SPECTROGRAM ANALYSIS VIA SELF-ATTENTION FOR REALIZING CROSS-MODEL VISUAL-AUDIO GENERATION

Huadong Tan, Guang Wu, Pengcheng Zhao, Yanxiang Chen

School of Computer Science and Information Engineering, Hefei University of Technology, *Hefei, 230009, China*

ABSTRACT

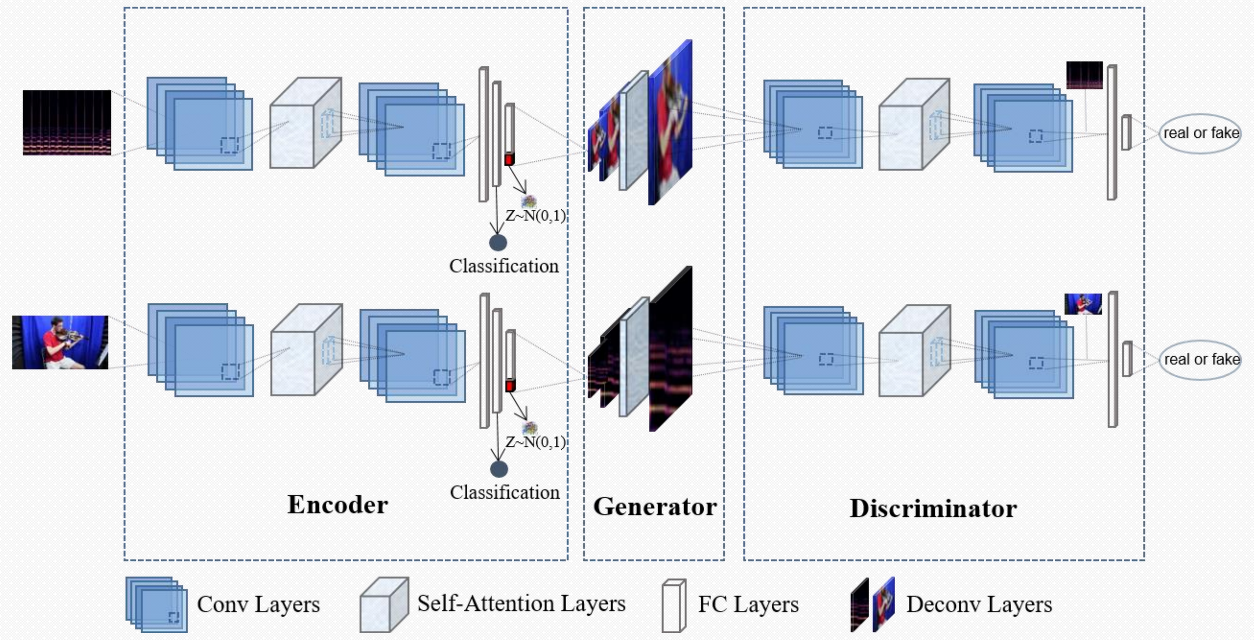
Human cognition is supported by the combination of multimodal information from different sources of perception. The two most important modalities are visual and audio. Cross-modal visual-audio generation enables the synthesis of data from one modality following the acquisition of data from another. This brings about the full experience that can only be achieved through the combination of the two. In this paper, the Self-Attention mechanism is applied to cross-modal visual-audio generation for the first time. This technique is implemented to assist in the analysis of the structural characteristics of the spectrogram. A series of experiments are conducted to discover the best performing configuration. The post-experimental comparison shows that the Self-Attention module greatly improves the generation and classification of audio data. Furthermore, the presented method achieves results that are superior to existing cross-modal visual-audio generative models.

MOTIVATION

- Cross-modal bidirectional generation has long-term value in applications of data restoration.
- It is hard for existing methods to distinguish features at pixel level by using CNN alone.
- There are many repetitive waveform structures in the Log-Mel spectrogram, which makes it suitable for the Self-Attention mechanism.

PROPOSED METHOD

- Illustration of the proposed model



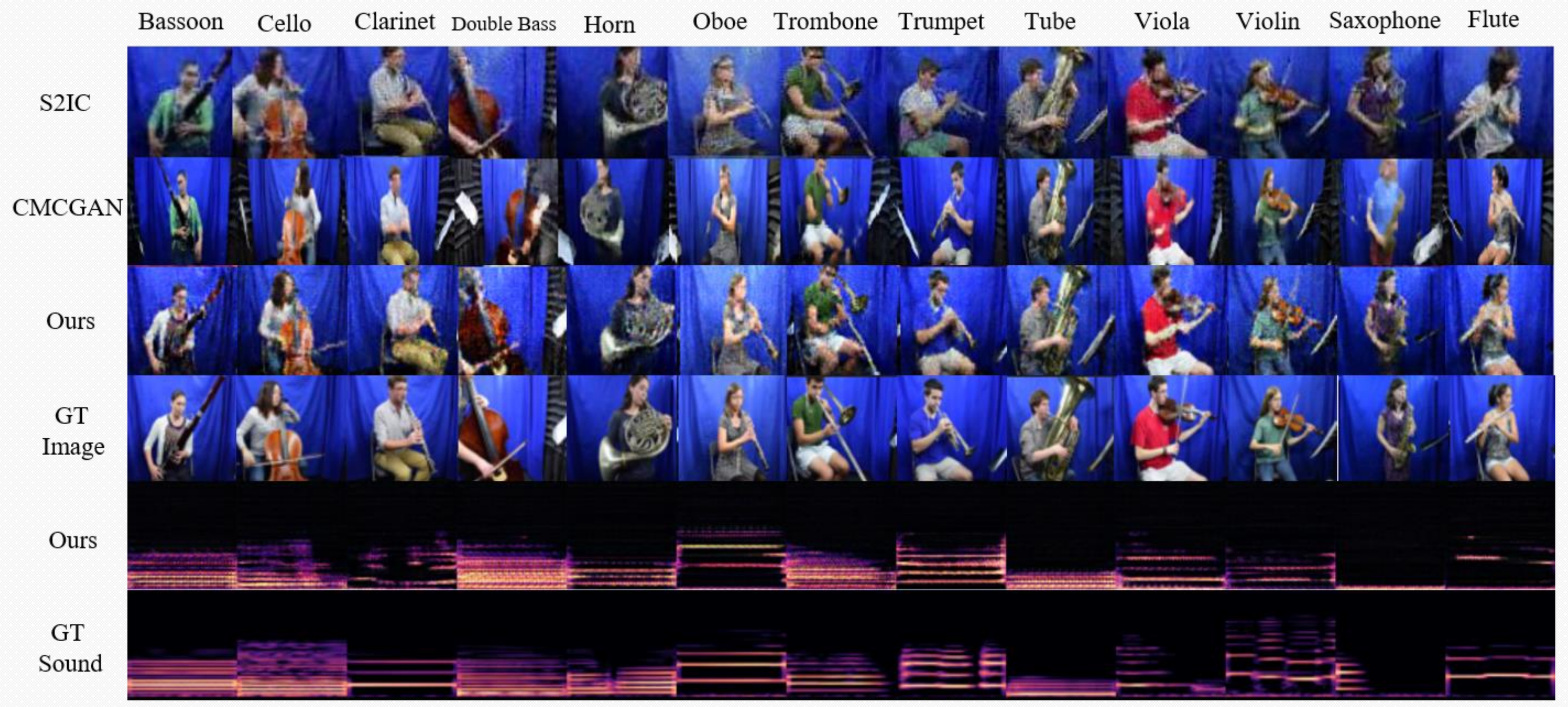
- Encoder: using the image or the spectrogram as input to extract the corresponding features, and a classifier is added to better optimize the encoder. The self-attention layer is connected after the first convolutional layer.
- Generator: taking the output of the encoder embedded with a random noise z~N(0, 1) as input to generate the new data. The self-attention layer is connected to the fourth deconvolutional layer.
- Discriminator: taking the image and the spectrogram as input, and outputs a probability value between 0 and 1 for discriminating the authenticity. The self-attention layer is connected after the first convolutional layer.

- Formulation

$$L_D = -E_{(x,y)\sim P_{data}}[min(0, -1 + D(x,y))]$$
$$-\frac{1}{2}(E_{Z\sim P_Z, y\sim P_{data}}[min(0, -1 - D(G(Z),y))]$$
$$+E_{(\hat{x},y)\sim P_{data}}[min(0, -1 - D(\hat{x},y))]$$

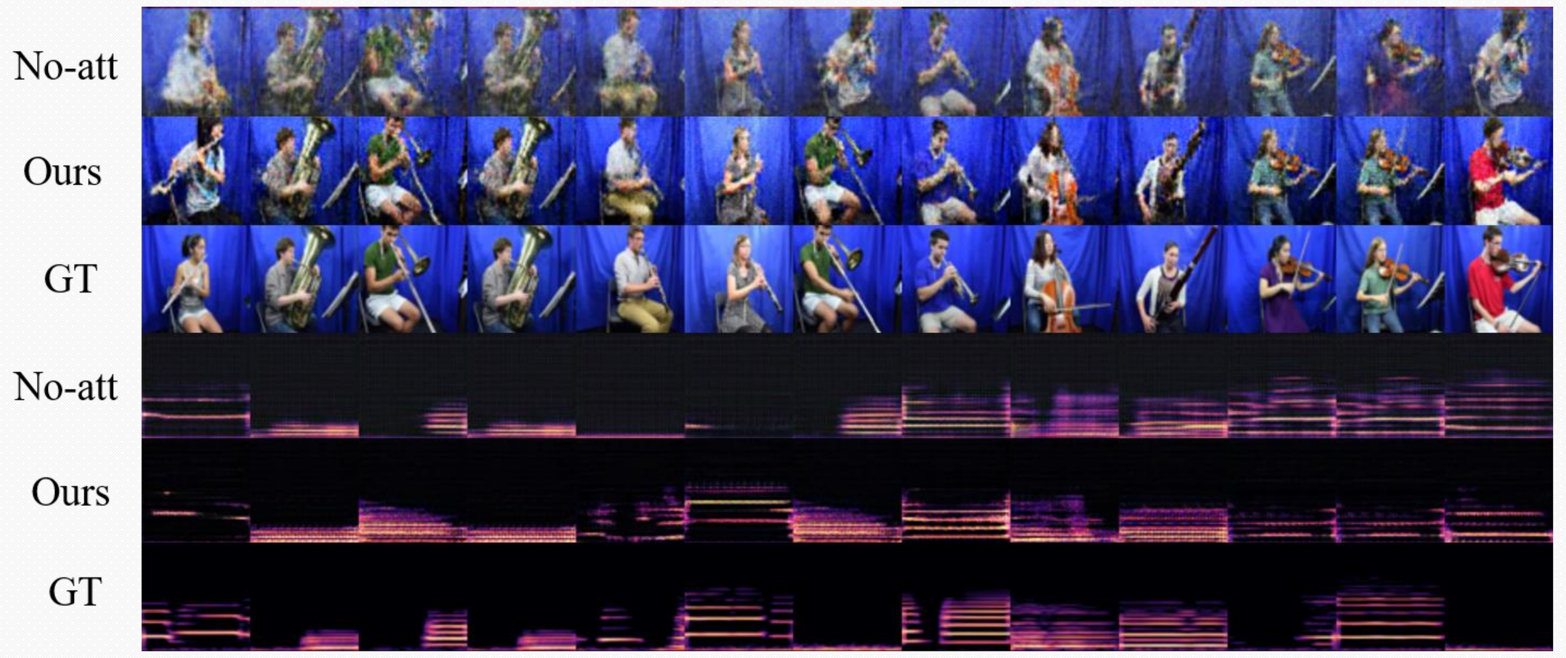
RESULTS

- The performance of SA-CMGAN and existing models



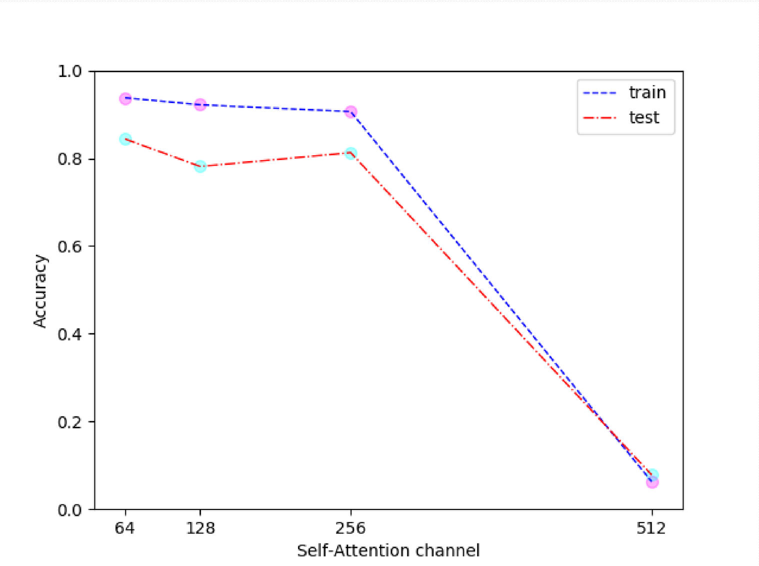
Models(S2I)	S2IC	CMCGAN	SA-CMGAN
Training Acc	0.8737	0.9105	**0.9375**
Testing Acc	0.7556	0.7661	**0.8438**
Models(I2S)	I2S	CMCGAN	SA-CMGAN
Training Acc	-	0.8109	**0.8750**
Testing Acc	0.1117	0.5189	**0.5937**

- The performance of SA-CMGAN with and without the self-attention module



Models(S2I)	No-att	SA-CMGAN
Training Acc	0.90625	**0.93750**
Testing Acc	0.79688	**0.84375**
Models(I2S)	No-att	SA-CMGAN
Training Acc	0.28125	**0.87500**
Testing Acc	0.23427	**0.59375**

- The results when embedding the self-attention module at different channels



In search of the best performing approach, the design was modified several times with the self-attention module embedded at different layers. The results show that the classification accuracy is highest when self-attention channel is set at 64. As the number increases beyond the peak at 64, the accuracy decreases proportionally. The effect is particularly poor when the channel reaches 512, which means that embedding the self-attention module in the high-dimensional layer is not effective.

CONLUSIONS

In this work, a first attempt to apply the self-attention mechanism to cross-model visual-audio generation is made. The networks are optimized using spectral normalization and several experiments were conducted in search of the best configuration. The results demonstrate that the proposed method performs superiorly in trems of both accuracy and training time.

REFERENCES

[1] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu, "Deep cross-modal audio-visual generation," ACM 2017.

[2] Wangli Hao, Zhaoxiang Zhang, and He Guan, "Cmcgan: A uniform framework for cross-modal visual-audio mutual generation," AAAI 2018.

[3] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena, "Self-attention generative adversarial networks," arXiv preprint arXiv: 1805.08318, 2018.

[4] Jae Hyun Lim and Jong Chul Ye, "Geometric gan," arXiv preprint arXiv: 1705.02894, 2017.

[5] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," arXiv preprint arXiv: 1411.178, 2014.

SOURCE CODES

The source codes can be downloaded at:**https:// github.com/TwistedW/SA-CMGAN.**

ACKNOWLEDGMENTS

This work is supported by the National Nature Science Foundation of China under grants 61672201, 61972127.