC3M	A+V	10.6	45.2

Table 5. **Results on the AudioCaps dataset for text-audio retrieval.** † Higher than reported in the paper, as these are provided by authors on our test set. Inputs refers to video inputs as follows: **A:** Audio spectrograms **V:** RGB video frames. Rows highlighted in light blue show Zero-shot (ZS) performance.

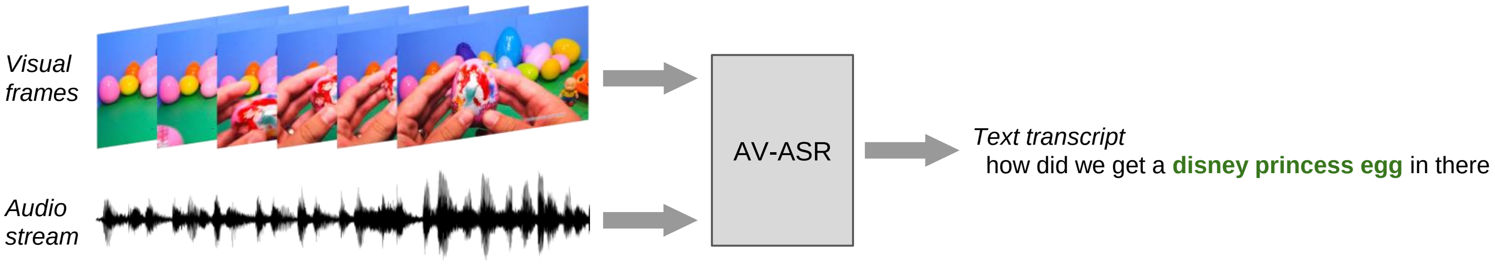
AVATAR: Unconstrained Audiovisual Speech Recognition

Valentin Gabeur*, Paul Hongsuck Seo*, Arsha Nagrani*, Chen Sun, Karteek Alahari, Cordelia Schmid

Google Research

Goal - Robust ASR

- Visual context (AV-ASR) can help with speech recognition
- When audio is noisy, corrupted etc.



Previous studies

Most AV-ASR works focus on using lip motion.

Fails

Speaker far away, blurry

Egocentric viewpoints

Face masks

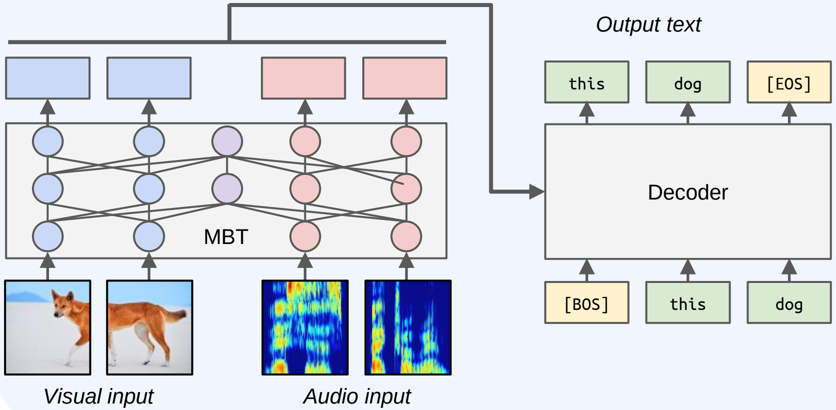
Lip motion is an obvious cue, but visual frames can also contain *objects, background info, actions* etc. that can help disambiguate

Google Research

AVATAR model and training

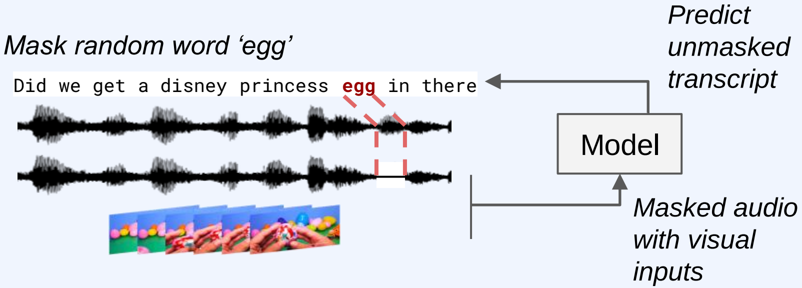
End-to-end trainable transformer with early modality fusion

- Early RGB + spectrogram fusion in the encoder.
- Trained from pixels directly



Novel training strategies based on word masking

- Prevent the audio stream from dominating training.
- Encourage the model to pay attention to the visual stream.



Evaluation

We evaluate with both **artificial** and **real** noise

- Artificial noise
  - Burst packet loss: randomly mask a chunk of the audio signal
  - Environment noise: add audio signals from AudioSet 'noise' category

Table 1: *Audiovisual ASR vs Audio only models under various evaluation noise conditions (Clean, Burst, Environment and Mixed) and with different training masking strategies (Random and Content). Percentage Word Error Rate (%WER) is reported on the How2 test set.* ***A:*** *Audio-only.* ***A+V:*** *Audiovisual.* ***Rel. Δ:*** *Relative improvement of A+V over A.*

Eval Noise / Training	Clean			Burst Loss			Environment Noise			Mixed Noise		
	A	**A+V**	**Rel. Δ**	**A**	**A+V**	**Rel. Δ**	**A**	**A+V**	**Rel. Δ**	**A**	**A+V**	**Rel. Δ**
No Pretraining	15.72	15.62	0.64%	29.59	28.69	3.05%	50.79	47.70	6.08%	60.51	57.49	5.0%
Vanilla	9.75	9.79	-0.33%	21.97	21.71	1.19%	25.97	25.55	1.61%	39.13	38.96	0.42%
Random Word Masking	9.19	9.11	0.93%	15.60	15.28	2.05%	23.39	22.35	4.45%	32.43	30.64	5.50%
Content Word Masking	9.58	9.25	3.48%	17.26	16.92	1.98%	23.77	22.67	4.65%	33.83	32.26	4.53%

Conclusions:

- Vision helps in all cases
- Masking strategies during training improve performance

Evaluation

We evaluate with both **artificial** and **real** noise

- Real world noise: we create a new test set called **VisSpeech**
  - Select challenging examples from YouTube where audio-ASR fails
  - Different accents, background sounds etc
  - Created from HowTo100M using a combination of automatic and manual techniques

1. Obtain videos with user-uploaded transcripts
2. Filter out cases where automatic ASR is perfect
3. Compute visual-text similarity and filter by thresholding
4. Manually correct user uploaded transcript

VisSpeech is available for download NOW at: https://gabeur.github.io/avatar-visspeech

Experimental Settings

HowTo100M
- Used for pretraining
- ~50M clips with their automatically-extracted speech transcriptions

How2
- Most widely used benchmark for unconstrained AV-ASR
- Each clip is accompanied by a user-uploaded (noisy) transcript
- To evaluate the use of visual stream, simulated noise is injected
  - *Burst packet loss*: randomly mask a chunk of the audio signal
  - *Environment noise*: add audio signals from AudioSet
- Train / val / test splits: 184,949 / 2,022 / 2,305 clips

VisSpeech
- 501 test examples in the wild with manually annotated transcripts.

Quantitative Results

Table 2: ***Comparison to the state-of-the-art on How2.*** *Our model outperforms all previous works when trained from scratch, and pretraining provides a significant boost. We report the best audio-visual numbers for all works.*

Model	**%WER**
BAS [10]	18.0
VAT [11]	18.0
MultiRes [17]	20.5
LLD [13]	16.7
AVATAR (scratch)	15.6
AVATAR (pretrained)	**9.1**

Table 3: *Results of AVATAR on our newly introduced test set Vis-Speech consisting of real-world noise. The models are trained on automatic ASR from HowTo100M, and finetuned on How2. Note here we do not add any artificial audio degradation at all.*

Training Strategy	**A**	**A+V**	**Rel. Δ**
No pretraining	51.70	49.73	3.81%
Vanilla	23.86	23.66	0.84%
Random Word Masking	22.13	21.08	4.78%
Content Word Masking	22.64	21.76	3.90%

Conclusions:

- Vision helps in all cases
- Masking strategies during training improve performance

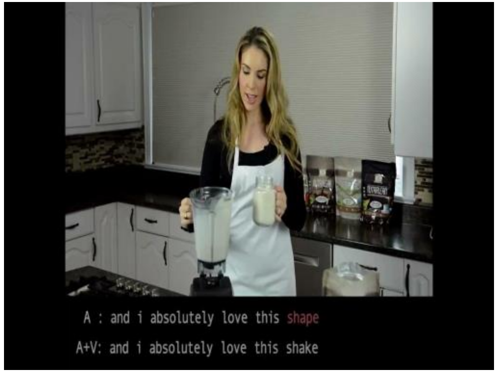
Google Research

Qualitative Results on VisSpeech

Visual context helps with objects ('eggplant', 'plane')

Google Research

Qualitative Results on VisSpeech



Visual context helps with objects (‘dessert’, ‘shake’, ‘coin’)

Google Research

Some more applications of fusion

Egocentric action recognition

- ❖ Audio is particularly useful in the egocentric (first person) domain
- ❖ Microphone is close to the person and may record sounds that are outside the view of the camera (eg. ‘eating’)
- ❖ Same object - different action – ‘wash steak’ vs ‘fry steak’

"With a Little Help from my Temporal Context: Multimodal Egocentric Action Recognition."
Kazakos, Evangelos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen.
BMVC (2021).

Google Research

Some more applications of fusion

Recognising chimpanzee behaviours in the wild

- Useful for conservation research
- Some actions like 'nut cracking', 'drumming' have distinct sounds
- Use an audio-visual CNN based model

Automated audiovisual behavior recognition in wild primates
Max Bain, Arsha Nagrani , Daniel Schofield, Sophie Berdugo, Joana Bessa, Jake Owens, Kimberley J. Hockings, Tetsuro Matsuzawa, Misato Hayashi, Dora Biro, Susana Carvalho, Andrew Zisserman
Science Advances, 2021

Challenges

- ❖ Different modalities learn at different rates
- ❖ Different input representations
    - ➢ Symbols, 1D waveform, 2D images, dense 3D point clouds
- ❖ Different noise topologies - how do we discard “irrelevant information?”
- ❖ Computational Complexity

- Our world is **multimodal** - it doesn't make sense to work with modalities in isolation
- Multimodal machine learning is an exciting area to do research in
- Transformers are a great flexible architecture for multimodal machine learning, can operate on any input that can be tokenized
- Audio can help action recognition and video retrieval
- Vision can be a an important cue for ASR

Thank you for listening! Questions?