AUDIO-VISUAL FUSION AND CONDITIONING WITH NEURAL NETWORKS FOR EVENT RECOGNITION

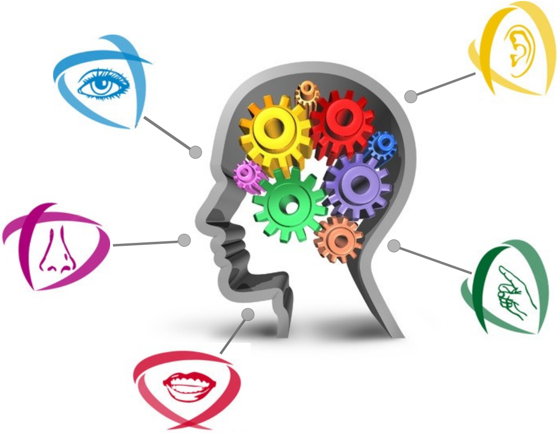
Mathilde Brousmiche [1] [2]   Jean Rouat [2]   Stéphane Dupont [1]

[1]Numediart Institute
University of Mons
Belgium

[2]Necotis Lab
University of Sherbrooke
Canada

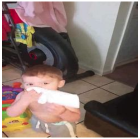
MLSP, October 2019

Multimodality



Problem setting : Audio-visual Event Classification

Subset of kinetics[1]:


blowing_nose

clapping

crying

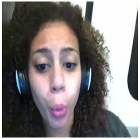
finger_snapping

playing_drums

playing_guitar

sneezing

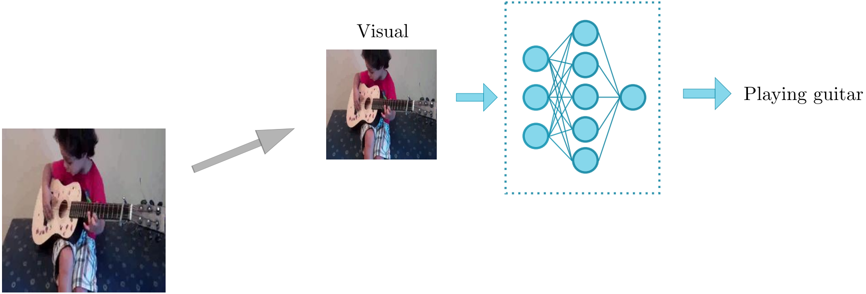
using_computer


whistling

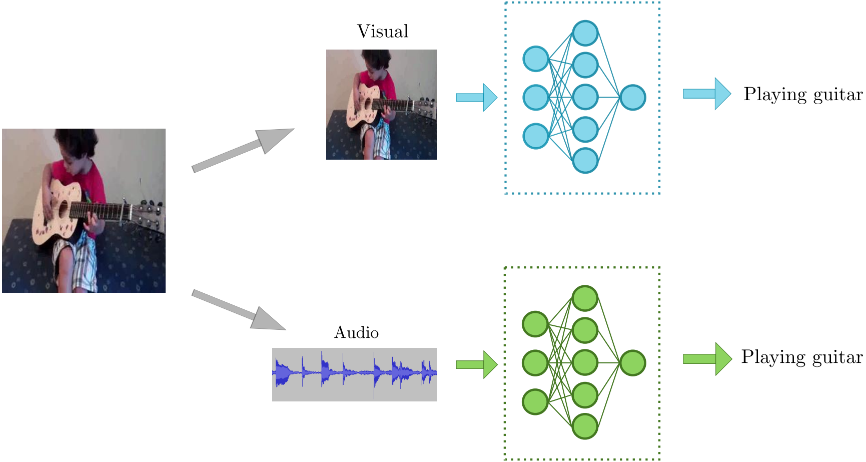
yawning

[1] W. Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

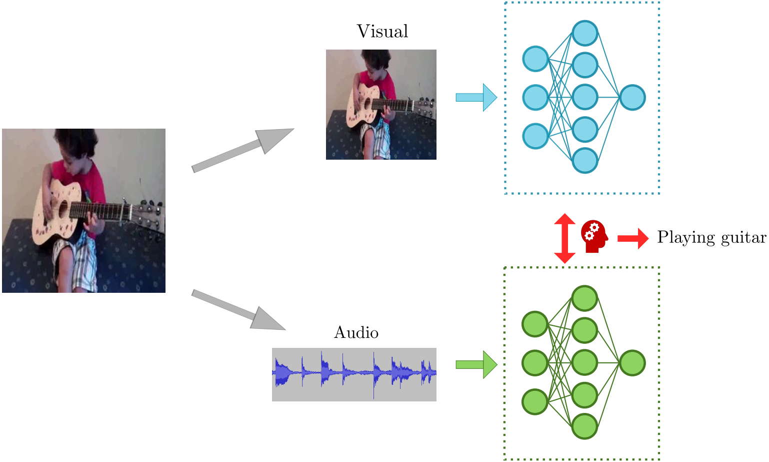
Problem setting : Audio-visual Event Classification



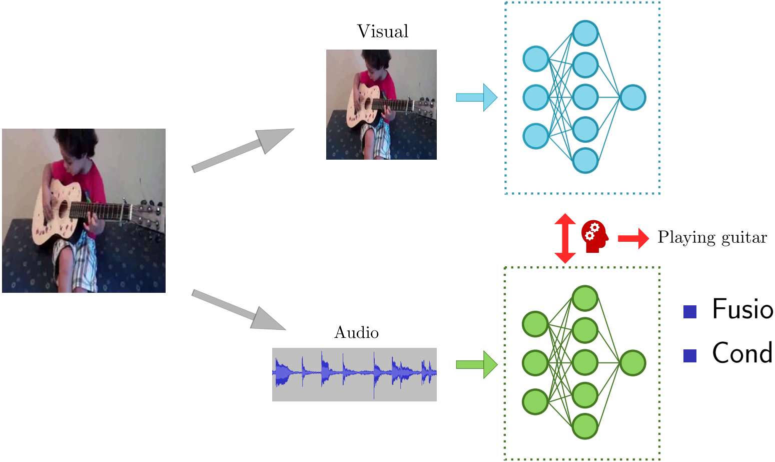
Problem setting : Audio-visual Event Classification



Problem setting : Audio-visual Event Classification



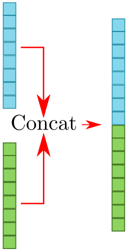
Problem setting : Audio-visual Event Classification



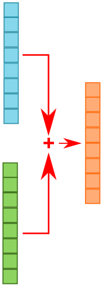
- Fusion ?
- Conditioning ?

Techniques of fusion

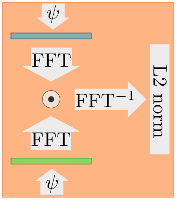
Concatenation



Element-wise addition

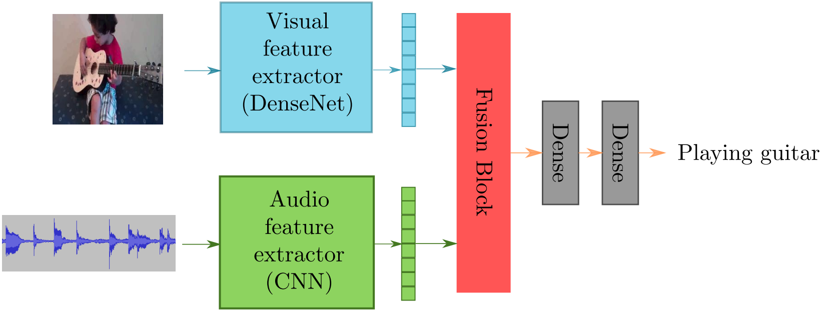


Multimodal Compact Bilinear pooling (MCB)[1]

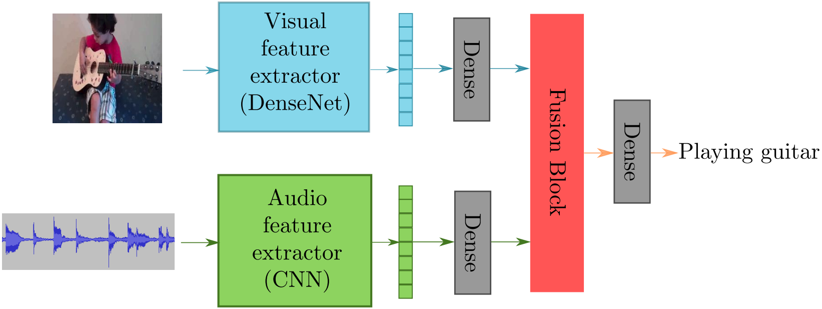


[1] Y. Gao et al. "Compact bilinear pooling". In: *IEEE Proc. CVPR*. 2016, pp. 317–326.

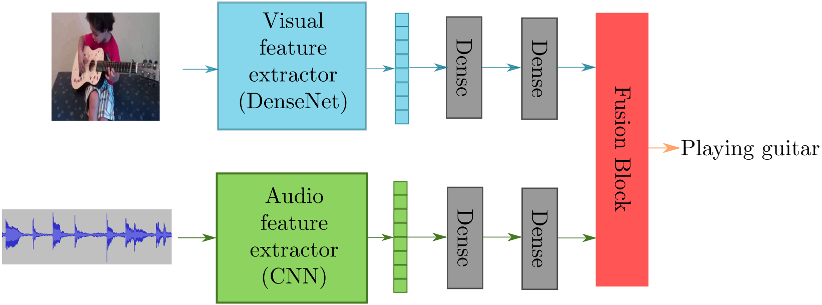
Fusion levels : Early fusion


Visual feature extractor (DenseNet)
Audio feature extractor (CNN)
Fusion Block
Dense
Dense
Playing guitar

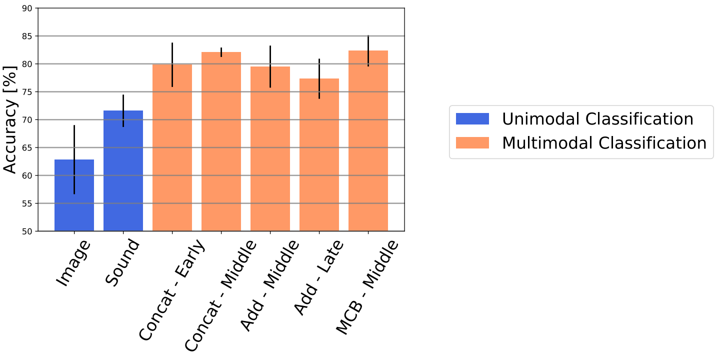
Fusion levels : Middle fusion


Visual feature extractor (DenseNet)
Dense
Fusion Block
Dense
Playing guitar
Audio feature extractor (CNN)
Dense

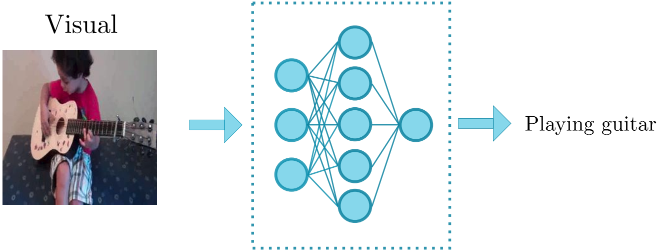
Fusion levels : Late fusion


Visual feature extractor (DenseNet)
Dense
Dense
Audio feature extractor (CNN)
Dense
Dense
Fusion Block
Playing guitar

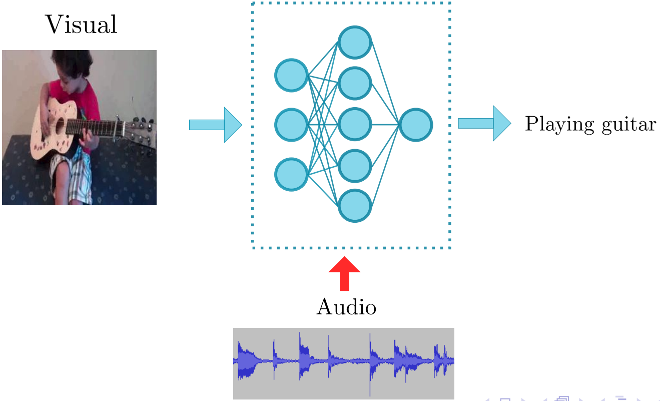
Audio-visual Fusion efficiency



What is conditioning ?



What is conditioning ?

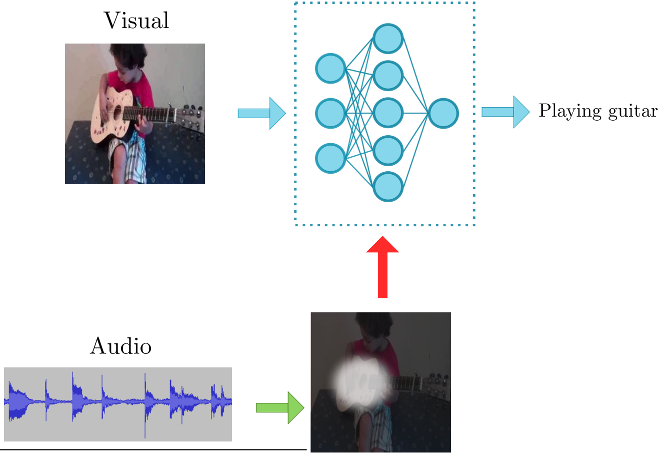


Visual

Playing guitar

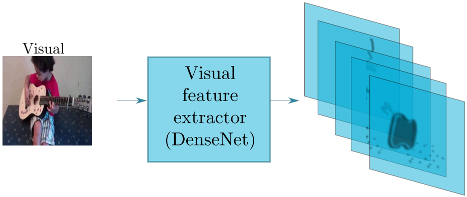
Audio

Modalities conditioning with a attention model[2]

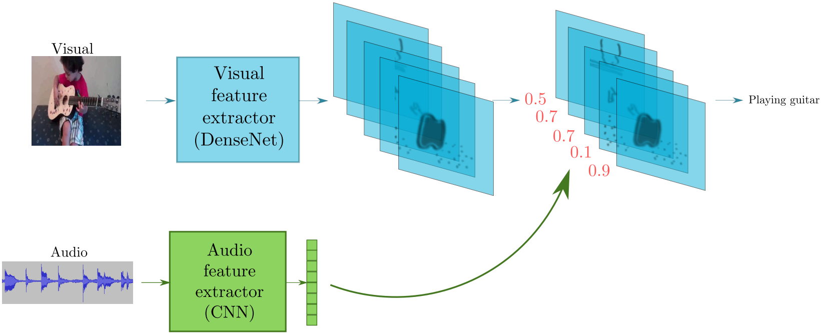


[2] Y. Tian et al. "Audio-visual event localization in unconstrained videos". In: *Proc. of ECCV*. 2018, pp. 247–263.

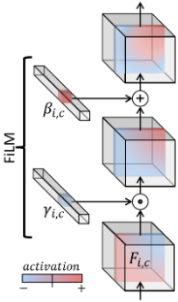
Proposal : Visual feature map modulation with audio information



Proposal : Visual feature map modulation with audio information


Visual
Visual feature extractor (DenseNet)
0.5
0.7
0.7
0.1
0.9
Playing guitar
Audio
Audio feature extractor (CNN)

Feature-wise Linear Modulation (FiLM)[3]



$\gamma_{i,c}$ and $\beta_{i,c}$ modulate the activations $\mathbf{F}_{i,c}$:

$$FiLM(\mathbf{F}_{i,c}|\gamma_{i,c},\beta_{i,c}) = \gamma_{i,c}\mathbf{F}_{i,c} + \beta_{i,c}$$

where

$$\gamma_{i,c} = f_c(x_i) \qquad \beta_{i,c} = h_c(x_i)$$

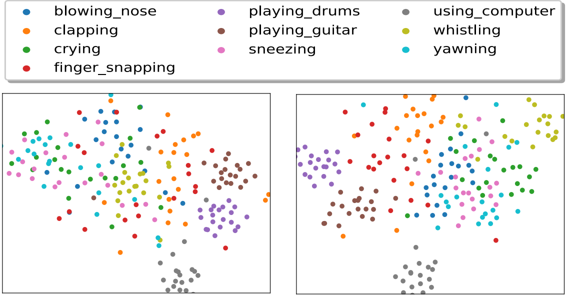
f and h can be arbitrary functions

[3] E. Perez et al. "Film: Visual reasoning with a general conditioning layer". In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

Contribution of FiLM in audio-visual event classification

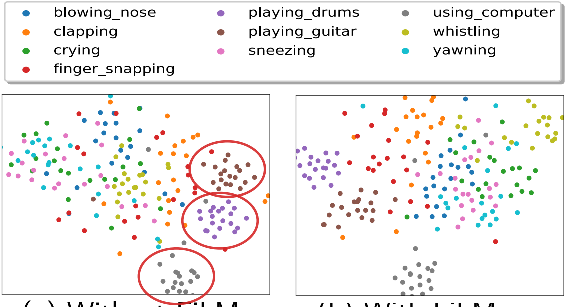
Accuracy [%]	Image	Sound
Without FiLM modulation	61.00 ± 5.11	66.67 ± 4.60
With FiLM modulation	75.75 ± 5.35	75.75 ± 3.14

Better embedding clustering with FiLM



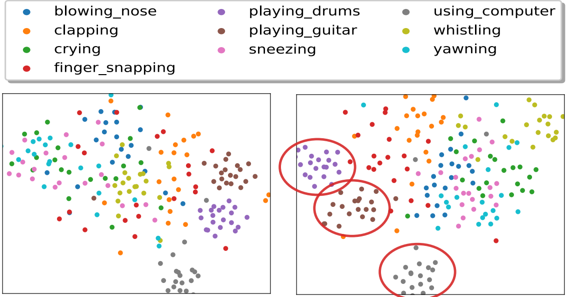
(a) Without FiLM (b) With FiLM

Better embedding clustering with FiLM



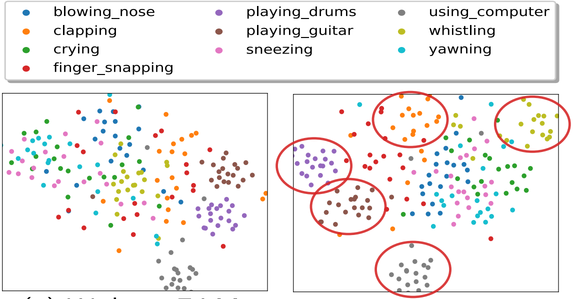
(a) Without FiLM (b) With FiLM

Better embedding clustering with FiLM



(a) Without FiLM (b) With FiLM

Better embedding clustering with FiLM



(a) Without FiLM (b) With FiLM

Conclusion

- Relevant information for event recognition exists both in visual and audio modalities.
- Exploiting both audio and visual modalities through fusion or conditioning improves event recognition performance
- The use of FiLM layers allows exploiting both audio and visual modalities without an explicit implementation of the fusion

Future Work

- Test another conditioning method based on multimodal Long Short-Term Memory (LSTM) neural networks
- Analyze the robustness of all methods in the presence of noise as well as in the absence of one modality.

Thank you !