Transfer Learning with Joint Fine-Tuning for Multimodal Sentiment Analysis

Guilherme L. Toledo and Ricardo M. Marcacini
Laboratory of Computational Intelligence
University of São Paulo

Motivation

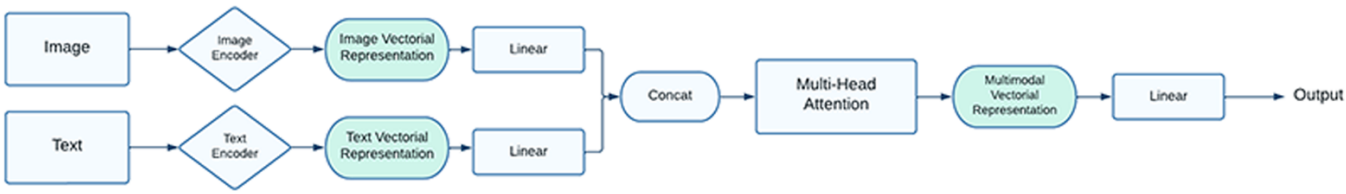
- Multimodal sentiment analysis is proposed to deal with complex emotion recognition scenarios.
- Recent state-of-the-art methods led to promising results, but pre-training became computationally expensive.
- How to fine-tune different pre-trained unimodal models considering a multimodal objective for sentiment analysis tasks?

Proposal

- Instead of looking for higher performance, we decided to take a step back and prioritize efficiency.
- We introduce a transfer learning approach using joint fine-tuning for sentiment analysis.
- Allowing the fine-tune of both pre-trained models as a single loss function makes this a flexible, simple and efficient aproach for multimodal tasks.

Architecture

- Use of pre-trained unimodal models to improve efficiency in multimodal classification.
- Multi-Head Attention as the fusion mechanism for the modalities.
- Both unimodal encoders are fine-tuned by one single loss function.



Results

Method	MVSA	HatefulMemes	Number of Parameters
SVM (concat)	0.64	0.54	89.6M
ResNet	0.63	0.53	23.5M
Distilbert	0.80	0.72	66.4M
ResNet + Distilbert	0.79	0.69	90.0M
CLIP	0.81	0.74	152M
Our proposal	0.81	0.74	90.3M

Conclusion

- High parameter reduction.
- Efficieny improvement.
- Competitive results to CLIP.
- Effective fusion method via attention mechanism.

Future Work

- Extensive evaluation of this strategy considering different tasks.
- Incorporate multimodal explainabilty in this approach.