Medical Vision Seminar

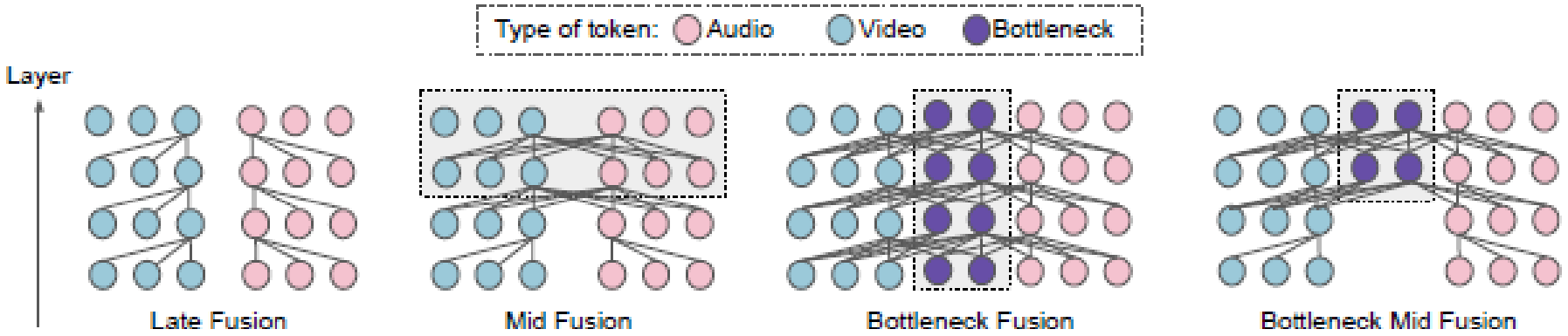
Lufei Gao

2021.12.29

- (NIPS2021) Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., & Sun, C. (2021). **Attention bottlenecks for multimodal fusion.** *Advances in Neural Information Processing Systems*, *34*.
- (ICLR2021) Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). **An image is worth 16x16 words: Transformers for image recognition at scale.** *arXiv preprint arXiv:2010.11929*.
- (Interspeech2021) Gong, Y., Chung, Y. A., & Glass, J. (2021). **AST: Audio Spectrogram Transformer.** *arXiv preprint arXiv:2104.01778*.

Challenges

- (i) variations in learning dynamics between modalities,
- (ii) different noise topologies, with some modality streams containing more information for the task at hand than others,
- (iii) specialized input representations: audio vs. vision
- Usually using late-fusion for simplicity



ViT and AST architectures

- Extract N nonoverlapping patches from the RGB image (or the audio spectrogram), $x_i \in \mathbb{R}^{h \times w}$, and convert them into a series of 1D tokens $z_i \in \mathbb{R}^d$:

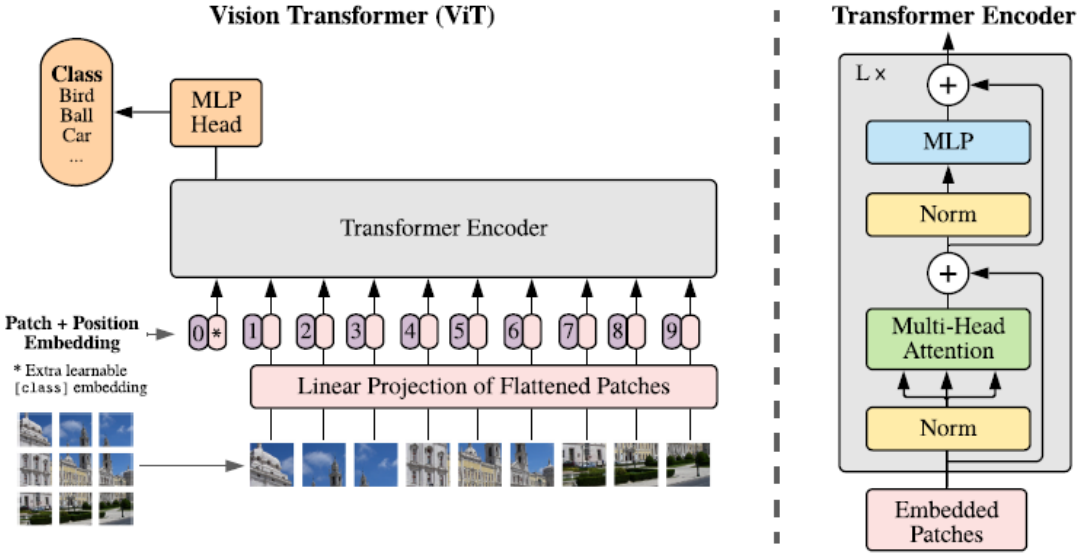
$$\mathbf{z} = g(\mathbf{x}; \mathbf{E}, z_{\text{cls}}) = [z_{\text{cls}}, \mathbf{E}x_1, \mathbf{E}x_2, ..., \mathbf{E}x_N] + \mathbf{p}.$$

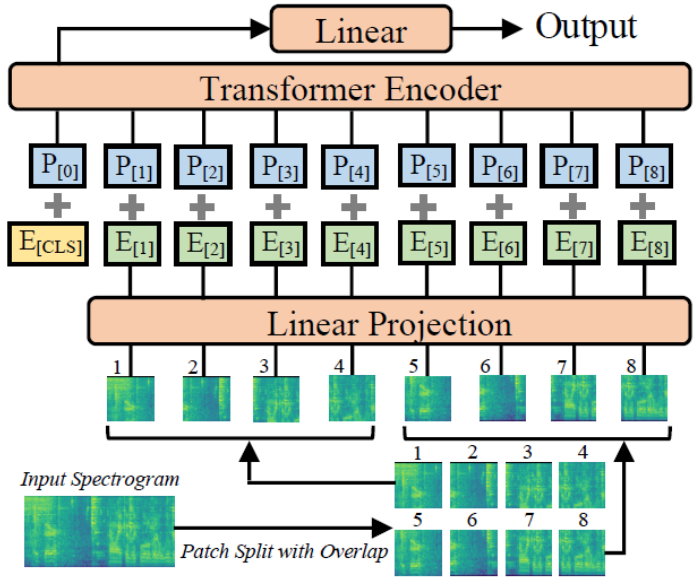
- Pass through an encoder (sequence of L transformer layers):
- Each transformer layer: Multi-Headed Self-Attention (MSA), Layer Normalization (LN) and Multilayer Perceptron (MLP) blocks applied using residual connections.

$$\mathbf{z}^{l+1} = \text{Transformer}(\mathbf{z}^l):$$

$$\mathbf{y}^l = \text{MSA}(\text{LN}(\mathbf{z}^l)) + \mathbf{z}^l$$
$$\mathbf{z}^{l+1} = \text{MLP}(\text{LN}(\mathbf{y}^l)) + \mathbf{y}^l.$$

- MSA operation: dot-product attention: $\text{MSA}(\mathbf{X}) = \text{Attention}(\mathbf{W}^{\text{Q}}\mathbf{X}, \mathbf{W}^{K}\mathbf{X}, \mathbf{W}^{\text{V}}\mathbf{X})$
- MCA (Multi-Headed Cross Attention: $\text{MCA}(\mathbf{X}, \mathbf{Y}) = \text{Attention}(\mathbf{W}^{\text{Q}}\mathbf{X}, \mathbf{W}^{K}\boldsymbol{Y}, \mathbf{W}^{\text{V}}\mathbf{Y})$





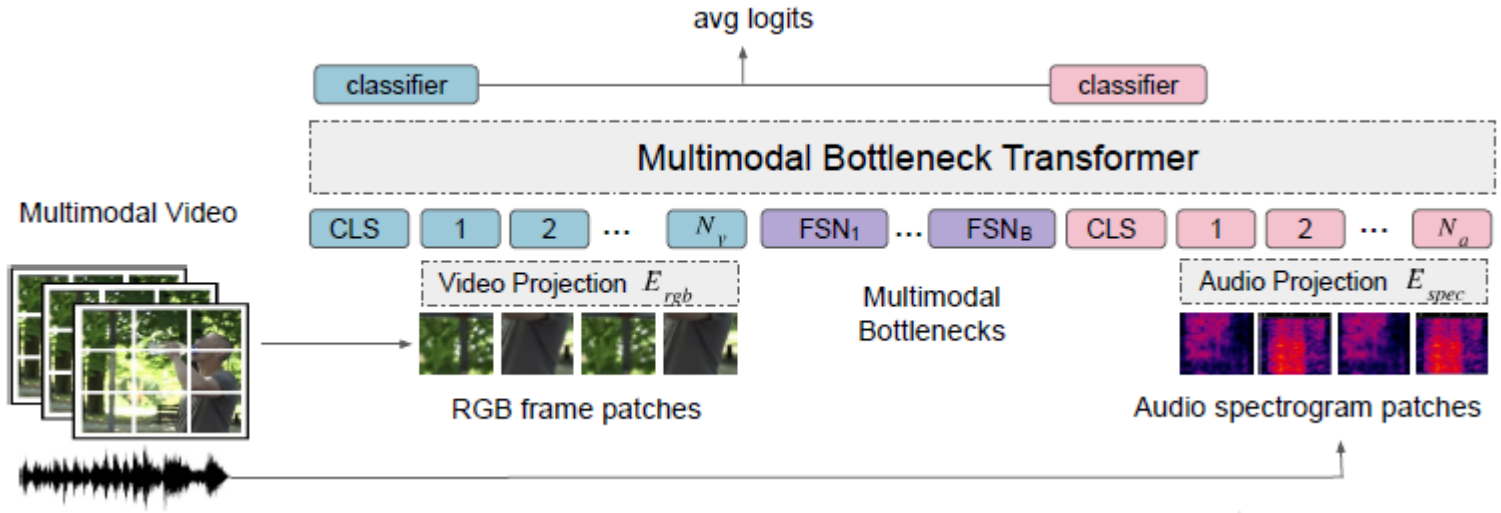
Multimodal transformer

- Fusion via vanilla self-attention
  - Tokenizing video:
    - Given a video clip of length t seconds, uniformly sample F RGB frames and convert the audio waveform into a single spectrogram.
    - Then embed each frame and the spectrogram independently following the encoding proposed in ViT,
    - Concatenate all tokens together into a single sequence.

  $\mathbf{x}_{\text{rgb}} \in \mathbb{R}^{N_v \times d}$, $\mathbf{x}_{\text{spec}} \in \mathbb{R}^{N_a \times d}$, $\mathbf{z}_{\text{rgb}} = g\left(\mathbf{x}_{\text{rgb}}; \mathbf{E}_{\text{rgb}}, z_{\text{cls-rgb}}\right), \mathbf{z}_{\text{spec}} = g\left(\mathbf{x}_{\text{spec}}; \mathbf{E}_{\text{spec}}, z_{\text{cls-spec}}\right)$

  - Sequence of tokens is :$\mathbf{z} = \left[\mathbf{z}_{\text{rgb}} \parallel \mathbf{z}_{\text{spec}}\right]$
  - Attention is allowed to flow freely through the network, i.e. each RGB token can attend to all other RGB and spectrogram tokens as follows: $\mathbf{z}^{l+1} = \text{Transformer}\left(\mathbf{z}^{l}; \theta\right)$ with model parameters θ.
- Cross-Transformer Layer

$$
\begin{aligned}
\mathbf{z}_{\text{rgb}}^{l+1} &= \text{Cross-Transformer}(\mathbf{z}_{\text{rgb}}^{l}, \mathbf{z}^{l}; \theta_{\text{rgb}}) \\
\mathbf{z}_{\text{spec}}^{l+1} &= \text{Cross-Transformer}(\mathbf{z}_{\text{spec}}^{l}, \mathbf{z}^{l}; \theta_{\text{spec}}), \\
\mathbf{y}^{l} &= \text{MCA}(\text{LN}(\mathbf{z}_{1}^{l}), \text{LN}(\mathbf{z}_{2}^{l})) + \mathbf{z}_{1}^{l}.
\end{aligned}
$$



- Fusion via attention bottlenecks
  - B fusion bottleneck tokens $\mathbf{z}_{\text{fsn}} = \left[\mathbf{z}_{\text{fsn}}^1, \mathbf{z}_{\text{fsn}}^2, \ldots, \mathbf{z}_{\text{fsn}}^B\right]$ (B ≪ Nv, B ≪ Na)
  - Input sequence: $\mathrm{z} = \left[\mathbf{z}_{\text{rgb}} \parallel \mathbf{z}_{\text{fsn}} \parallel \mathbf{z}_{\text{spec}}\right]$
  - Restrict all cross-modal attention flow in our model to be via these bottleneck tokens.
  - For layer l, token representation:

$$[\mathrm{z}_{\mathrm{i}}^{l+1} || \hat{\mathrm{z}}_{\mathrm{fsn}_i}^{l+1}] = \mathrm{Transformer}([\mathrm{z}_{\mathrm{i}}^{l} || \mathrm{z}_{\mathrm{fsn}}^{l}]; \theta_{\mathrm{i}}) \quad (8)$$

$$\mathrm{z}_{\mathrm{fsn}}^{l+1} = \mathrm{Avg}_{\mathrm{i}}(\hat{\mathrm{z}}_{\mathrm{fsn}_i}^{l+1}) \quad (9)$$

Create modality specific temporary bottleneck fusion tokens $\hat{\mathbf{z}}_{\text{fsn}}^{i}$, which are updated separately and simultaneously with audio and visual information (Equation 8).
The final fusion tokens from each cross-modal update are then averaged in Equation 9.

$$\mathrm{z}_{\mathrm{rgb}}^{l+1} = \mathrm{Transformer}(\mathrm{z}_{\mathrm{rgb}}^{l}; \theta_{\mathrm{rgb}}), \mathrm{z}_{\mathrm{spec}}^{l+1} = \mathrm{Transformer}(\mathrm{z}_{\mathrm{spec}}^{l}; \theta_{\mathrm{spec}}) \quad \text{if } l < L_f$$

$$\mathrm{z}^{l} = [\mathrm{z}_{\mathrm{rgb}}^{l} || \mathrm{z}_{\mathrm{spec}}^{l}], \quad \mathrm{z}^{l+1} = \text{Multimodal-Transformer}(\mathrm{z}^{l}; \theta_{\mathrm{spec}}, \theta_{\mathrm{rgb}}) \quad \text{otherwise}$$

'early-fusion' model: $L_f = 0$;
'late-fusion' model: $L_f = L$;
'mid-fusion' model: $0 < L_f < L$

Exp: Datasets

- 3 video classification datasets:
  - AudioSet
    - Almost 2 million 10-second video clips from YouTube, annotated with 527 classes.
    - 20,361 clips for the balanced train set and 18,589 clips for the test set.
    - Train on a (slightly more) balanced subset consisting of 500K samples (AS-500K). with a binary cross-entropy (BCE) loss
    - mean average precision (mAP) over all classes
  - Epic-Kitchens-100
    - 90,000 variable length clips spanning 100 hours
    - Actions are mainly short-term (average length is 2.6s with minimum length 0.25s).
    - Predict verb and noun in each action label using a single network using two “heads”, both trained with a CE loss.
    - TOP-1 action accuracy
  - VGGSound
    - Almost 200K video clips of length 10s, annotated with 309 sound classes consisting of human actions, sound-emitting objects and human-object interactions
    - 172,427 training and 14,448 test clips
    - Standard CE loss for classfiication
    - TOP-1 and TOP-5 classification accuracy

Exp: Set

- ViT-Base (#transformer layers $L = 12$, #self-attention heads $N_H = 2$, with #hidden dimension $d = 3072$,) initialized from ImageNet-21K
- B=4
  - Bottleneck tokens are initialized using a Gaussian with mean of 0 and standard deviation of 0.02
- Video: randomly sample clips of t seconds for training.
  - Extract RGB frames for all datasets at 25 fps.
  - For AudioSet and VGGSound
    - Sample 8 RGB frames over the sampling window of length t with a uniform stride of length (t × 25)/8.
    - Extract 16 × 16 patches from each frame of size 224 × 224, giving a total of 8 × 14 × 14 = 1568 patches per video.
  - For Epic-Kitchens
    - Sample 32 frames with stride 1.
- Audio: sampled at 16kHz and converted to mono channel.
  - Extract log-mel spec. with a freq. dim. of 128 (25ms Hamming window with hop length 10ms).
  - Input of size 128 × 100t for t seconds of audio.
  - Extract spectrogram patches with size 16 × 16, giving 8 × 50 = 400 patches for 8 seconds of audio.
- Data augmentation:
  - For images: standard data augmentations (random crop, flip, colour jitter)
  - for specs: SpecAugment with a max time mask length of 192 frames and max frequency mask length of 48 bins following AST
- Mixup with α = 0.3 and stochastic depth regularization with probability p = 0.3.
- Batch size of 64, synchronous SGD with momentum of 0.9, cosine learning rate schedule with warmup of 2.5 epochs on TPU accelerators.
- Base learning rate to 0.5; train for 50 epochs.

Exp: Fusion Strategies

- 3 strategies:
  1. Vanilla self-attention
  2. Vanilla cross-attention with separate weights
  3. Bottleneck fusion
- Investigate:
  - Impact of sharing the encoder weights for both modalities ((1) vs (2)) —— Fig1: use separate modal weights
  - Impact of varying the fusion layer $L_f(0,2,4,6,8,10,12)$, for the latter two strategies——Fig3: mid>early, later
  - Effect of bottleneck attention vs vanilla cross-attention for multimodal fusion.——GFLOPS
  - Number of bottleneck tokens B: B=4.
  - The impact of different modality sampling strategies

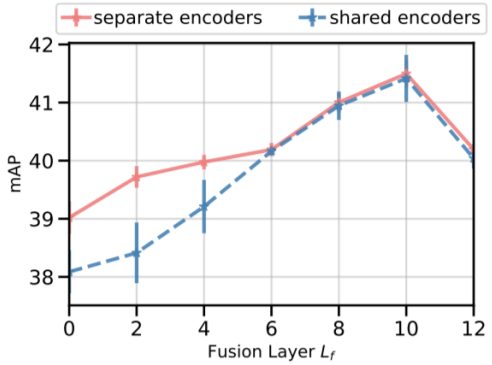


Figure 1: The effect of sharing weights for vanilla fusion.

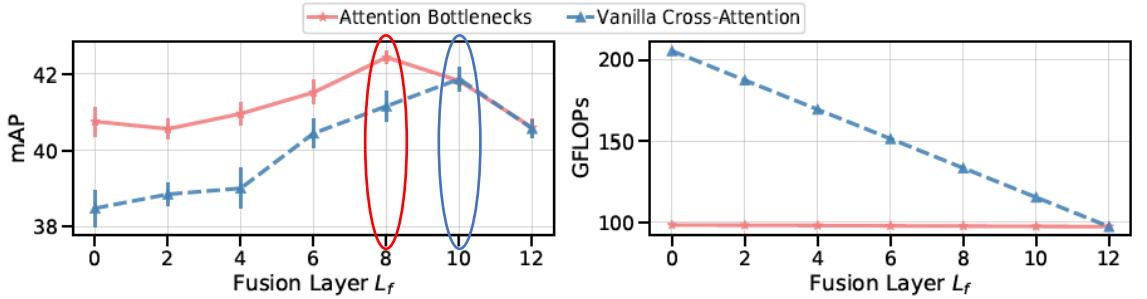


Figure 3: The impact of using attention bottlenecks for fusion on performance (left) and compute (right) at different fusion layers L_f on AudioSet, using clip span $t = 4$ and $B = 4$ bottleneck tokens. Attention bottlenecks improve performance at lower computational cost.

Exp: Fusion Strategies

- The impact of different modality sampling strategies
  - Sampling window size t: Fig4
  - Synchronous vs Asynchronous sampling: Fig2
  - Modality MixUp
    - Standard mixup
    - Modality mixup: 42.6 mAP →43.9 mAP
  - Impact of Dataset Size: Fig5

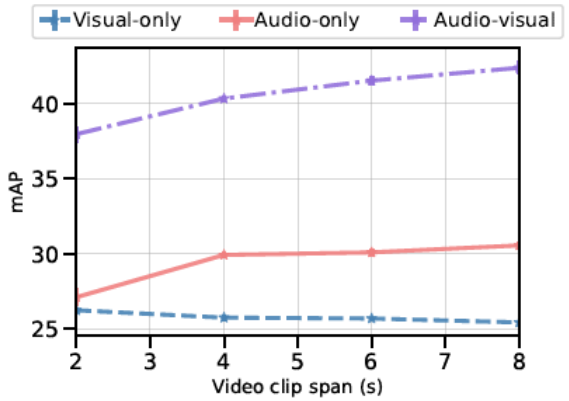


Figure 4: The effect of varying input clip span t on the AudioSet test set.

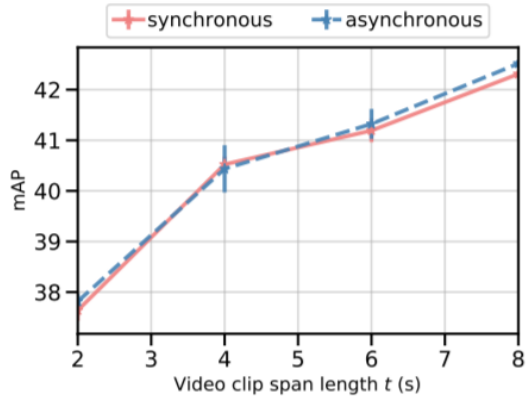


Figure 2: Asynchronous vs synchronous sampling of RGB and spectrogram inputs.

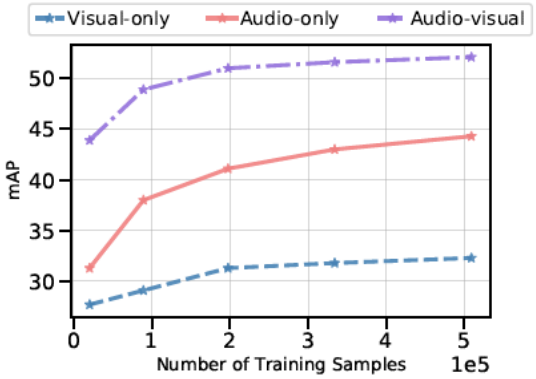


Figure 5: The effect of training data size on the AudioSet test set.

Model	Training Set	A only	V only	AV Fusion
GBlend [56]	MiniAS	29.1	22.1	37.8
GBlend [56]	FullAS-2M	32.4	18.8	41.8
Attn Audio-Visual [18]	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 SOTA on AudioSet [21].** We report mean average precision (mAP). We outperform works that train on the full Audioset (2M samples), while we train on only 500K samples.

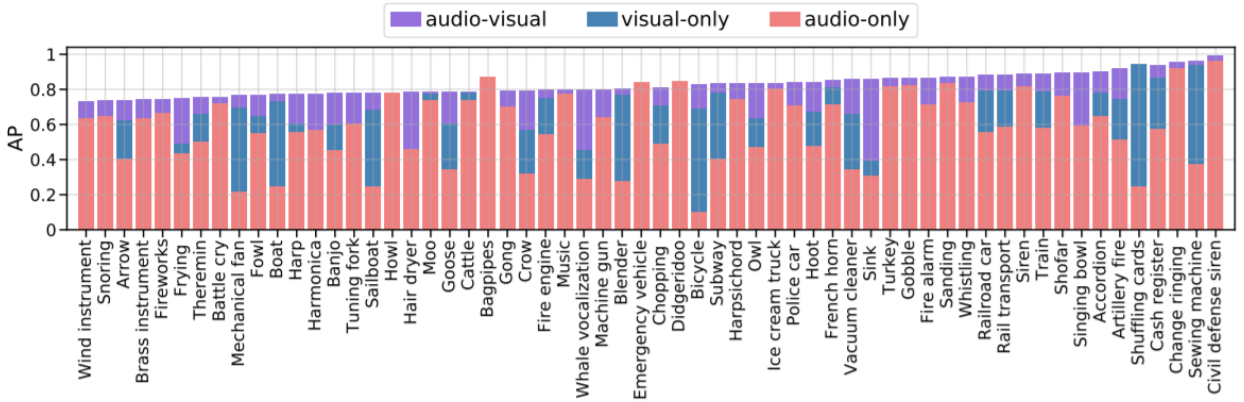


Figure 3: Per-class average precision for the top 60 classes in Audioset ranked by mAP. Best viewed in colour and zoomed in. Note how audio-visual fusion helps improve performance over audio only for almost all classes. The visual only model performs well for classes that have a stronger visual signature than audio, eg 'bicycle', 'mechanical fan', 'boat' and 'arrow'.

Worse: 'bagpiping', 'emergency vehicle' and 'didgeridoo' which have **strong audio signatures**.

Extremely better: 'bicycle' and 'shuffling cards' where **audio signals are weaker**

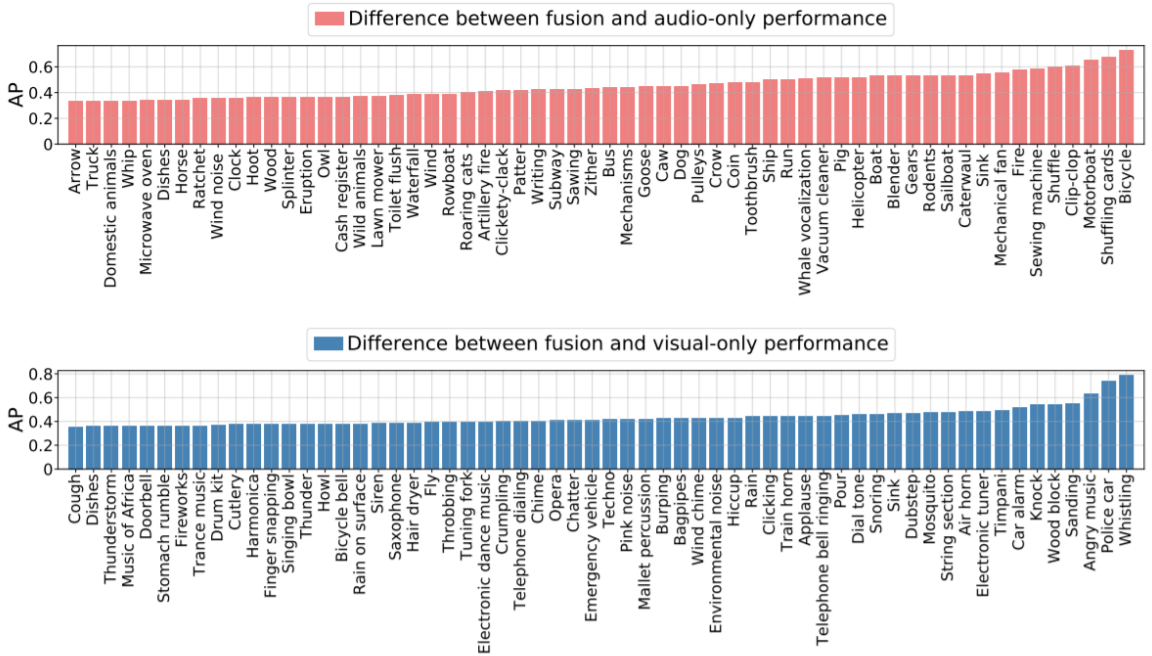


Figure 4: Top 60 classes that have the highest gain with fusion over a audio only (top) and visual only (bottom) baseline. Note how fusion improves the per class AP for certain classes by over 50% over a unimodal model. As expected, the classes that benefit most from visual information are 'bicycle' and 'shuffling cards' and the class that benefits most from audio is 'Whistling'.

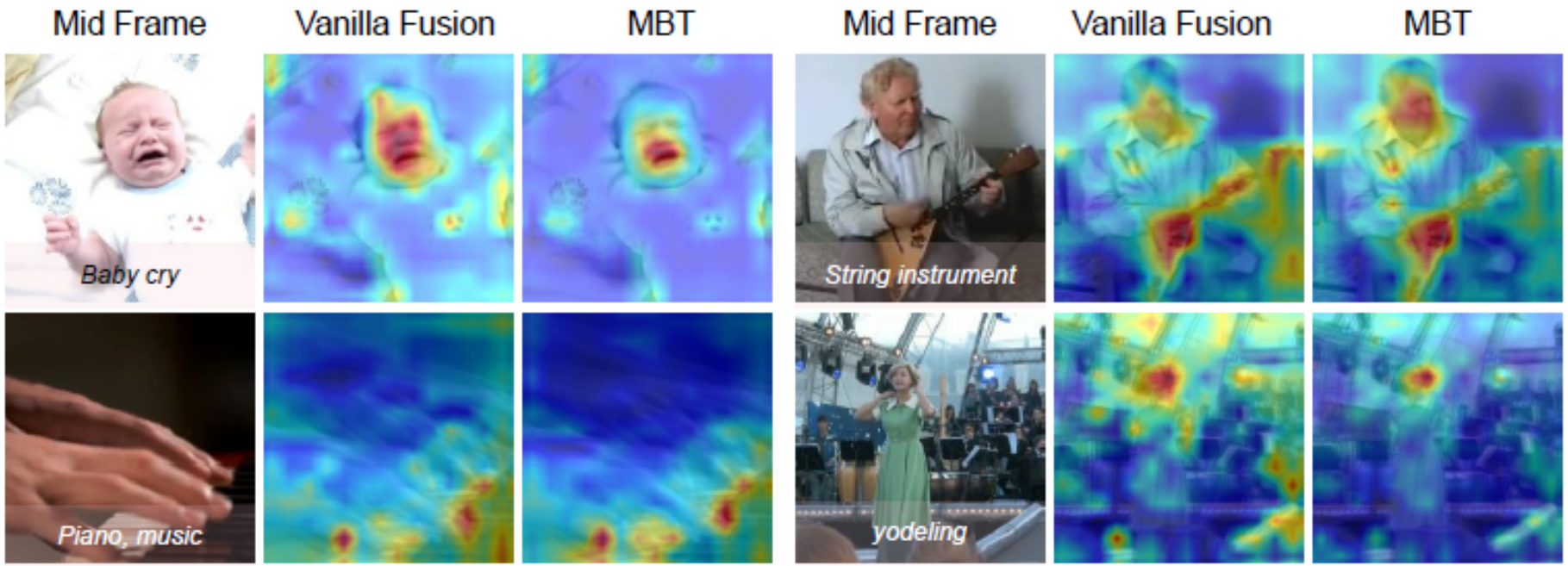


Figure 6: **Attention Maps.** We compute maps of the attention from the output CLS tokens to the RGB image input space for a vanilla self-attention model and MBT on the Audioset test set. For each video clip, we show the original middle frame on the left with the ground truth labels overlayed at the bottom. The attention is particularly focused on sound source regions in the video that contain motion, eg. the fingertips on the piano, the hands on the string instrument, faces of humans. The bottlenecks in MBT further force the attention to be localised to smaller regions of the images (i.e the mouth of the baby on the top left and the mouth of the woman singing on the bottom right).

The model focuses on semantically salient regions in the video for audio classification, particularly regions where there is motion that creates or modifies sound.

The tight bottlenecks do force the model **to focus only on the image patches that are actually relevant for the audio classification task** and which benefit from early fusion with audio.