Interspeech 2022 – Multimodal

Renmiao Chen

2022/12/23

Catalog

1. Context-aware Multimodal Fusion for Emotion Recognition
* The Chinese University of Hong Kong and Lightspeed & Quantum Studios

2. Robust Self-Supervised Audio-Visual Speech Recognition
* Toyota Technological Institute at Chicago and Meta AI

3. Audio-Visual Scene Classification Based on Multi-modal Graph Fusion
* East China University of Science and Technology

4. Interactive Co-Learning with Cross-Modal Transformer for Audio-Visual Emotion Recognition
* NTT Corporation

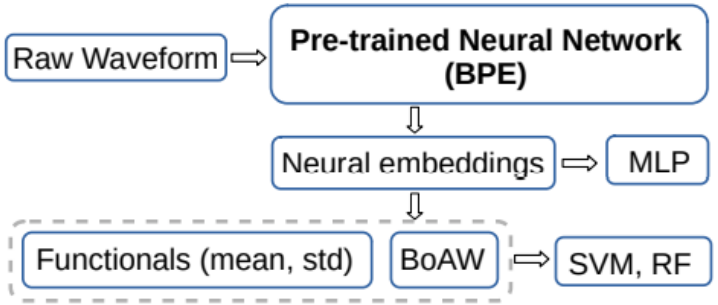
5. Perceiver: General Perception with Iterative Attention
* DeepMind

6. Perceiver IO: A general architecture for structured inputs & outputs
* DeepMind

1. On Breathing Pattern Information in Synthetic Speech

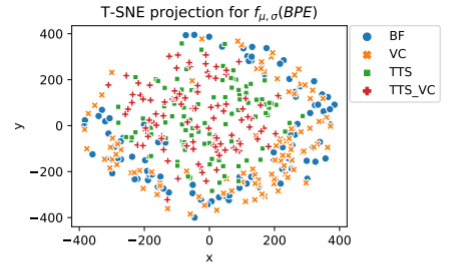
* Idiap Research Institute and Ecole polytechnique fed´ erale de Lausanne

- investigate whether synthetic speech carries breathing pattern related information in the same way as natural human speech

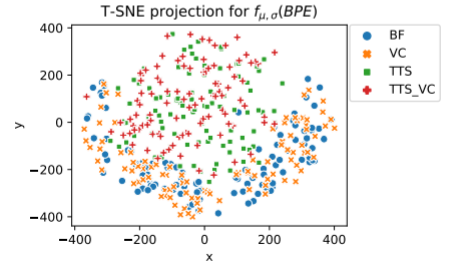


Features	Classifier	Measure	All	VC	TTS	TTS_VC
Embeddings from CNN pre-trained on Philips database						
3 seconds speech input						
BPE	MLP	AUC	90.35	59.92	99.4	99.7
		EER	16.88	42.75	2.53	1.32
$f_{\mu\sigma}$(BPE)	SVM	AUC	89.51	59.42	98.42	98.78
		EER	16.98	43.54	4.29	3.48
	RF	AUC	90.65	62.44	98.93	99.54
		EER	17.02	41.02	4.22	2.6
BoAW(BPE)	SVM	AUC	89.35	61.43	97.54	98.17
		EER	17.69	41.62	7	6.01
	RF	AUC	90.86	62.72	99.16	99.62
		EER	17.69	40.04	4.14	2.5
2 seconds speech input						
BPE	MLP	AUC	84.56	47.94	95.08	96.63
		EER	21.5	51.59	10.89	8.72
$f_{\mu\sigma}$(BPE)	SVM	AUC	87.52	57.92	95.85	97.7
		EER	20.08	44.39	10.63	7.49
	RF	AUC	89.15	56.68	98.61	99.55
		EER	18.23	45.41	5.25	2.7
BoAW(BPE)	SVM	AUC	88.18	52.17	98.91	99.18
		EER	19.28	48.57	4.24	2.97
	RF	AUC	88.04	51.14	99.01	99.34
		EER	19.51	48.46	4.61	3.2
Embeddings from CNN pre-trained on UCL_SBM database						
3 seconds speech input						
BPE	MLP	AUC	90.02	58.33	99.43	99.73
		EER	17.27	43.65	2.56	1.65
$f_{\mu\sigma}$(BPE)	SVM	AUC	90.23	59.86	99.2	99.66
		EER	17.48	42.96	3.62	2.35
	RF	AUC	90.76	60.85	99.64	99.93
		EER	17.29	41.77	1.6	0.59
BoAW(BPE)	SVM	AUC	90.11	58.32	99.58	99.8
		EER	17.29	44.44	2.75	1.74
	RF	AUC	90.5	60.11	99.49	99.9
		EER	17.07	42.62	2.46	0.63
2 seconds speech input						
BPE	MLP	AUC	89.84	60.24	98.42	99.41
		EER	17.48	43.14	5.56	3.08
$f_{\mu\sigma}$(BPE)	SVM	AUC	88.07	58.28	97.25	96.45
		EER	18.97	43.92	8.67	9.44
	RF	AUC	88.48	54.97	98.55	98.46
		EER	18.71	47.1	5.85	6.25
BoAW(BPE)	SVM	AUC	89.25	55.91	99.21	99.34
		EER	17.69	45.79	3.53	3.03
	RF	AUC	90.01	57.92	99.62	99.7
		EER	17.19	43.72	2.49	2.44

	Attack type	Our study AUC	Our study EER	ASVSpoof2019 EER [21]
A07	TTS	99.75	1.48 [0.38 - 5.39]	0.02
A08	TTS	98.65	4.76 [2.38 - 7.55]	0.09
A09	TTS	99.3	3.31 [0.45 - 12.36]	0.06
A10	TTS	99.74	1.39 [0.74 - 5.44]	12.21
A11	TTS	99.68	1.64 [0.79 - 5.35]	0.59
A12	TTS	98.85	4.6 [1.6 - 10.13]	3.75
A13	TTS_VC	99.75	1.31 [0.31 - 7.22]	12.41
A14	TTS_VC	99.67	2.13 [0.73 - 4.82]	2.88
A15	TTS_VC	99.68	1.79 [0.53 - 4.78]	3.22
A16	TTS	99.32	3.27 [1.69 - 5.37]	0.02
A17	VC	52.96	48.07 [42.93 - 51.22]	15.93
A18	VC	61.82	41.51 [38.9 - 45.38]	5.59
A19	VC	65.8	38.25 [35.66 - 39.74]	0.06



(a) *Philips*



(b) *UCL_SBM*

- irrespective of the TTS approach tends to not carry breathing pattern related information in the same way as natural human speech.
- the alterations done to the natural human speech signal during voice conversion is not strongly altering the breathing pattern related information.

2. Does Audio Deepfake Detection Generalize?

* Fraunhofer AISEC and Technical University Munich and why do birds GmbH

- Which factors contribute to success, and which are accidental for audio deepfake detection

1. that average EER on ASVspoof drops from 19.89% to 9.85% when the full-length input is used. four-second clip is insufficient for the model to extract useful information compared to using the full audio file as input. (the numerous works that use fixed-length inputs suggest otherwise.)
2. The 'raw' models outperform the feature-based models
3. Simply replacing melspec with cqtspec increases the average performance by 37%, all other factors constant.

- Evaluate generalization capabilities
  - Often, the models do not perform better than random guessing

Model Name	Feature Type	Input Length	ASVspoof19 eval EER%	ASVspoof19 eval t-DCF	In-the-Wild Data EER%
LCNN	cqtspec	Full	**6.354±0.39**	0.174±0.03	**65.559±11.14**
LCNN	cqtspec	4s	25.534±0.10	0.512±0.00	70.015±4.74
LCNN	logspec	Full	7.537±0.42	**0.141±0.02**	72.515±2.15
LCNN	logspec	4s	22.271±2.36	0.377±0.01	91.110±2.17
LCNN	melspec	Full	15.093±2.73	0.428±0.05	70.311±2.15
LCNN	melspec	4s	30.258±3.38	0.503±0.04	81.942±3.50
LCNN-Attention	cqtspec	Full	**6.762±0.27**	**0.178±0.01**	**66.684±1.08**
LCNN-Attention	cqtspec	4s	23.228±3.98	0.468±0.06	75.317±8.25
LCNN-Attention	logspec	Full	7.888±0.57	0.180±0.05	77.122±4.91
LCNN-Attention	logspec	4s	14.958±2.37	0.354±0.03	80.651±6.14
LCNN-Attention	melspec	Full	13.487±5.59	0.374±0.14	70.986±9.73
LCNN-Attention	melspec	4s	19.534±2.57	0.449±0.02	85.118±1.01
LCNN-LSTM	cqtspec	Full	**6.228±0.50**	**0.113±0.01**	**61.500±1.37**
LCNN-LSTM	cqtspec	4s	20.857±0.14	0.478±0.01	72.251±2.97
LCNN-LSTM	logspec	Full	9.936±1.74	0.158±0.01	79.109±0.84
LCNN-LSTM	logspec	4s	13.018±3.08	0.330±0.05	79.706±15.80
LCNN-LSTM	melspec	Full	9.260±1.33	0.240±0.04	62.304±0.17
LCNN-LSTM	melspec	4s	27.948±4.64	0.483±0.03	82.857±3.49
LSTM	cqtspec	Full	**7.162±0.27**	**0.127±0.00**	**53.711±11.68**
LSTM	cqtspec	4s	14.409±2.19	0.382±0.05	55.880±0.88
LSTM	logspec	Full	10.314±0.81	0.160±0.00	73.111±2.52
LSTM	logspec	4s	23.232±0.32	0.512±0.00	78.071±0.49
LSTM	melspec	Full	16.216±2.92	0.358±0.00	65.957±7.70
LSTM	melspec	4s	37.463±0.46	0.553±0.01	64.297±2.23
MesoInception	cqtspec	Full	11.353±1.00	0.326±0.03	50.007±14.69
MesoInception	cqtspec	4s	21.973±4.96	0.453±0.09	68.192±12.47
MesoInception	logspec	Full	**10.019±0.18**	**0.238±0.02**	**37.414±9.16**
MesoInception	logspec	4s	16.377±3.72	0.375±0.09	72.753±6.62
MesoInception	melspec	Full	14.058±5.67	0.331±0.11	61.996±12.65
MesoInception	melspec	4s	21.484±3.51	0.408±0.03	51.980±15.32
MesoNet	cqtspec	Full	**7.422±1.61**	0.219±0.07	54.544±11.50
MesoNet	cqtspec	4s	20.395±2.03	0.426±0.06	65.928±2.57
MesoNet	logspec	Full	8.369±1.06	**0.170±0.05**	**46.939±5.81**
MesoNet	logspec	4s	11.124±0.79	0.263±0.03	80.707±12.03
MesoNet	melspec	Full	11.305±1.80	0.321±0.06	58.405±11.28
MesoNet	melspec	4s	21.761±0.26	0.467±0.00	64.415±15.68
ResNet18	cqtspec	Full	**6.552±0.49**	**0.140±0.01**	**49.759±0.17**
ResNet18	cqtspec	4s	18.378±1.76	0.432±0.07	61.827±7.46
ResNet18	logspec	Full	7.386±0.42	**0.139±0.02**	80.212±0.23
ResNet18	logspec	4s	15.521±1.83	0.387±0.02	88.729±2.88
ResNet18	melspec	Full	21.658±2.56	0.551±0.04	77.614±1.47
ResNet18	melspec	4s	28.178±0.33	0.489±0.01	83.006±7.17
Transformer	cqtspec	Full	**7.498±0.34**	**0.129±0.01**	**43.775±2.85**
Transformer	cqtspec	4s	11.256±0.07	0.329±0.00	48.208±1.49
Transformer	logspec	Full	9.949±1.77	0.210±0.06	64.789±0.88
Transformer	logspec	4s	13.935±1.70	0.320±0.03	44.406±2.17
Transformer	melspec	Full	20.813±6.44	0.394±0.10	73.307±2.81
Transformer	melspec	4s	26.495±1.76	0.495±0.00	68.407±5.53
CRNNSpoof	raw	Full	**15.658±0.35**	**0.312±0.01**	44.500±8.13
CRNNSpoof	raw	4s	19.640±1.62	0.360±0.04	**41.710±4.86**
RawNet2	raw	Full	**3.154±0.87**	**0.078±0.02**	37.819±2.23
RawNet2	raw	4s	4.351±0.29	0.132±0.01	**33.943±2.59**
RawPC	raw	Full	**3.092±0.36**	**0.071±0.00**	**45.715±12.20**
RawPC	raw	4s	3.067±0.91	0.097±0.03	52.884±6.08
RawGAT-ST	raw	Full	**1.229±0.43**	**0.036±0.01**	**37.154±1.95**
RawGAT-ST	raw	4s	2.297±0.98	0.074±0.03	38.767±1.28

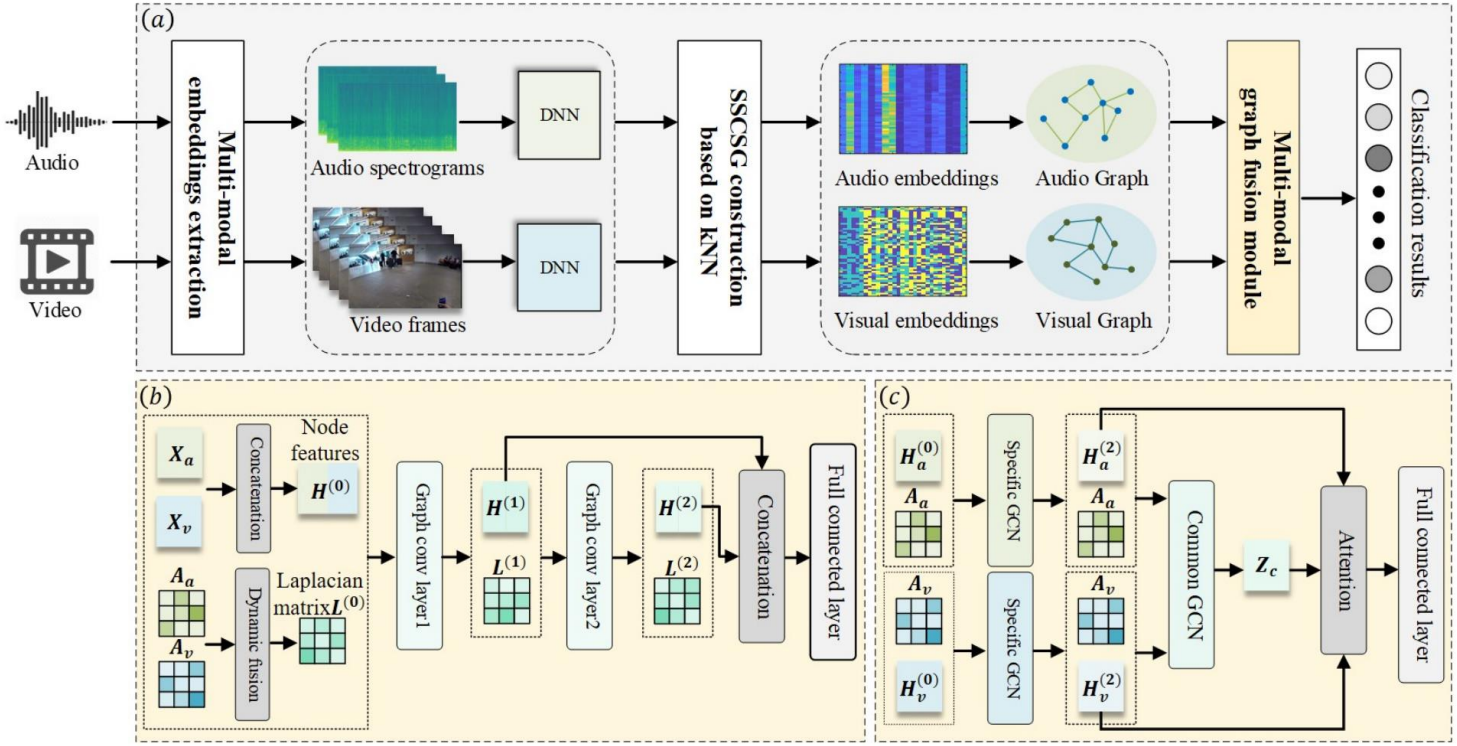
3. Audio-Visual Scene Classification Based on Multi-modal Graph Fusion

* East China University of Science and Technology

- **adjacency matrix**

$$A_a(i,j) = \exp\left(-\frac{d(\mathbf{x}_{a,i}, \mathbf{x}_{a,j})}{\mu_a}\right)$$

- Only the k-top edges with high similarity of each sample are retained



- **DM-GCN**

$$\mathbf{L}^{(t)} = \theta_a^{(t)} \tilde{\mathbf{D}}_a^{-\frac{1}{2}} \tilde{\mathbf{A}}_a \tilde{\mathbf{D}}_a^{-\frac{1}{2}} + \theta_v^{(t)} \tilde{\mathbf{D}}_v^{-\frac{1}{2}} \tilde{\mathbf{A}}_v \tilde{\mathbf{D}}_v^{-\frac{1}{2}}$$

$$\mathbf{H}^{(t+1)} = \sigma(\mathbf{L}^{(t)} \mathbf{H}^{(t)} \mathbf{W}^{(t)})$$

- **AT-GCN**

$$\mathbf{H}_a^{(t+1)} = \sigma(\tilde{\mathbf{D}}_a^{-\frac{1}{2}} \tilde{\mathbf{A}}_a \tilde{\mathbf{D}}_a^{-\frac{1}{2}} \mathbf{H}_a^{(t)} \mathbf{W}_a^{(t)})$$

$$\mathbf{Z}_a^{(t+1)} = \sigma(\tilde{\mathbf{D}}_a^{-\frac{1}{2}} \tilde{\mathbf{A}}_a \tilde{\mathbf{D}}_a^{-\frac{1}{2}} \mathbf{Z}_a^{(t)} \mathbf{W}_c^{(t)})$$

$$\mathbf{Z}_c = \frac{\mathbf{Z}_a^{(T_c)} + \mathbf{Z}_v^{(T_c)}}{2}$$

3. Audio-Visual Scene Classification Based on Multi-modal Graph Fusion

* East China University of Science and Technology

Table 1: *The effectiveness of GCN in embedding optimization.*

Model	Audio-only		Visual-only	
	Log-loss	Acc	Log-loss	Acc
without GCN	0.962	68.4%	0.720	86.2%
with GCN	0.929	72.1%	0.592	88.1%

Table 2: *Performance comparison on the development dataset.*

Model	Audio-Visual	
	Log-loss	Acc
[13]	0.658	77.0%
[15]	0.688	88.3%
Mean-GCN	0.358	90.4%
AT-GCN(ours)	0.304	90.7%
DM-GCN(ours)	0.329	91.3%

Table 3: *Complexity comparison of fusion modules.*

Model	Model Complexity	
	Parameters	Time
[15]	272k	89 hours
Mean-GCN	86k	5 hours
AT-GCN(ours)	182k	5 hours
DM-GCN(ours)	86k	5 hours

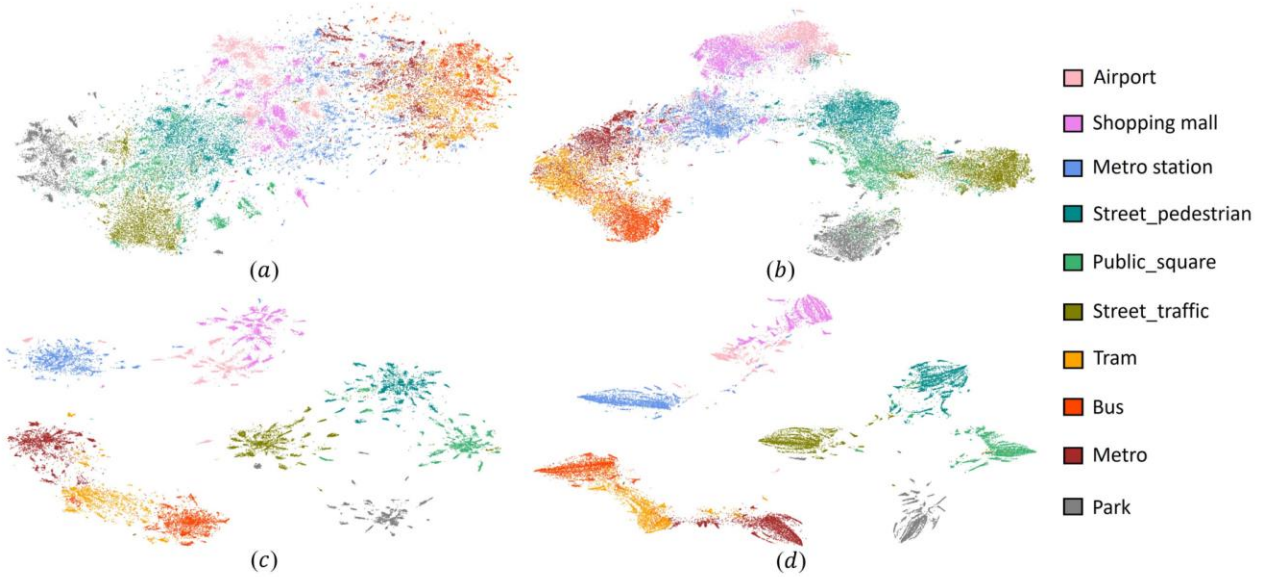


Figure 2: *Visualization of embeddings by t-SNE. (a) and (b) are output embeddings before and after GCN optimization in audio modality, (c) and (d) are output embeddings before and after GCN optimization in video modality.*

4. Interactive Co-Learning with Cross-Modal Transformer for Audio-Visual Emotion Recognition

* NTT Corporation

- **Audio encoder**

$$
\begin{aligned}
\boldsymbol{A}_{\mathrm{co}} &= \mathrm{ConvolutionPooling}(\boldsymbol{S}; \boldsymbol{\theta}^{\mathrm{co}}_{\mathrm{audio}}), \\
\boldsymbol{A}_{\mathrm{po}} &= \mathrm{AddPosition}(\boldsymbol{A}_{\mathrm{co}}), \\
\boldsymbol{A}_{\mathrm{tr}} &= \mathrm{TransformerEnc}(\boldsymbol{A}_{\mathrm{po}}; \boldsymbol{\theta}^{\mathrm{tr}}_{\mathrm{audio}}), \\
\boldsymbol{A} &= \mathrm{AddAudioSegment}(\boldsymbol{A}_{\mathrm{tr}}; \boldsymbol{\theta}^{\mathrm{se}}_{\mathrm{audio}}),
\end{aligned}
$$

networks, and AddAudioSegment() is a function that adds a continuous vector in which speech segment information is embedded.

- **Visual encoder**

$$
\begin{aligned}
\boldsymbol{V}_{\mathrm{cnn}} &= \mathrm{CNN}(\boldsymbol{C}; \boldsymbol{\theta}^{\mathrm{cnn}}_{\mathrm{visual}}), \\
\boldsymbol{V}_{\mathrm{po}} &= \mathrm{AddPosition}(\boldsymbol{V}_{\mathrm{cnn}}), \\
\boldsymbol{V}_{\mathrm{tr}} &= \mathrm{TransformerEnc}(\boldsymbol{V}_{\mathrm{po}}; \boldsymbol{\theta}^{\mathrm{tr}}_{\mathrm{visual}}), \\
\boldsymbol{V} &= \mathrm{AddVisualSegment}(\boldsymbol{V}_{\mathrm{tr}}; \boldsymbol{\theta}^{\mathrm{se}}_{\mathrm{visual}}),
\end{aligned}
$$

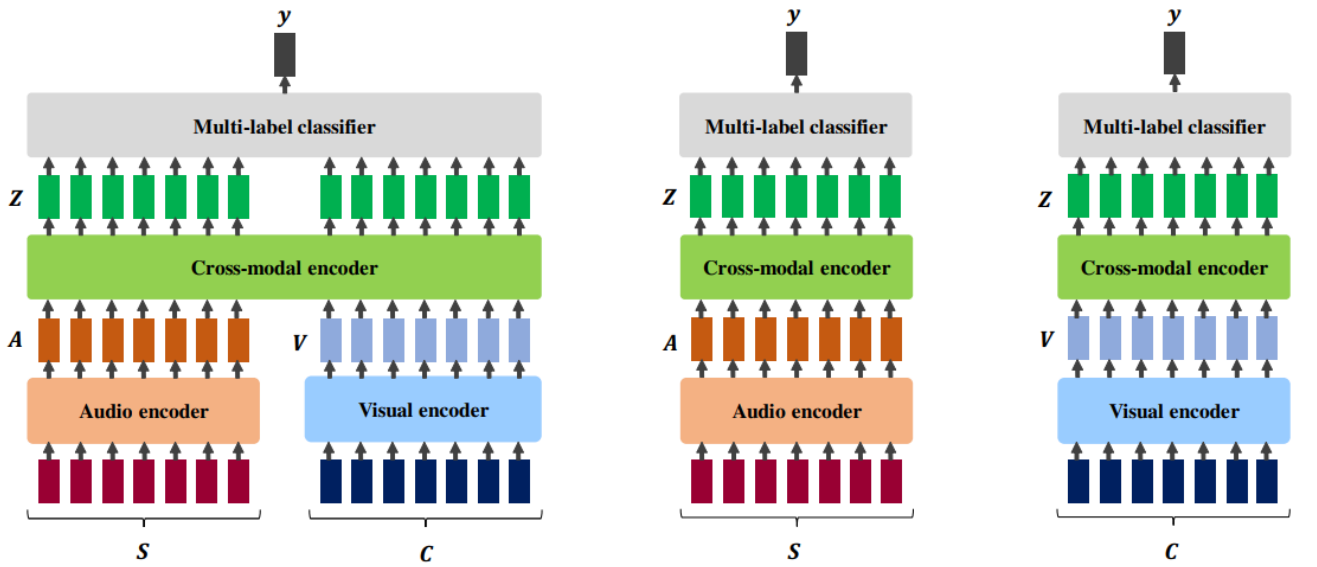
- **Cross-modal encoder**

$$
\boldsymbol{Z}_0 = \begin{cases} \mathrm{Concat}(\boldsymbol{A}, \boldsymbol{V}) & \text{if audio-visual features are used,} \\ \boldsymbol{A} & \text{if audio features are only used,} \\ \boldsymbol{V} & \text{if visual features are only used,} \end{cases}
$$

$$
\boldsymbol{Z} = \mathrm{TransformerEnc}(\boldsymbol{Z}_0; \boldsymbol{\theta}_{\mathrm{cross}}),
$$

- **Multi-label classifier**

$$
\begin{aligned}
\boldsymbol{o} &= \mathrm{AttentivePooling}(\boldsymbol{Z}; \boldsymbol{\theta}^{\mathrm{att}}_{\mathrm{label}}), \\
\bar{\boldsymbol{y}} &= \mathrm{Swish}(\boldsymbol{o}; \boldsymbol{\theta}^{\mathrm{sw}}_{\mathrm{label}}), \\
\boldsymbol{y} &= \mathrm{Sigmoid}(\bar{\boldsymbol{y}}; \boldsymbol{\theta}^{\mathrm{sig}}_{\mathrm{label}}),
\end{aligned}
$$



(a). Audio-visual emotion recognition *(b). Audio emotion recognition* *(c). Visual emotion recognition*

Table 2: *Recognition performances for target emotions. "A", "V", and "A+V" represent the usage of the audio features alone, the visual features alone, and both of the audio and visual features. wF1 and mF1 are the weighted and macro F1s, respectively.*

	Training	Inference	*Happy* wF1	*Happy* mF1	*Sad* wF1	*Sad* mF1	*Anger* wF1	*Anger* mF1	*Surprise* wF1	*Surprise* mF1	*Disgust* wF1	*Disgust* mF1	*Fear* wF1	*Fear* mF1	Average wF1	Average mF1
Baseline	A	A	64.2	64.1	70.9	59.1	74.0	59.4	86.1	48.3	79.8	61.0	88.0	50.5	77.2	57.1
	V	V	60.5	60.4	70.2	57.0	74.9	61.0	85.9	48.0	78.2	57.0	87.8	49.8	76.3	55.5
Conventional	A+V	A	58.2	57.9	70.0	54.5	69.4	53.1	86.4	51.1	77.5	57.5	87.8	48.1	74.9	53.7
		V	61.9	61.7	69.9	57.0	73.7	61.6	86.1	49.8	76.4	58.4	87.4	50.7	75.9	56.5
		A+V	63.6	63.3	72.1	58.9	74.6	62.0	**86.2**	**50.3**	79.4	61.8	**88.0**	50.7	77.3	57.8
Proposed	Interactive co-learning	A	63.9	63.7	71.9	59.8	72.4	60.3	86.3	54.9	80.1	64.4	87.9	48.8	77.1	58.7
		V	60.0	60.2	68.5	56.0	74.2	61.6	86.1	48.0	77.8	58.0	87.5	53.9	75.7	56.3
		A+V	**66.0**	**66.0**	**72.2**	**62.2**	**75.5**	**63.8**	86.0	47.5	**81.0**	**66.5**	87.4	**55.0**	**78.0**	**60.2**

- **Interactive co-learning**

$$
\begin{aligned}
\mathcal{L}_{\mathrm{AV}}(\Theta) &= -\sum_{t=1}^{T}\sum_{k=1}^{K} \log P(l^t_k | \boldsymbol{S}^t, \boldsymbol{C}^t, \Theta), \\
\mathcal{L}_{\mathrm{A}}(\Theta) &= -\sum_{t=1}^{T}\sum_{k=1}^{K} \log P(l^t_k | \boldsymbol{S}^t, \Theta), \\
\mathcal{L}_{\mathrm{V}}(\Theta) &= -\sum_{t=1}^{T}\sum_{k=1}^{K} \log P(l^t_k | \boldsymbol{C}^t, \Theta),
\end{aligned}
$$

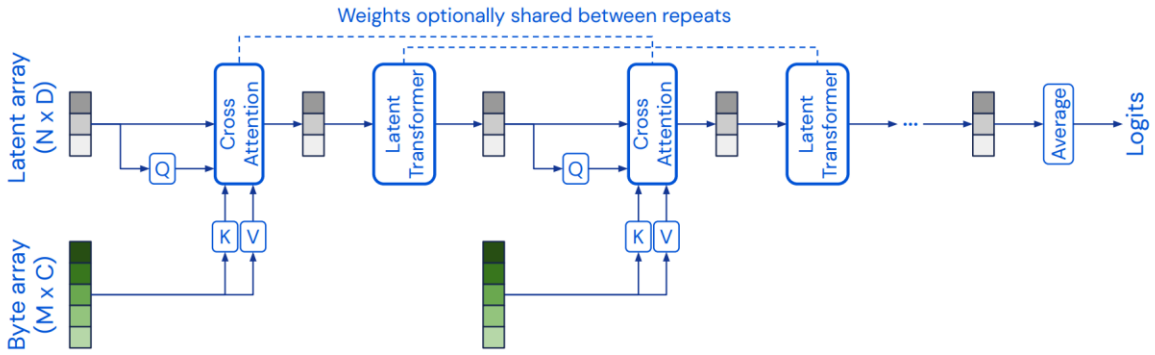
$$
\hat{\Theta} = \underset{\Theta}{\mathrm{argmin}} \{\mathcal{L}_{\mathrm{AV}}(\Theta) + \mathcal{L}_{\mathrm{A}}(\Theta) + \mathcal{L}_{\mathrm{V}}(\Theta)\}.
$$

5. Perceiver: General Perception with Iterative Attention

* DeepMind

- Motivation
  - The perception models are designed for individual modalities, often relying on domain-specific assumptions
    - **How to build a universal modal?**
  - Scale quadratically with the number of inputs, in terms of both memory and computation.
    - **How to handle very large inputs?**

- Methods



$Q\ (N \times C)$

$K\ (M \times C)$

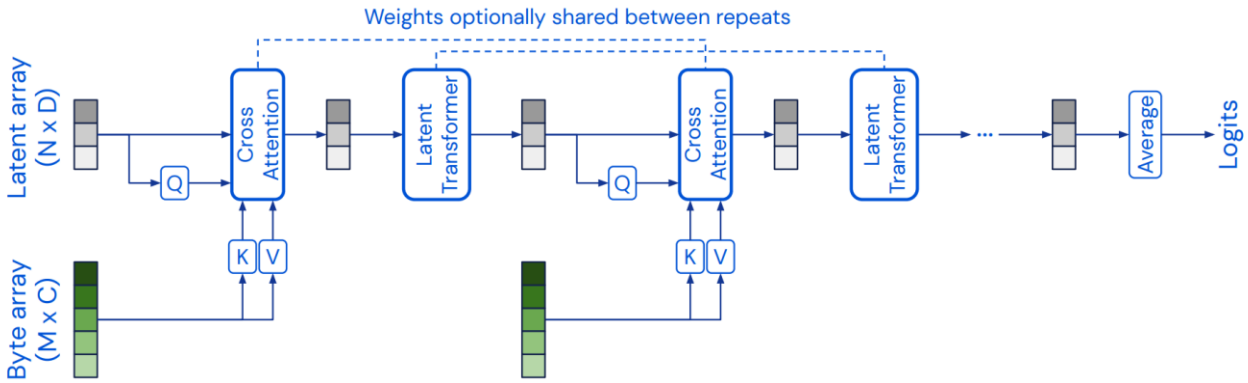
$V\ (M \times C)$

$\rightarrow QK^T (N \times M) \rightarrow (QK^T)V\ (N \times C)$

5. Perceiver: General Perception with Iterative Attention

* DeepMind

- Methods



Origin: $O(M^2)$

New: $O(MN + LN^2)$

- Experiments

ResNet-50 (He et al., 2016)	77.6
ViT-B-16 (Dosovitskiy et al., 2021)	77.9
ResNet-50 (FF)	73.5
ViT-B-16 (FF)	76.7
Transformer (64x64, FF)	57.0
Perceiver (FF)	78.0

Table 1. Top-1 validation accuracy (in %) on ImageNet. Models

Model / Inputs	Audio	Video	A+V
Benchmark (Gemmeke et al., 2017)	31.4	-	-
Attention (Kong et al., 2018)	32.7	-	-
Multi-level Attention (Yu et al., 2018)	36.0	-	-
ResNet-50 (Ford et al., 2019)	38.0	-	-
CNN-14 (Kong et al., 2020)	43.1	-	-
CNN-14 (no balancing & no mixup) (Kong et al., 2020)	37.5	-	-
G-blend (Wang et al., 2020c)	32.4	18.8	41.8
Attention AV-fusion (Fayek & Kumar, 2020)	38.4	25.7	46.2
Perceiver (raw audio)	38.3	25.8	43.4
Perceiver (mel spectrogram)	38.4	25.8	43.2

Table 3. Perceiver performance on AudioSet, compared to state-of-the-art models (mAP, higher is better).

	Accuracy
PointNet++ (Qi et al., 2017)	**91.9**
ResNet-50 (FF)	66.3
ViT-B-2 (FF)	78.9
ViT-B-4 (FF)	73.4
ViT-B-8 (FF)	65.3
ViT-B-16 (FF)	59.6
Transformer (44x44)	82.1
Perceiver	**85.7**

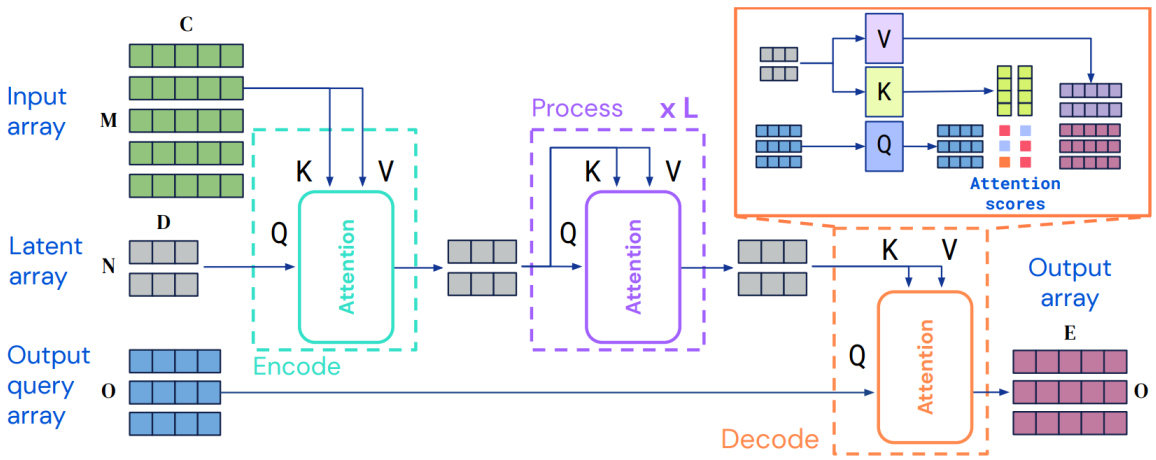
Table 4. Top-1 test-set classification accuracy (in %) on ModelNet40. Higher is better. We report best result per model class,

6. Perceiver IO: A general architecture for structured inputs & outputs

* DeepMind

- Motivation
  - Perceiver can only handle simple output spaces like classification
    - **how to handle a host of new domains without sacrificing the benefits of deep, domain-agnostic processing?**

- Methods



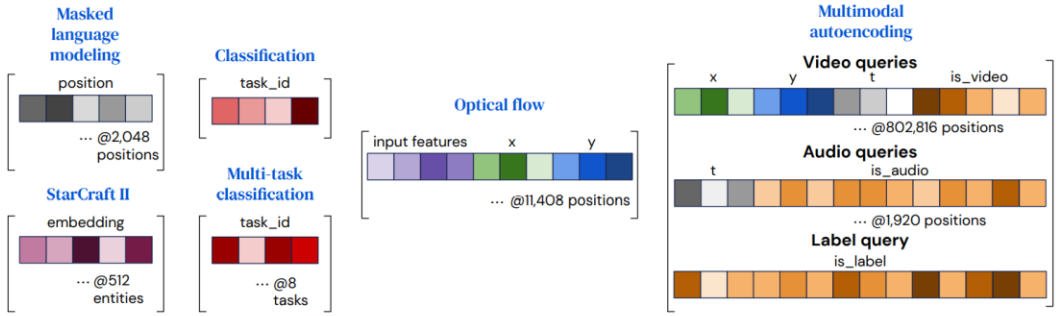


Figure 3: We construct queries with output-specific features to produce outputs with different semantics. For settings where each output point differs only in its position, like language, a position embedding can be used. Input features for the target output can also be used to query, either alone (as for StarCraft II) or alongside position features (as for flow). For multi-{task, modal} settings we use one embedding for each {task, modality} instead of each position. A single learned embedding suffices for simple classification tasks, like ImageNet. For tasks with heterogeneous outputs like multimodal autoencoding, features that are specific to some queries (like xy position) can be combined with modality embeddings, which also pad embeddings to fixed length.

6. Perceiver IO: A general architecture for structured inputs & outputs

* DeepMind

- Experiments

Model	Tokenization	M	N	Depth	Params	FLOPs	SPS	Avg.
BERT Base (test)	SentencePiece	512	512	12	110M	109B	-	81.0
BERT Base (ours)	SentencePiece	512	512	12	110M	109B	7.3	81.1
Perceiver IO Base	SentencePiece	512	256	26	223M	119B	7.4	**81.2**
BERT (matching FLOPs)	UTF-8 bytes	2048	2048	6	20M	130B	2.9	71.5
Perceiver IO	UTF-8 bytes	2048	256	26	201M	113B	7.6	81.0
Perceiver IO++	UTF-8 bytes	2048	256	40	425M	241B	4.2	**81.8**

Network	Sintel.clean	Sintel.final	KITTI
PWCNet (Sun et al., 2018)	2.17	2.91	5.76
RAFT (Teed & Deng, 2020)	1.95	2.57	**4.23**
Perceiver IO	**1.81**	**2.42**	4.98