MAEST: Open music representation n for music understanding

Efficient Supervised Training of Audio Transformers for Music Representation Learning

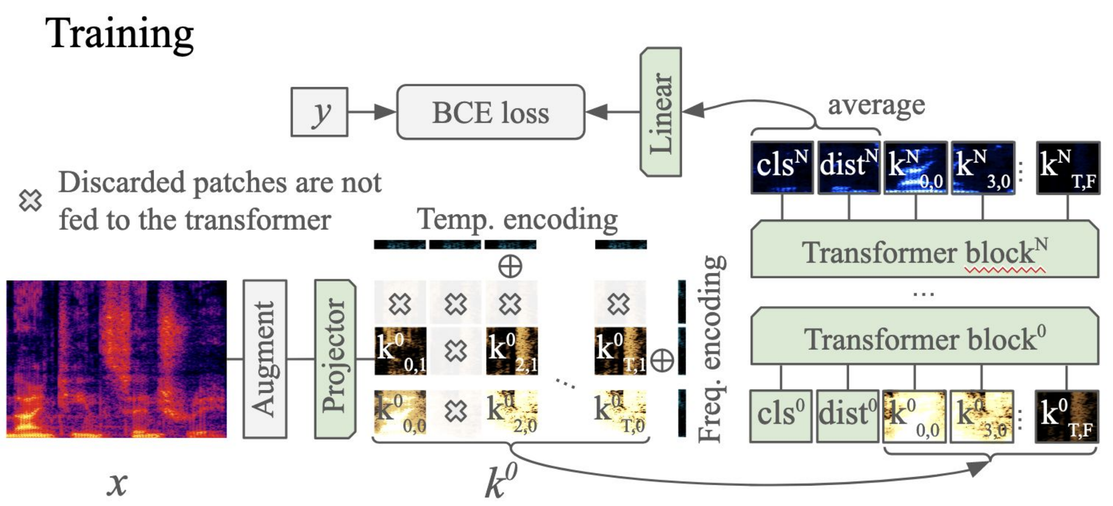
Pablo Alonso-Jiménez, Xavier Serra, Dmitry Bogdanov

Music Technology Group, Universitat Pompeu Fabra

Motivation

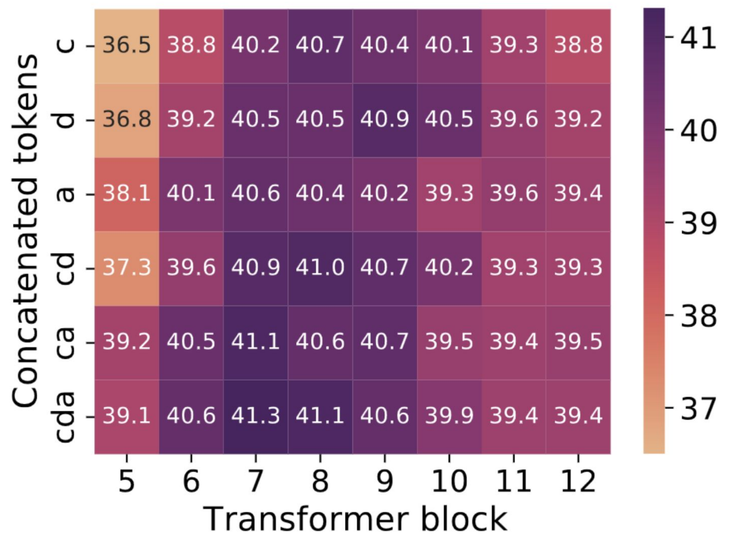
Descriptive tags are **difficult to obtain** and **noisy**. We need alternative ways of generating training targets for large music collections and suitable training approaches to develop **music representation** models.

Experimental setup



Experiments and results

Extracting embeddings from the transformer



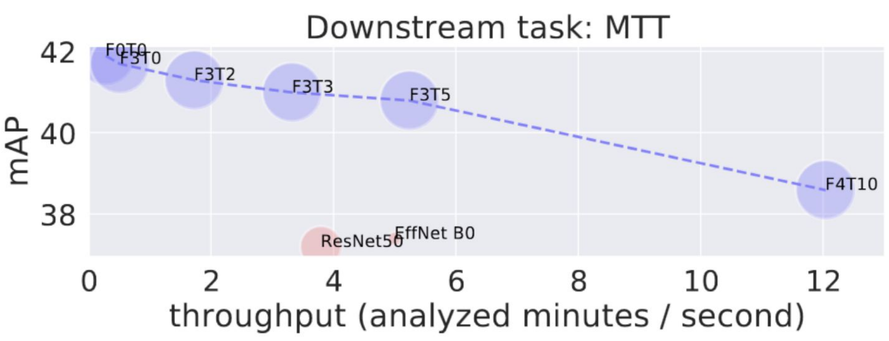
Impact of the initial weights

Model	RW	DeiT	PaSST
Pre-training task: Discogs20			
MAEST-10s	20.5	22.7	22.8
MAEST-10s-swa	20.1	23.2	23.5
Downstream task: MTT			
MAEST-10s	38.7	40.4	41.1
MAEST-10s-swa	39.0	40.2	41.0

Effect of the input segment length

Model	5s	10s	20s	30s
Pre-training task: Discogs20				
MAEST-*T*	21.1	22.8	24.8	26.1
MAEST-*T*-swa	21.3	23.5	25.8	27.0
Downstream task: MTT				
MAEST-*T*	40.8	41.1	41.2	41.7
MAEST-*T*-swa	40.9	41.0	41.2	41.5

Faster feature extraction with inference patchout



Conclusions

1. Patchout allows for **efficient training** and **inference** with Transformers
2. Transformers allow for **better music representations** than CNNs
3. We propose **MAEST**, an **publicly available music representation** model

Contact: pablo.alonso@upf.edu @pablo__alonso

[1] Park, J. Lee, J.W. Ha, and J. Nam. "Representation learning of music using artist label.," *in Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. 2018.

[2] Saeed, Aaqib, Grangier, David, and Zeghidour, Neil. "Contrastive learning of general-purpose audio representations." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021.

[3] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International Conference on Machine Learning (ICML)*. 2019.

[4] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

This research was carried out under the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

MAEST: Open music representation n for music understanding

Efficient Supervised Training of Audio Transformers for Music Representation Learning

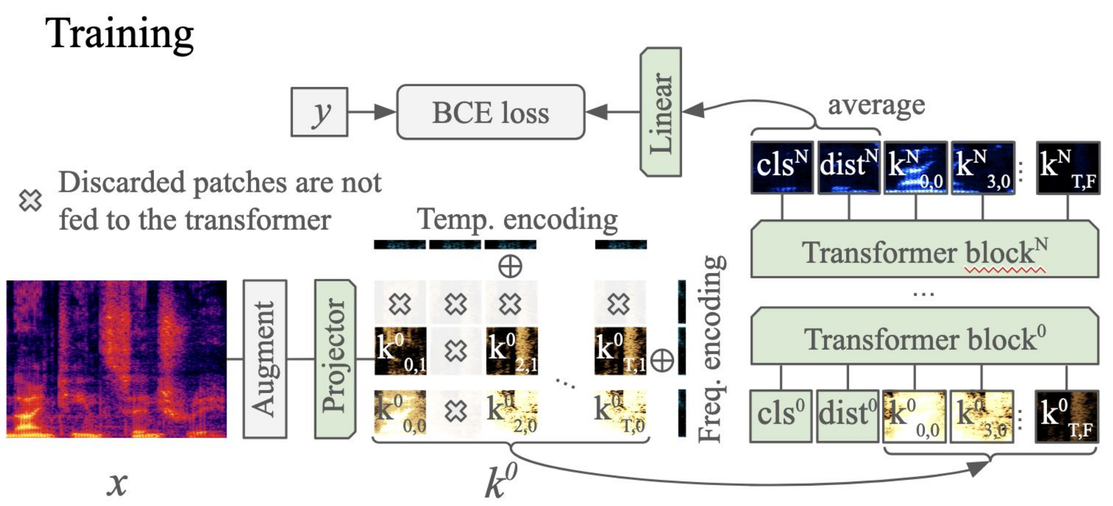
Pablo Alonso-Jiménez, Xavier Serra, Dmitry Bogdanov

Music Technology Group, Universitat Pompeu Fabra

Motivation

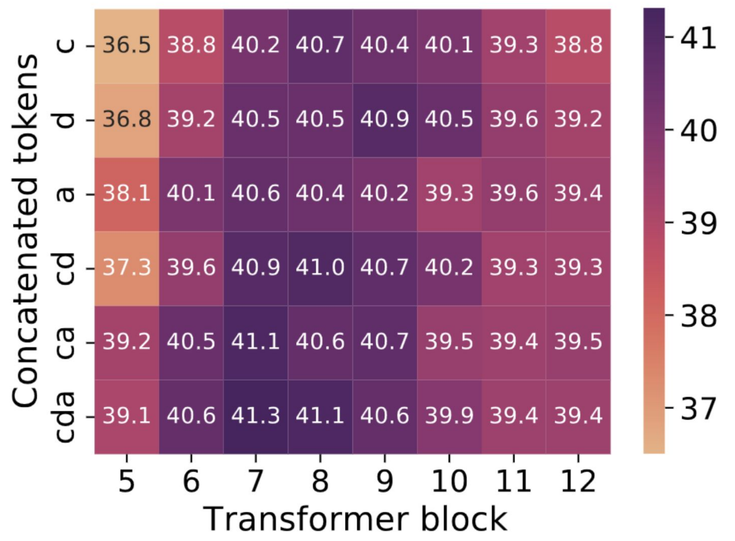
Descriptive tags are **difficult to obtain** and **noisy**. We need alternative ways of generating training targets for large music collections and suitable training approaches to develop **music representation** models.

Experimental setup



Experiments and results

Extracting embeddings from the transformer



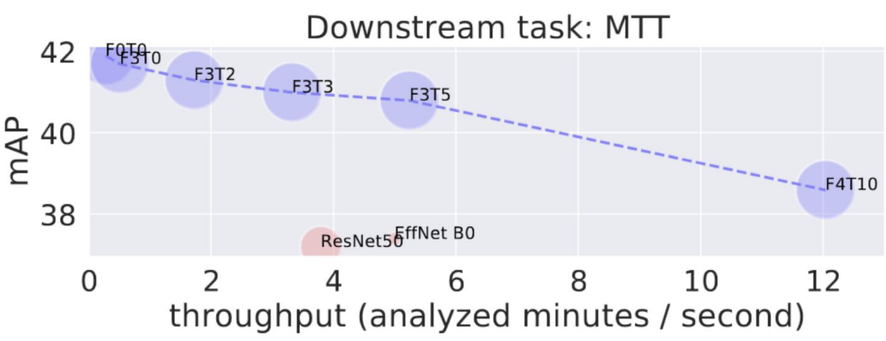
Impact of the initial weights

Model	RW	DeiT	PaSST
Pre-training task: Discogs20			
MAEST-10s	20.5	22.7	22.8
MAEST-10s-swa	20.1	23.2	23.5
Downstream task: MTT			
MAEST-10s	38.7	40.4	41.1
MAEST-10s-swa	39.0	40.2	41.0

Effect of the input segment length

Model	5s	10s	20s	30s
Pre-training task: Discogs20				
MAEST-*T*	21.1	22.8	24.8	26.1
MAEST-*T*-swa	21.3	23.5	25.8	27.0
Downstream task: MTT				
MAEST-*T*	40.8	41.1	41.2	41.7
MAEST-*T*-swa	40.9	41.0	41.2	41.5

Faster feature extraction with inference patchout



Conclusions

1. Patchout allows for **efficient training** and **inference** with Transformers
2. Transformers allow for **better music representations** than CNNs
3. We propose **MAEST**, an **publicly available music representation** model

Contact: pablo.alonso@upf.edu @pablo__alonso

[1] Park, J. Lee, J.W. Ha, and J. Nam. "Representation learning of music using artist label.," *in Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. 2018.

[2] Saeed, Aaqib, Grangier, David, and Zeghidour, Neil. "Contrastive learning of general-purpose audio representations." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021.

[3] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International Conference on Machine Learning (ICML)*. 2019.

[4] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

This research was carried out under the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

MAEST: Music Audio Efficient Spectrogram Transformer

Efficient Supervised Training of Audio Transformers for Music Representation Learning

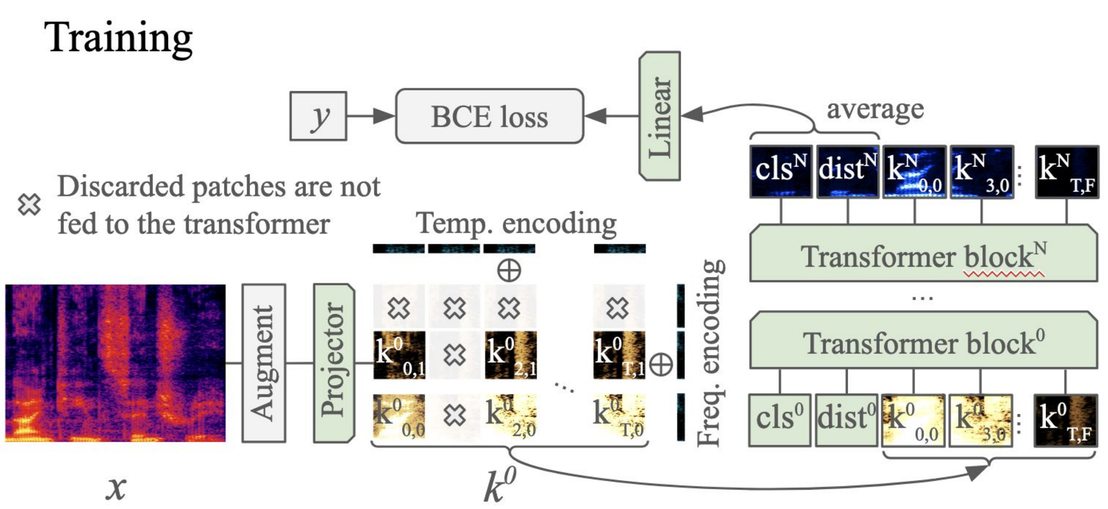
Pablo Alonso-Jiménez, Xavier Serra, Dmitry Bogdanov

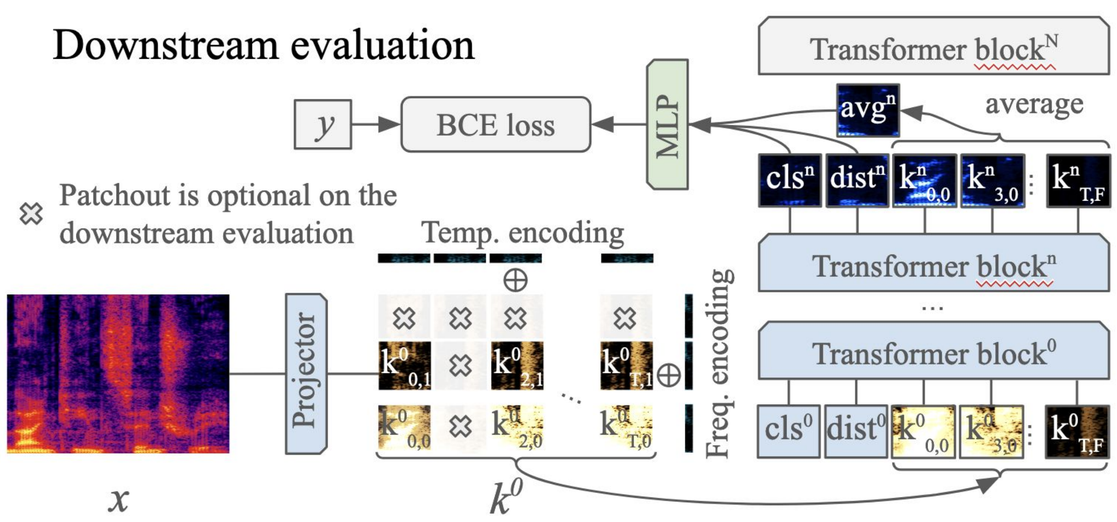
Music Technology Group, Universitat Pompeu Fabra

Motivation

Our goal is to propose **music representation models** with a focus in **semantic music description**. We rely on **convolution-free Transformers**, and propose experiments to understand which conditions **optimize** the **downstream performance**.

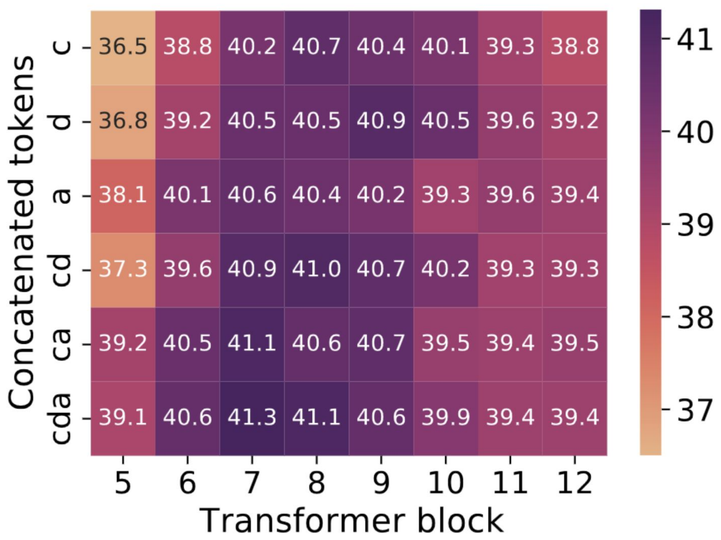
Experimental setup





Experiments and results

Extracting embeddings from the transformer



- We experimented with the class c), distillation d), and average a) tokens. Staking the three tokens produces the best performance.
- The optimal features are in the middle blocks of the transformer.

Impact of the initial weights

Model	RW	DeiT	PaSST
Pre-training task: Discogs20			
MAEST-10s	20.5	22.7	22.8
MAEST-10s-swa	20.1	23.2	23.5
Downstream task: MTT			
MAEST-10s	38.7	40.4	41.1
MAEST-10s-swa	39.0	40.2	41.0

- Starting the training from the **PaSST pre-trained weights** produces the best performance in the downstream task.

Effect of the input segment length

Model	5s	10s	20s	30s
Pre-training task: Discogs20				
MAEST-*T*	21.1	22.8	24.8	26.1
MAEST-*T*-swa	21.3	23.5	25.8	27.0
Downstream task: MTT				
MAEST-*T*	40.8	41.1	41.2	41.7
MAEST-*T*-swa	40.9	41.0	41.2	41.5

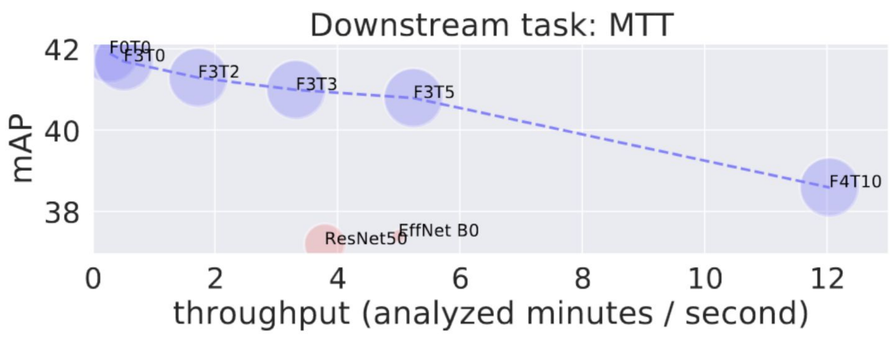
- We modify the input sequence length, from 5 up to 30 seconds finding that the results **consistently increase** in the downstream task.

Performance in downstream tasks

	MTGJ-Genre ROC	MTGJ-Genre mAP	MTGJ-Inst ROC	MTGJ-Inst mAP	MTGJ-Mood ROC	MTGJ-Mood mAP	MTGJ-T50 ROC	MTGJ-T50 mAP	MTAT ROC	MTAT mAP	MSDs ROC	MSDs mAP	MSDc ROC	MSDc mAP
State of the art														
Fully-trained	- -	- -	- -	- -	77.8 *[42]*	15.6 *[42]*	83.2 *[34]*	29.8 *[34]*	90.69 *[41]*	38.44 *[41]*	92.2 *[40]*	38.9 *[40]*	89.7 *[40]*	34.8 *[40]*
Embeddings	87.7 *[6]*	19.9 *[6]*	77.6 *[6]*	19.8 *[6]*	78.6 *[5]* †	16.1 *[5]* †	84.3 *[5]* †	32.1 *[5]* †	92.7 *[7]* †	41.4 *[5]* †	- -	- -	90.3 *[5]* †	36.3 *[5]* †
Baseline														
EffNet-B0	87.7	19.9	77.6	19.8	75.6	13.6	83.1	29.7	90.2	37.4	90.4	32.8	88.9	32.8
Our models														
MAEST-10s	88.1	21.1	79.7	22.4	77.9	15.1	84.0	31.3	91.8	41.0	91.5	36.9	88.9	32.7
MAEST-20s	88.1	21.4	79.9	22.6	77.9	15.2	**84.1**	**31.5**	91.8	41.0	92.1	39.2	89.5	34.5
MAEST-30s	**88.2**	**21.6**	**80.0**	**22.9**	**78.1**	**15.4**	84.0	**31.5**	**92.0**	**41.9**	**92.4**	**40.7**	**89.8**	**35.4**

- we evaluate our models in the MTG Jamendo Dataset, MagnaTagAtune and the Million Song Dataset.
- We find that our models are the **best performing open solution**.

Faster feature extraction with inference patchout



- We experiment applying **patchout** at **inference time.**
- It is possible to reach scenarios where the **throughput** is **higher** than the fully convolutional baselines while keeping higher performance.

Conclusions

- Patchout allows for **efficient training** and **inference** with Transformers.
- Transformers allow for **better music representations** than CNNs
- We propose **MAEST**, an **publicly available music representation** model.

Contact: pablo.alonso@upf.edu @pablo__alonso

This research was carried out under the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

MTG Music Technology Group

ESSENTIA

MUSICAL AI

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

AGENCIA ESTATAL DE INVESTIGACIÓN

MAEST: Open music representation n for music understanding

Efficient Supervised Training of Audio Transformers for Music Representation Learning

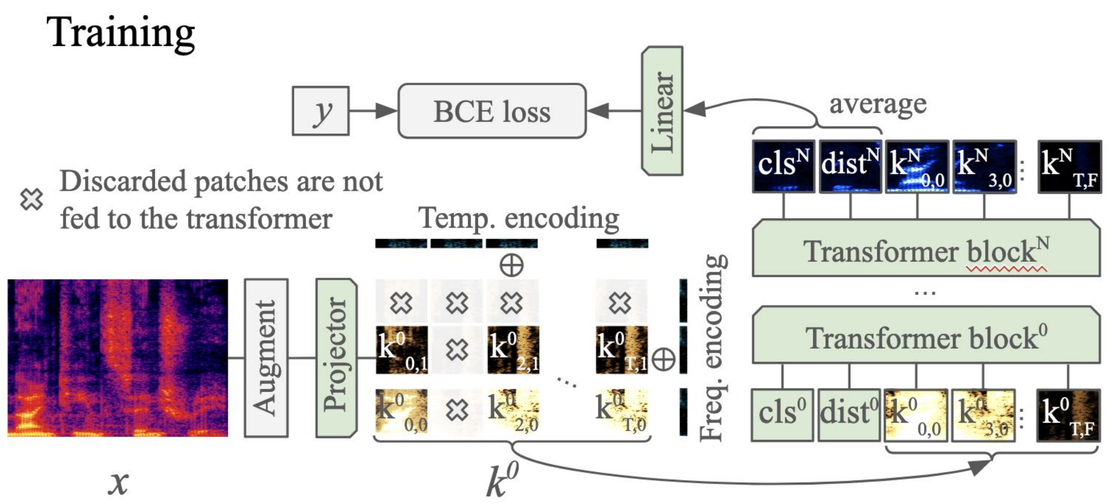
Pablo Alonso-Jiménez, Xavier Serra, Dmitry Bogdanov

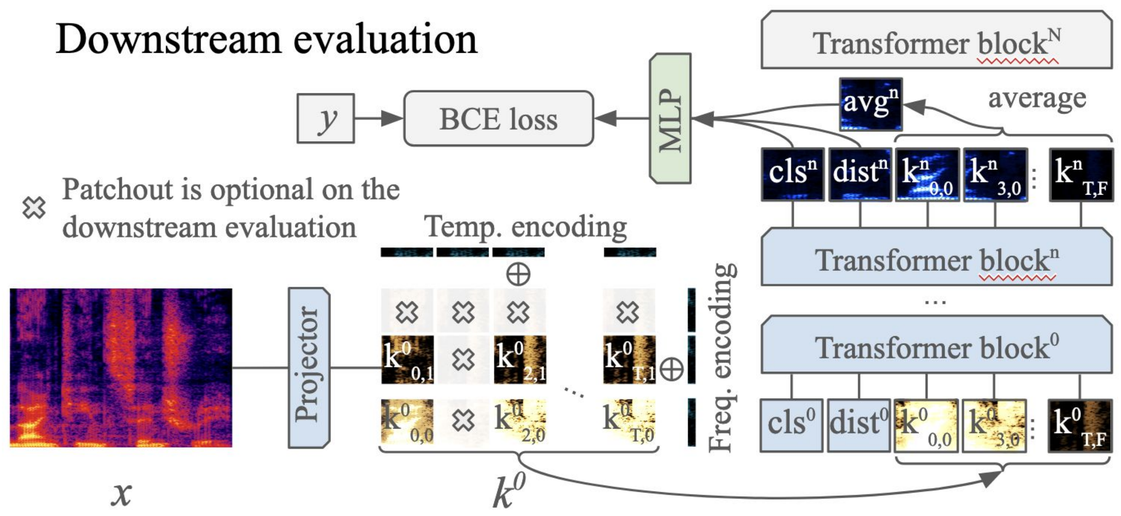
Music Technology Group, Universitat Pompeu Fabra

Motivation

Descriptive tags are **difficult to obtain** and **noisy**. We need alternative ways of generating training targets for large music collections and suitable training approaches to develop **music representation** models.

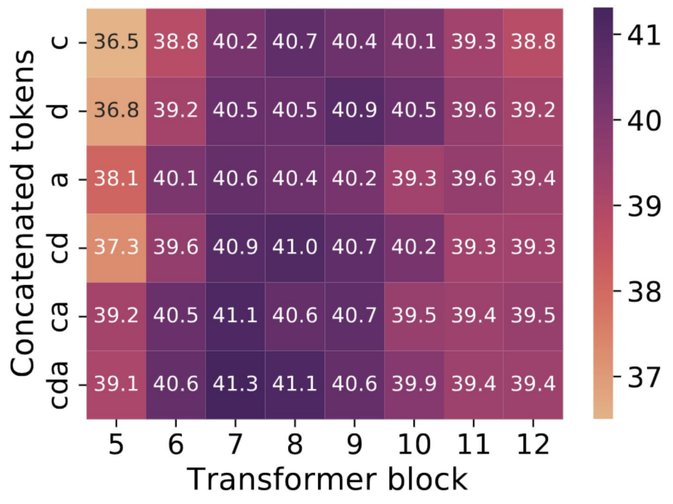
Experimental setup





Experiments and results

Extracting embeddings from the transformer



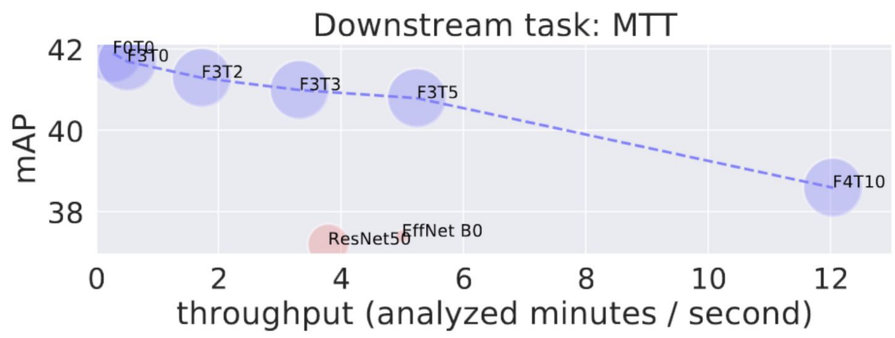
Impact of the initial weights

Model	RW	DeiT	PaSST
Pre-training task: Discogs20			
MAEST-10s	20.5	22.7	22.8
MAEST-10s-swa	20.1	23.2	23.5
Downstream task: MTT			
MAEST-10s	38.7	40.4	41.1
MAEST-10s-swa	39.0	40.2	41.0

Effect of the input segment length

Model	5s	10s	20s	30s
Pre-training task: Discogs20				
MAEST-*T*	21.1	22.8	24.8	26.1
MAEST-*T*-swa	21.3	23.5	25.8	27.0
Downstream task: MTT				
MAEST-*T*	40.8	41.1	41.2	41.7
MAEST-*T*-swa	40.9	41.0	41.2	41.5

Faster feature extraction with inference patchout



Conclusions

1. Patchout allows for **efficient training** and **inference** with Transformers
2. Transformers allow for **better music representations** than CNNs
3. We propose **MAEST**, an **publicly available music representation** model

Contact: pablo.alonso@upf.edu @pablo__alonso

[1] Park, J. Lee, J.W. Ha, and J. Nam. "Representation learning of music using artist label.," *in Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. 2018.
[2] Saeed, Aaqib, Grangier, David, and Zeghidour, Neil. "Contrastive learning of general-purpose audio representations." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021.
[3] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International Conference on Machine Learning (ICML)*. 2019.
[4] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

This research was carried out under the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.