ACMMM 2023

TMac: Temporal Multi-Modal Graph Learning for Acoustic Event Classification

Meng Liu[1], Ke Liang[1], Dayu Hu, Hao Yu, Yue Liu, Lingyuan Meng, Wenxuan Tu, Sihang Zhou, Xinwang Liu*

National University of Defense Technology, Changsha, China

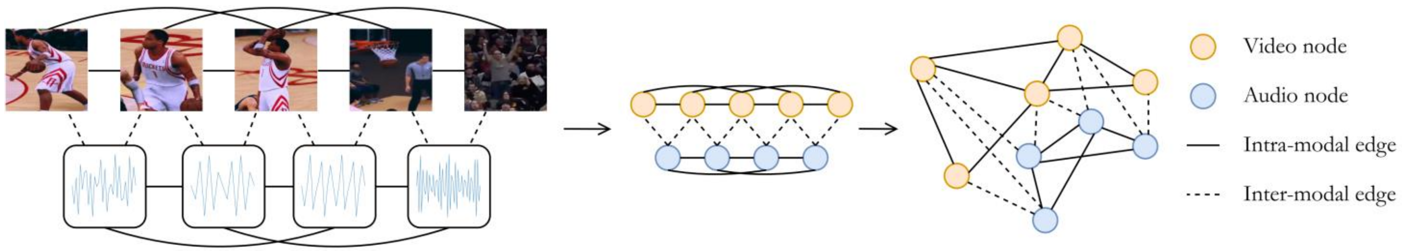
Presenter: Meng Liu

Email: mengliuedu@163.com

Audiovisual Graph Learning

- Audio is an indispensable part of information expression in the real world, and its appearance is often accompanied by visual information.
- Visual information is an effective complement to audio information.
- Audiovisual data can be well split into multiple segments and constructed as a graph.

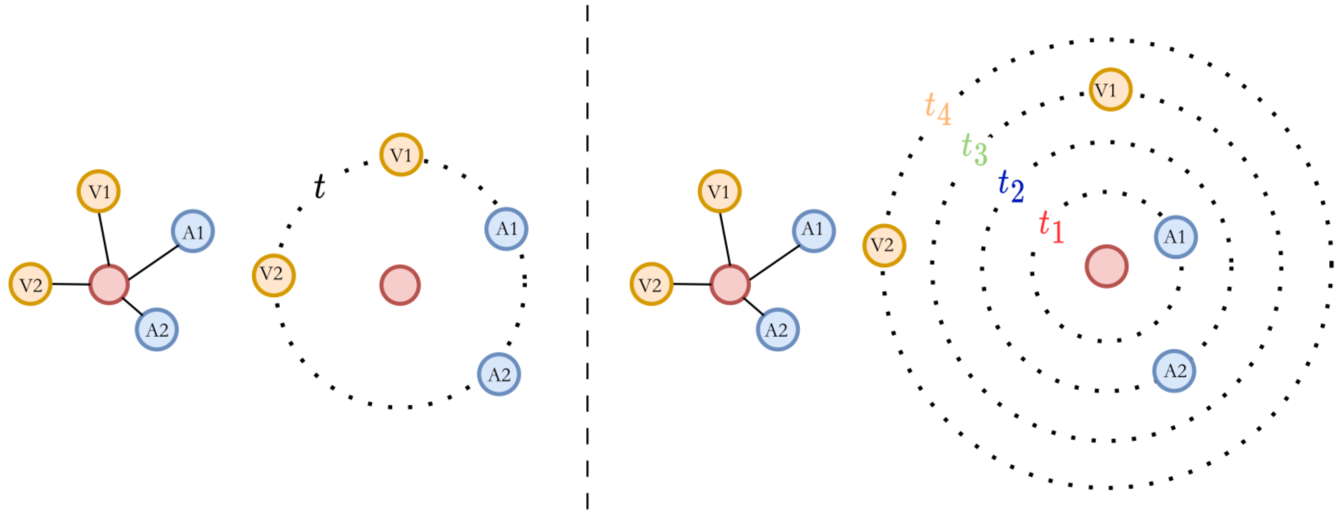
(a) Audio and video segments over time.



(b) Audiovisual graph construction.

Static and Temporal

- Temporal equipotential lines in different graph constructions.
- Nodes on the same equipotential line are treated as having equal time status.

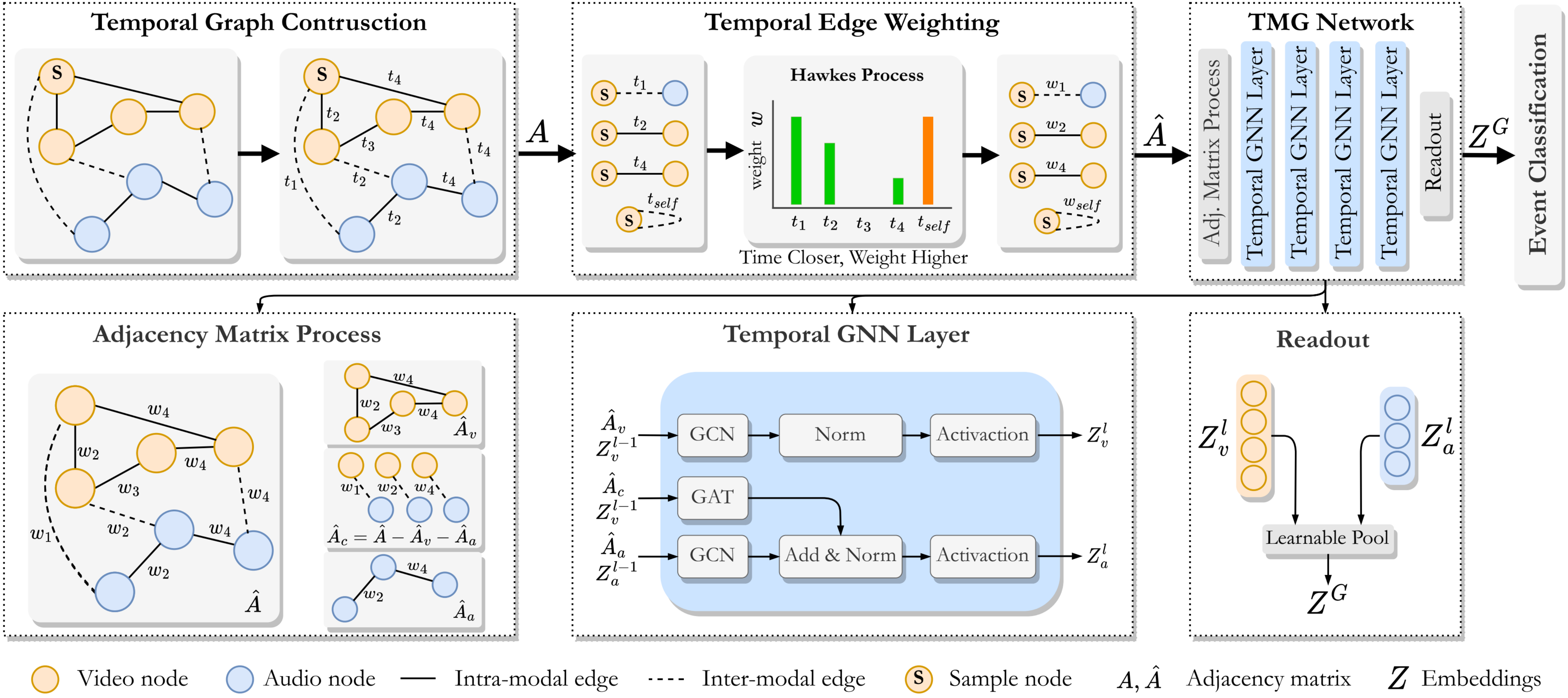


(a) Temporal equipotential lines in existing methods

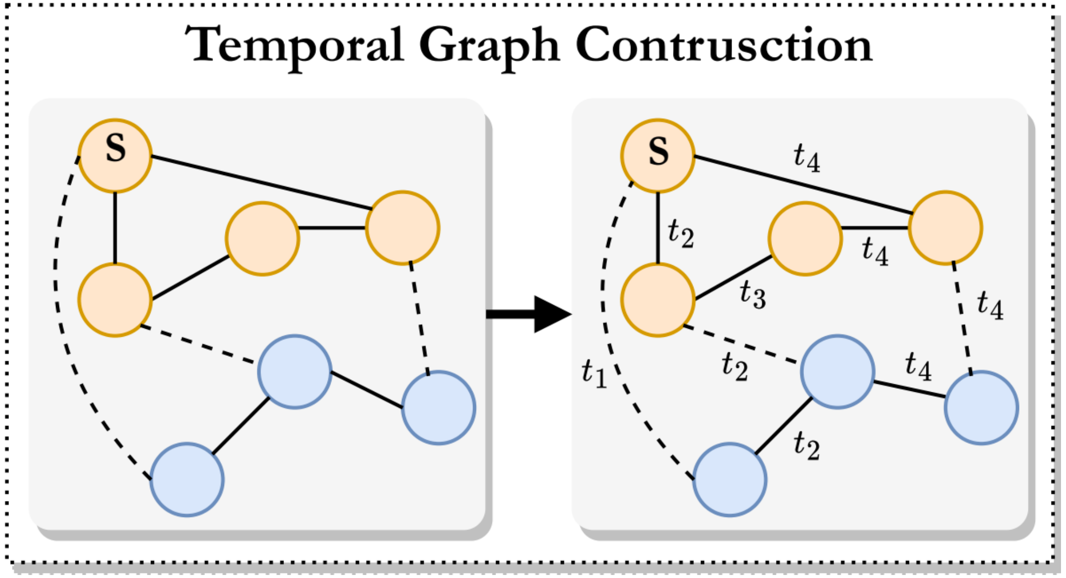
(b) Temporal equipotential lines in actual data

- **Problem.** We point out that audiovisual multi-modal data is well-suited for temporal graph modeling.
- **Algorithm.** We propose the TMac method to construct the temporal graph structure for audiovisual data, which introduce the Hawkes process to capture the dynamic information of intra-modal and inter-modal.
- **Evaluation.** We compare TMac with multiple methods with several experiments.

TMac


Temporal Graph Contrusction
Temporal Edge Weighting
Hawkes Process
weight w
t_1 t_2 t_3 t_4 t_{self}
Time Closer, Weight Higher
TMG Network
Adj. Matrix Process
Temporal GNN Layer
Readout
Event Classification
A
$\hat{A}$
Z^G
Adjacency Matrix Process
$\hat{A}_c = \hat{A} - \hat{A}_v - \hat{A}_a$
Temporal GNN Layer
GCN
Norm
Activaction
GAT
Add & Norm
Readout
Learnable Pool
Video node
Audio node
Intra-modal edge
Inter-modal edge
Sample node
$A, \hat{A}$ Adjacency matrix
Z Embeddings

Graph Construction



Feature Extraction

- Each audio clip is partitioned into segments of 960 ms duration, with an overlap of 764 ms.
- Each video is split into non-overlapping chunks of 250 ms duration.

Acoustic Events as Temporal Graphs

- Given an audiovisual event, we uniformly split it into video and audio segments to construct a temporal graph. The graph consists of two types of nodes (audio and video) and two types of edges (intra-modal and inter-modal).
- Due to these different types, we can construct three adjacency matrices: video matrix A_v, audio matrix A_a, and cross-modal matrix A_c.

Temporal Multi-Modal Graph Network

Temporal Edge Weighting

- For a node u's neighborhood, if neighboring nodes are closer in time to the source node, they tend to be more similar in terms of semantics and features. In this way, when a node receives messages from neighborhood in GNNs, these messages should be weighted by time information.

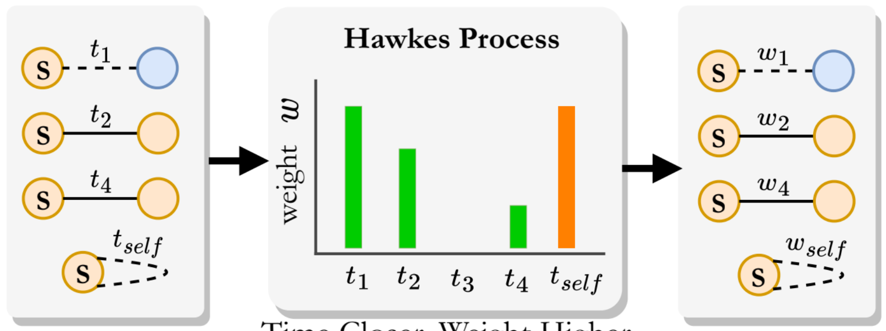
$$\hat{A} = [A_{i,j}W_{i,j}]_{n\times n}, \qquad W_{i,j} = \exp(-\frac{t_{max} - t_i + 1}{t_{max} - t_{min} + 1})$$

GNN Paradigm

- GNN utilizes the adjacency matrix for node features aggregation, the l-layer of GNNs can be formulated as follows.

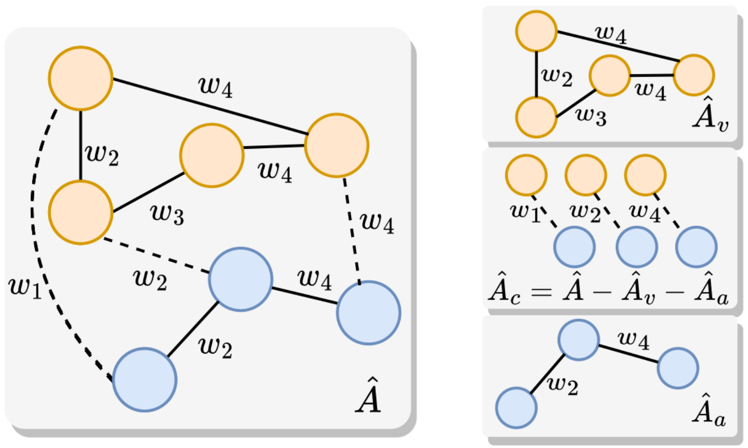
$$Z^l = \mathrm{GNN}(A, Z^{l-1}) = \sigma(AZ^{l-1}W^{l-1})$$

Temporal Edge Weighting



Time Closer, Weight Higher

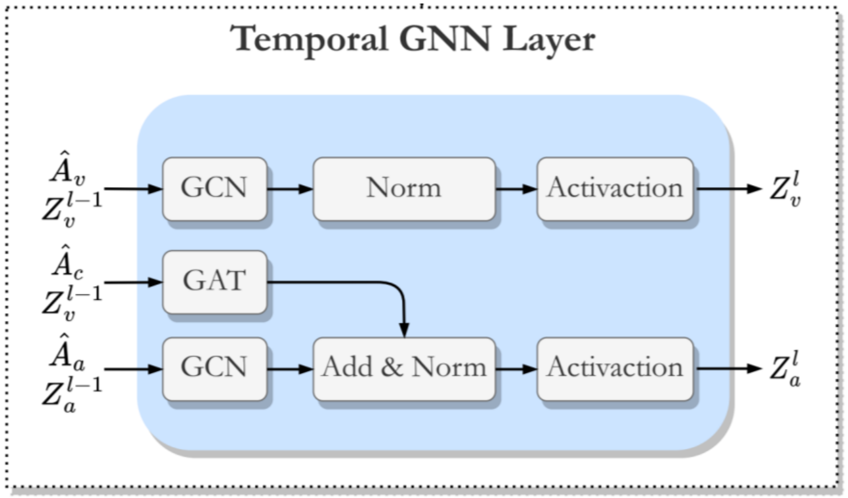
Adjacency Matrix Process

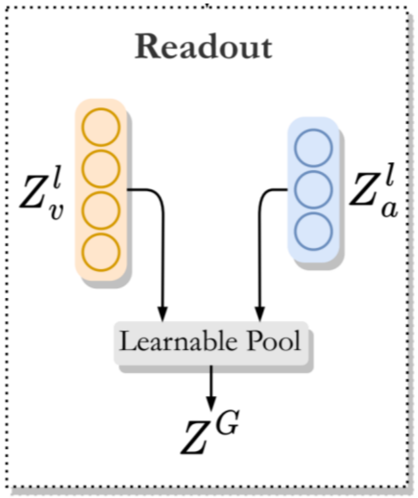


Temporal Multi-Modal Graph Network

Embedding Generation

- A temporal multi-modal graph network can be regarded as using several GNNs to extract and aggregate different modal information separately. Note that our final objective is acoustic event classification, thus we need propagate video embeddings to audio embeddings.





$$Z_a^l = \mathrm{GCN}_a\left(\hat{A}_a, Z_a^{l-1}\right) + \mathrm{GAT}_a\left(\hat{A}_c, Z_v^{l-1}\right), \qquad Z_v^l = \mathrm{GCN}_a\left(\hat{A}_v, Z_v^{l-1}\right)$$

- After modeling the node information, we need to learn a graph readout function for this event to pool all node embeddings into one final embedding. For the i-th event G_i, its graph embedding can be calculated as follows.

$$Z^{G_i} = \mathrm{Readout}(G_i) = \left[\mathrm{P}\left(Z_a^l\right) \mid \mathrm{P}\left(Z_v^l\right)\right] = Z_a^l \mathrm{P}^a + Z_v^l \mathrm{P}^v$$

Acoustic Event Classification

Model (Year)	mAP	AUC	Params
Spectrogram-VGG	0.26±0.01	-	6M
ResNet-1D-audio	0.35±0.01	0.90±0.00	40.4M
ResNet-1D-both	0.38±0.03	0.89±0.02	81.2M
DaiNet	0.25±0.07	-	1.8M
R(2+1)D-video	0.36±0.00	0.81±0.00	33.4M
Wav2vec2-audio	0.42±0.02	0.88±0.00	94.4M
Wave-Logmel	0.43±0.04	-	81M
VATT	0.39±0.02	-	87M
AST	0.44±0.00	-	88M
PaSST-S	0.49±0.01	0.90±0.01	87M
VAED	<u>0.50±0.01</u>	<u>0.91±0.01</u>	2.1M
Audio-MAE	0.47±0.01	-	86M
MaskSpec	0.47±0.02	-	86M
SSL graph	0.42±0.02	-	218K
HGCN	0.44±0.01	0.88±0.01	42.4M
TMac (improv.)	**0.56±0.01** (+12.00%)	**0.94±0.01** (+3.29%)	4.3M -

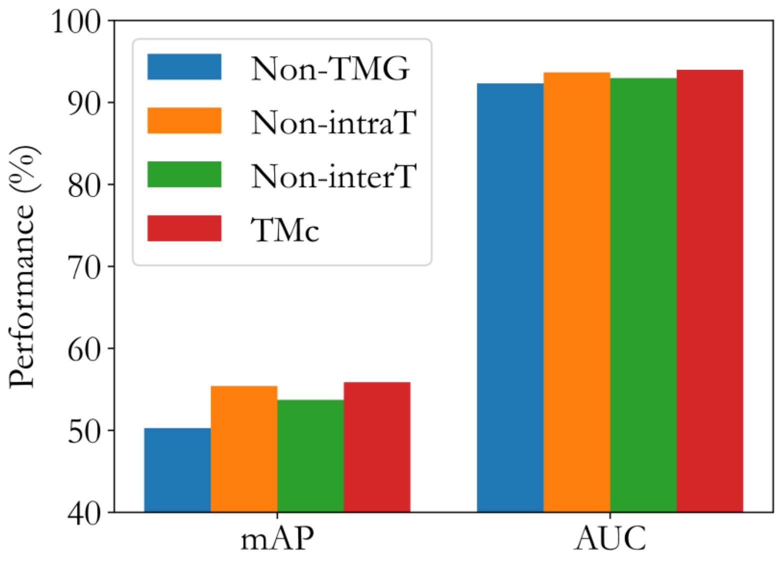
- **Q1:** Is temporal information really useful for the acoustic event classification task?
- **Answer:** Temporal information is useful, and combining it does not result in massive model inflation.

(1) TMac outperforms other all methods on the mAP metric.

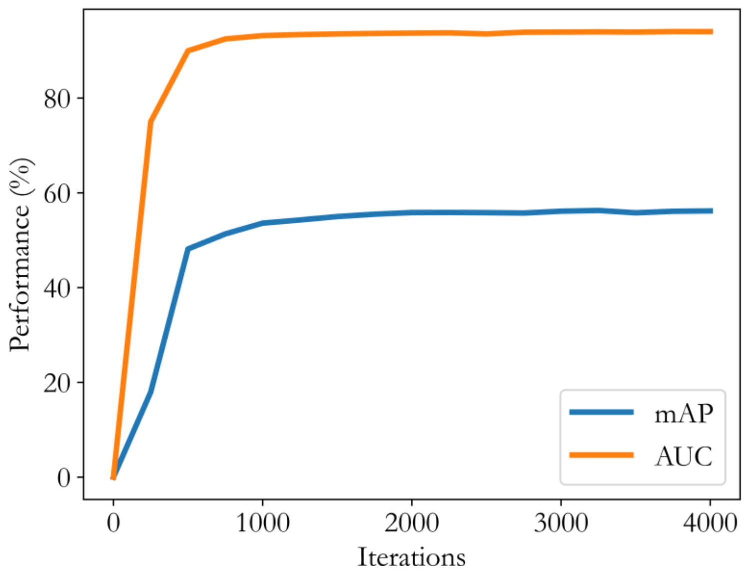
(2) The highest ROC-AUC score of TMac indicating that it produces more reliable predictions at various thresholds.

(3) The small parameter number of TMac means that it is easy to implement.

Ablation Study



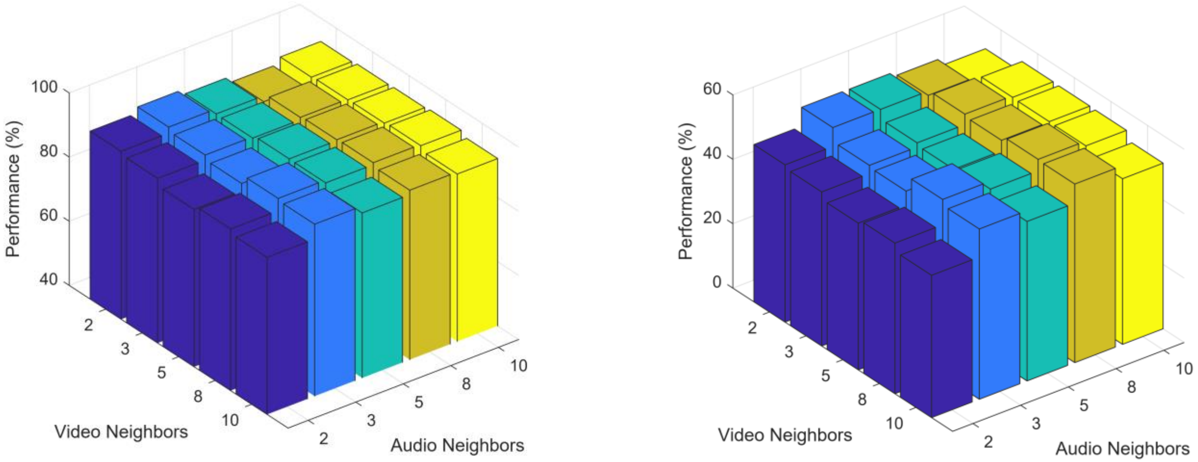
Ablation study on temporal information



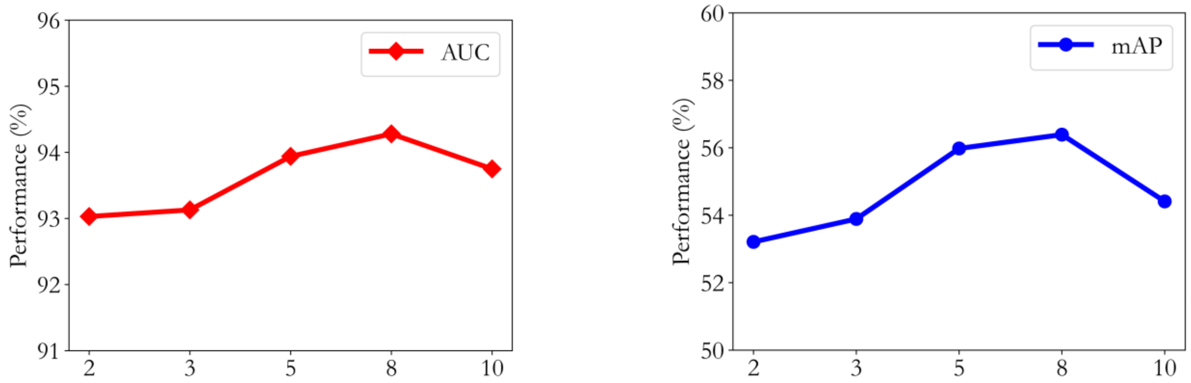
Performance changes over the number of iterations

- **Q2：** Is temporal information both intra-modal and inter-modal meaningful?
- **Answer:** Both intra-modal and inter-modal temporal information are meaningful, with inter-modal information being more meaningful.
- **Q3：** Will the extra consideration of time information affects the convergence of TMac?
- **Answer:** TMac can achieves convergence in a relatively short period.

Parameter Sensitivity Study



Ablation study on intra-modal neighbor numbers



Ablation study on inter-modal neighbor numbers

- **Q4:** Does the model's effect remain stable when faced with different parameter choices?
- **Answer:** matter how the super-parameters are combined, the model can maintain a good performance.
- Our code is available at: https://github.com/MGitHubL/TMac

Thanks!

Presenter: Meng Liu

Contact: mengliuedu@163.com