Team 19: Multilingual Multimodal Sentiment Analysis

Yuxin Pei, Zhiyi Kuang, Zhe Chen, Yuchen Zhou, Shuhao Zhang

Carnegie Mellon University

Motivation & Objectives

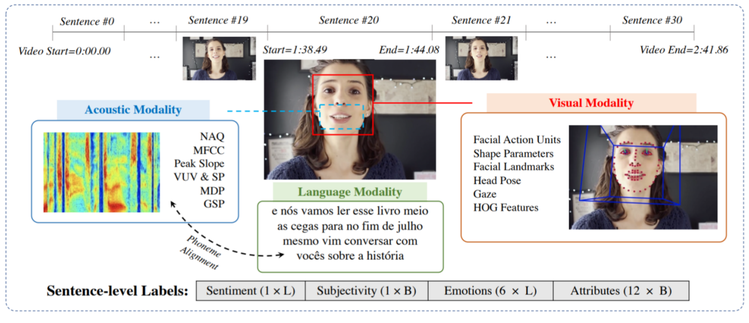
Multimodal sentiment analysis, which is the idea of using language, acoustic, and visual features to detect sentiment, is a well-established research question. However, very few work has been done on languages other than English, partly because there was no cross-linguistic sentiment analysis dataset available in the past, and because low-resource languages are so many and diverse that we cannot build models on every of them.

With the release of CMU-MOSEAS[1] dataset, which contains sentiment labels for video clips in four European languages, we propose to build a multilingual multimodal sentiment detection model by transfer learning. Specifically, we generate **cross-linguistic paired texts** by means of translation and adopt **Noise Contrastive Estimation (NCE)** as the learning objective. We also explore **Adaptive MML** that filters out non-representative modalities in the flow for efficient and robust sentiment analysis.

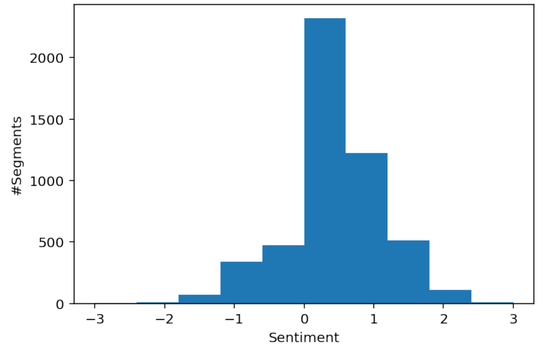
Dataset

CMU-MOSEAS contains 10,000 video segments for each of four European languages - Spanish, Portuguese, German, and French. French and Spanish are picked as training and test languages.

Each video segment is scored by multiple human raters according to the speaker's sentiment. For each video segment, we take the average scores from all raters as its label.

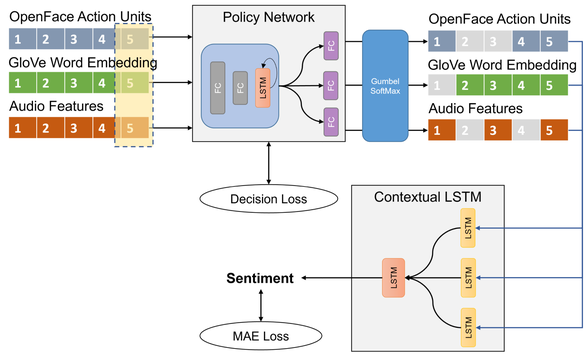


Sentiment label distribution is highly skewed towards neutral sentiments. Mildly positive sentiments are more common than negative sentiments. Therefore, we removed sentiment $\in [-0.5, 0.5]$ for most of our works.



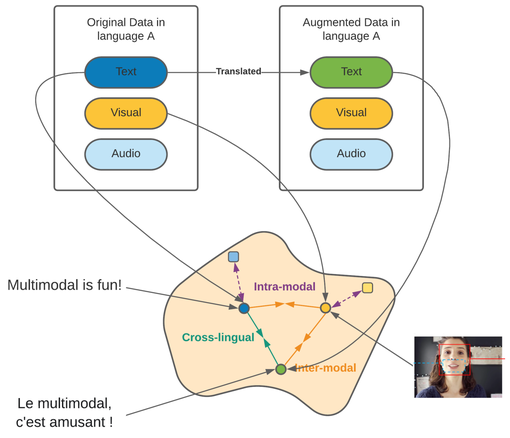
Research Ideas

- Few-shot Transfer Learning: AdaMML [3] + Contextual LSTM [4]



Goal: use a much smaller subset of Spanish training data together with full French training data to reach similar level of model accuracy as compared to learning from the full Spanish training set. Use an adaptive policy to select relevant modalities for prediction.

- Noise Contrastive Estimation (NCE) Objective + Translation Augmentation.



Goal: learn a multimodal cross-lingual representation that is transferable across languages. Key intuition: conditioned on a video V, the multilingual descriptions (translated pairs (T_1, T_2) in our case) should be semantically similar [2]. We set the visually-pivoted cross-lingual NCE loss as

$$\mathcal{L}^{\text{cross}} = \mathcal{L}(T_1|V, T_2|V)$$

We optimize the model by

$$\min MSE + \mathcal{L}^{\text{cross}}$$

Discussions

- Gumbel SoftMax gating can improve computational efficiency of fusion model.
- Fine-tuning our model on a few samples from a new language can make our AdaMML model generalize on the new language.
- NCE objective with translation augmentation allows us to learn cross-lingual representations in a monolingual setting that enables zero-shot transfer.

Results

- AdaMML + Contextual LSTM

Training Data	Eval Data	DL	r(↑)	MAE(↓)
FR	FR	N.A.*	0.38	0.54
FR	FR	2	0.31	0.53
ES	ES	2	0.28	0.57
FR	ES	2	0.23	0.60
FR + 13% ES	ES	2	0.27	0.53
FR + 1% ES	ES	2	0.28	0.56

Table 1:Model performances evaluated on CMU-MOSEAS dataset. DL: decision loss.
* Depends on where dropout is placed

- NCE + Translation DA

Training Data	Eval Data	r(↑)	MAE(↓)
FR only	FR	0.35	0.50
FR only	ES	0.28	0.83
FR + DA	FR	0.39	0.49
FR + DA	ES	0.33	0.77
FR + ES	FR	0.48	0.46
FR + ES	ES	0.54	0.55

Table 2:Model performances evaluated on CMU-MOSEAS dataset

r stands for Pearson's correlation coefficient.

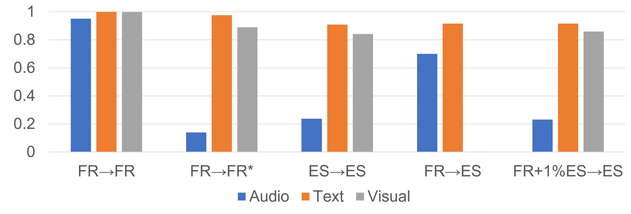


Figure 1:AdaMML Modality Utilization Rate

References

[1] A. Bagher Zadeh, Y. Cao, S. Hessner, P. P. Liang, S. Poria, and L.-P. Morency, *CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French*, in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, Association for Computational Linguistics, pp. 1801–1812.

[2] P. Huang, M. Patrick, J. Hu, G. Neubig, F. Metze, and A. G. Hauptmann, *Multilingual multimodal pre-training for zero-shot cross-lingual transfer of vision-language models*, CoRR, abs/2103.08849 (2021).

[3] R. Panda, C.-F. Chen, Q. Fan, X. Sun, K. Saenko, A. Oliva, and R. Feris, *AdaMML: Adaptive Multi-Modal Learning for Efficient Video Recognition*, in International Conference on Computer Vision (ICCV), 2021.

[4] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, *Context-dependent sentiment analysis in user-generated videos*, in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada, July 2017, Association for Computational Linguistics, pp. 873–883.

Carnegie Mellon University