Multimodal Emotion Recognition in Videos

Student: Galbenus Andrei Emanuel
Coordonator: Sl. Dr. Ing. Dumitru Clementin Cercel

Introduction

- Objective: Automated system, remove subjective human factor
- Problem: Different emotional expressiveness styles
- Motivation:
  - Available datasets (e.g. CMU-MOSEI, EmoReact)
  - Computational power
  - Useful in: medicine, retail, educational

https://www2.informatik.uni-hamburg.de/wtm/OMG-EmotionChallenge/

Related Work

- Z. Zheng et al., GammaLab, 2018
    - A: SoundNet
    - V: VGG16
- D. Deng et al., HKUST, 2018
    - A: openSMILE + LL
    - V: OpenFace + VGG16
- A. Triantafyllopoulos et al., audEERING, 2018
    - A: openSMILE + BiLSTM
    - V: MTCNN + VGG16

Datasets (One-Minute Gradual Emotion)

- 2018
- 420 YouTube videos
- 1 minute per video, on average, split in utterances
- Approximately 10 hours of content (monologues, auditions, dialogues)
- Valence and arousal
- 3 baseline models
- CCC (Concordance Correlation Coefficient)
- Train/Validation/Test: 1790/550/1450

$$ccc(y, \hat{y}) = \frac{2\rho(y,\hat{y})\sigma_y\sigma_{\hat{y}}}{\sigma_y^2\sigma_{\hat{y}}^2+(\mu_y-\mu_{\hat{y}})^2}$$

$$\rho(y, \hat{y}) = \frac{\sum(y-\mu_y)(\hat{y}-\mu_{\hat{y}})}{\sqrt{\sum(y-\mu_y)^2(\hat{y}-\mu_{\hat{y}})^2}}$$

https://www2.informatik.uni-hamburg.de/wtm/OMG-EmotionChallenge/

Data processing (A)

- .wav files/3s/16khz
- Spectrograms/MS/LMS (STFT)
  - Window: 2048
  - Step: 512
  - Mels: 90
- 40 MFCCs (Mel Frequency Cepstral Coef)

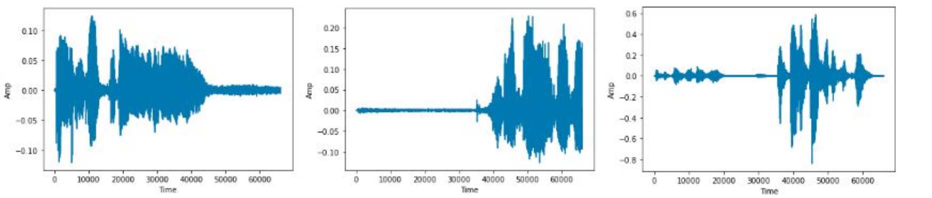


Figura 1: Semnale audio neprocesate (tristete/fericire/nervozitate)

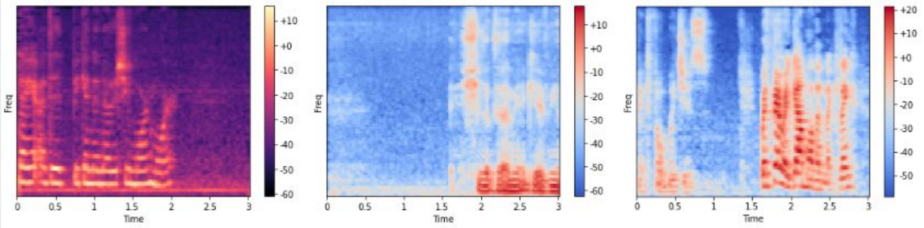


Figura 3: Spectrograme Mel logaritmate (tristete/fericire/nervozitate)

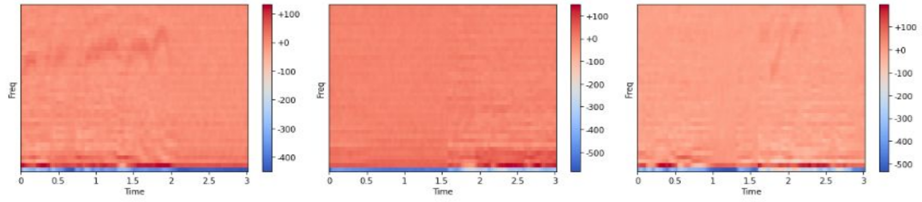
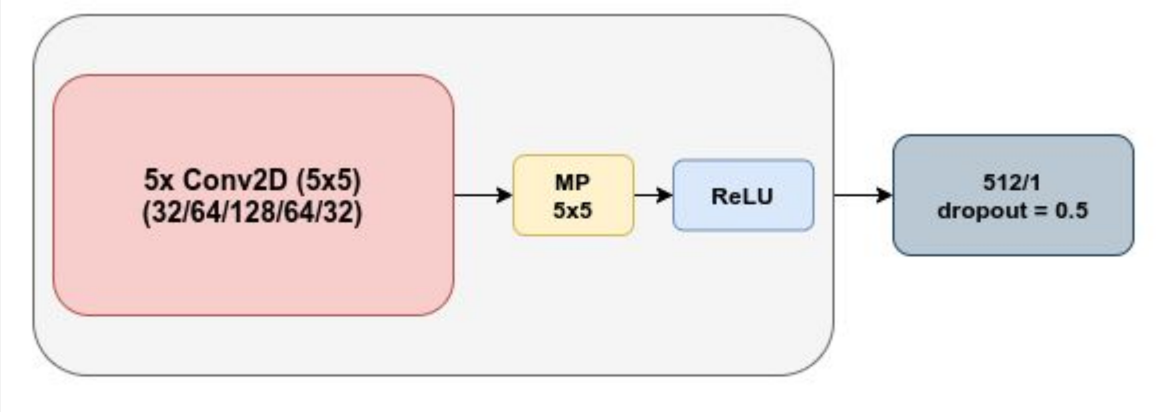


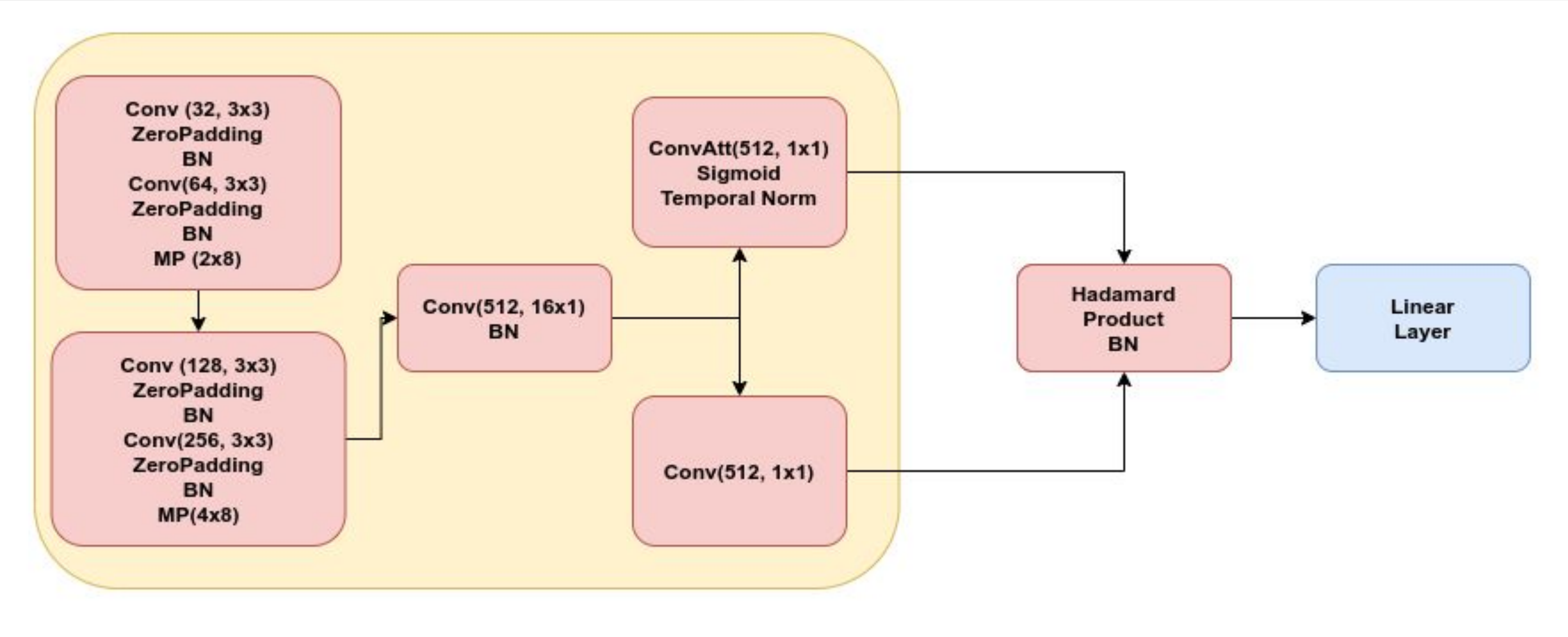
Figura 4: Spectrograme MFCC (40) (tristete/fericire/nervozitate)

Audio Models (1)

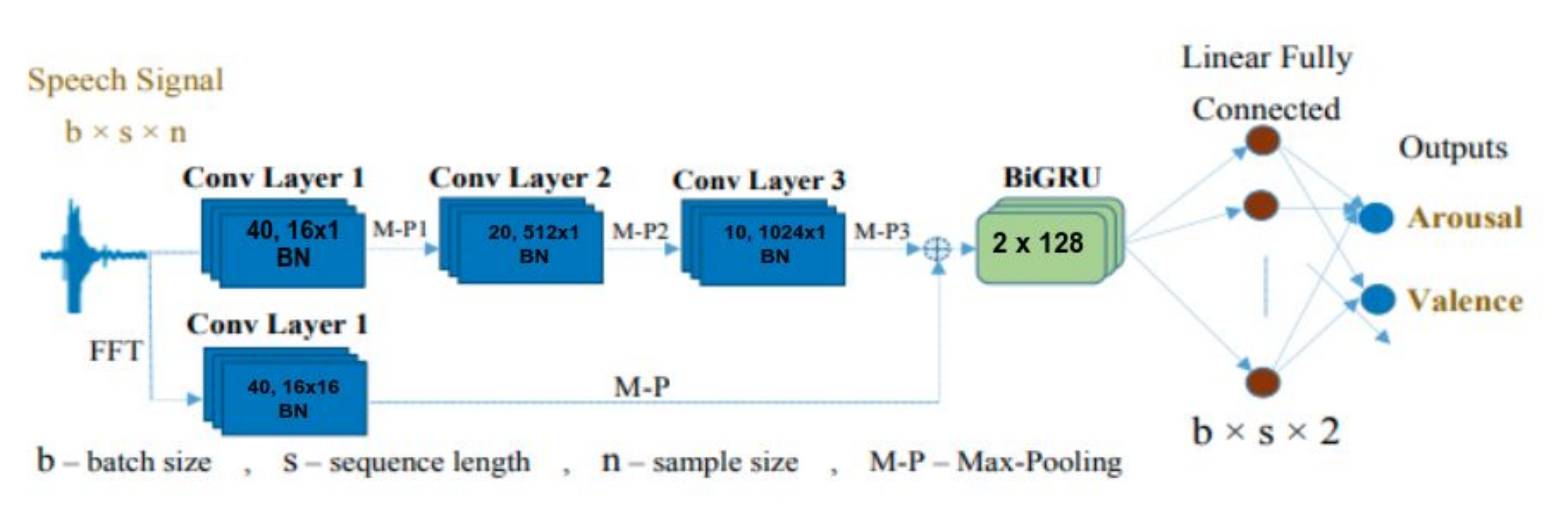


- VGG16/ResNet50/AlexNet (P/NP)

Audio Models (2)



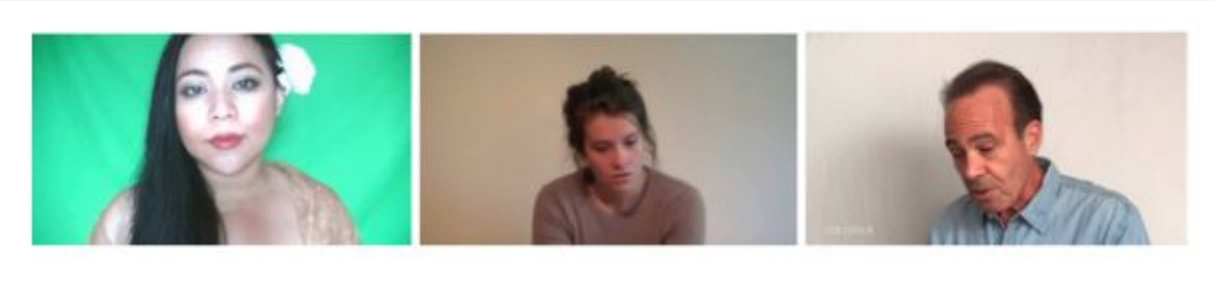
Audio Models (3)

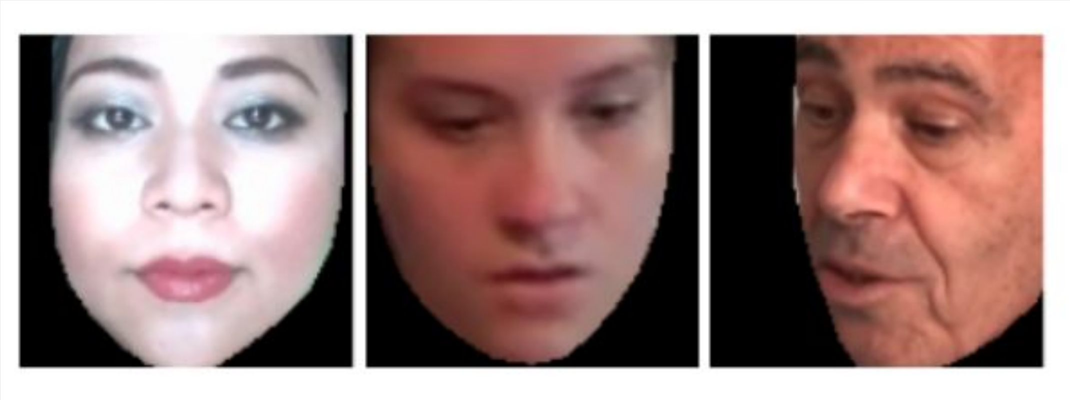


https://ecai2020.eu/papers/574_paper.pdf

Data processing (V)

- 16 frames extracted
- Cropped
- Aligned faces
- OpenFace





Visual Models

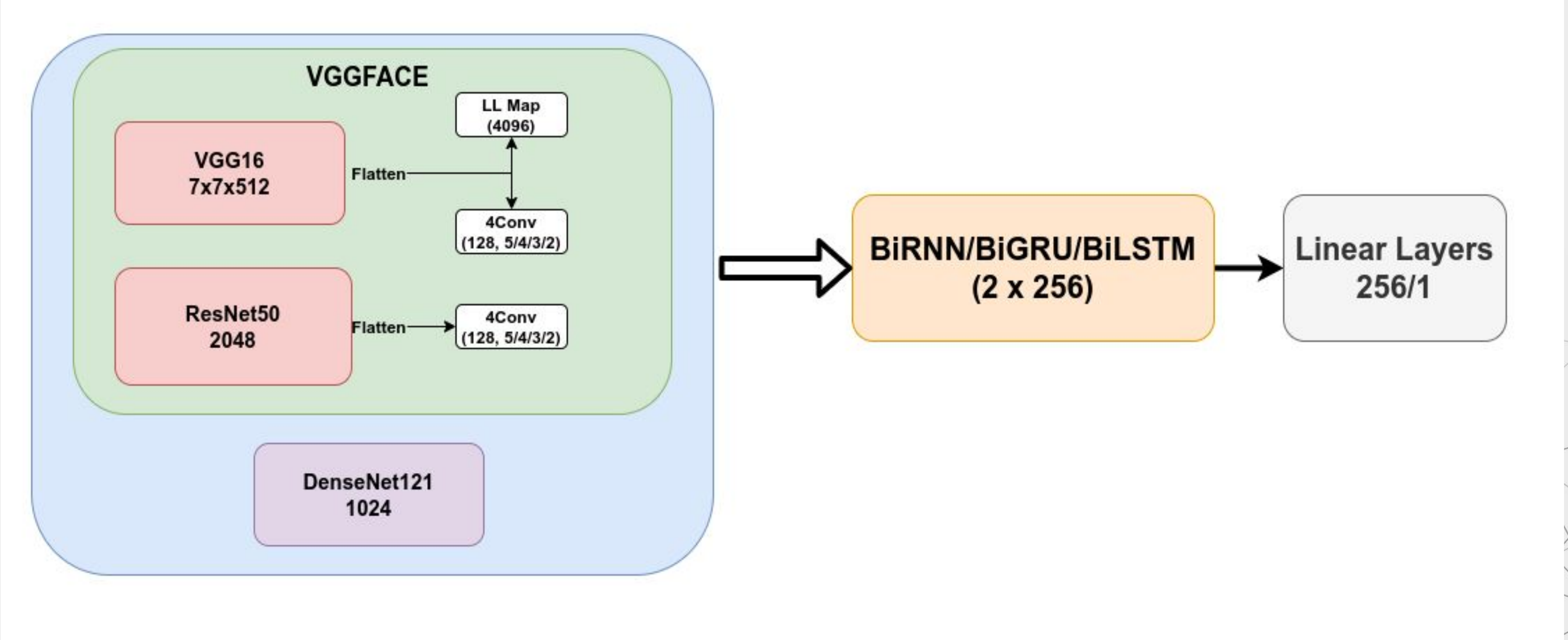


Table 3: Models results comparison

Model	Arousal(CCC)	Valence(CCC)
GammaLab (VA)	0.361	0.498
Audeering (VA)	0.286	0.368
HKUST (VAT)	0.276	0.359
UMONS (VAT)	0.175	0.262
ADSC (VA)	0.236	0.442
ADSC (V)	0.244	0.437
Audeering (A)	0.292	0.361
iBug (V)	0.130	0.400
Baseline (A)	0.08	0.10
BaseLine (V)	0.12	0.23
VGG16(P)(A)	0.285	0.195
ResNet50+GRU (V)	0.180	0.446

Model	Arousal	Valence	Total
5Conv2D	0.211	0.122	0.333
ResNet50(P)	0.251	0.205	0.456
ResNet50(NP)	0.194	0.210	0.404
VGG16(P)	0.285	0.195	0.480
VGG16(NP)	0.078	0.251	0.329
AlexNet(P)	0.159	0.152	0.311
AlexNet(NP)	0.248	0.077	0.325
Conv + ATT	0.066	0.047	0.113
Conv 1D + 2D	0.199	0.041	0.240

Model	Arousal	Valence	Total
ResNet50 + RNN	0.112	0.322	0.434
ResNet50 + GRU	0.18	0.446	0.626
ResNet50 + LSTM	0.249	0.325	0.574
VGG16 + RNN	0.194	0.303	0.497
VGG16 + GRU	0.274	0.254	0.528
VGG16 + LSTM	0.159	0.347	0.506
DenseNet121 + RNN	0.083	0.133	0.216
DenseNet121 + GRU	0.063	0.052	0.115
DenseNet121 + LSTM	0.098	0.021	0.119
ResNet50 + 4Conv1D	0.188	0.25	0.438
VGG16 + 4Conv1D	0.113	0.345	0.458

Hyperparams

- Tanh/Sigmoid
- SGD (Nesterov, momentum, 1e-3 - 1e-6)
- Batch size 8 - 32
- Colab (12GB Testa K80, 12 GB RAM) / Locally (4GB Nvidia 960M, 16GB)
- Python 3.7
- PyTorch 1.9.0/1.4.0

Conclusion

- Deep convolutional networks
- GRU - smaller datasets
- LMS - best results
- Improvements:
    - Bigger dataset
    - Multimodal model
    - Additional cues (textual, kinetic)

THANK YOU!

Any questions?

References

[1] Zheng, Z., Cao, C., Chen, X., and Xu, G. (2018). Multimodal emotion recognition for one-minute-gradual emotion challenge. arXiv preprint arXiv:1805.01060.

[2] Deng, D., Zhou, Y., Pi, J., and Shi, B. E. (2018). Multimodal utterance-level affect analysis using visual, audio and text features. arXiv preprint arXiv:1805.00625.

[3] Triantafyllopoulos, A., Sagha, H., Eyben, F., and Schuller, B. (2018). audEERING's approach to the One-Minute-Gradual emotion challenge. arXiv preprint arXiv:1805.01222.