Multimodal Humor Detection Final Report

Peilin Li
peilinl@usc.edu
University of Southern California
Los Angeles, California, USA

Jiaqi Lu
jiaqilu@usc.edu
University of Southern California
Los Angeles, California, USA

Yihang Yin
yihangyi@usc.edu
University of Southern California
Los Angeles, California, USA

Jayavibhav Niranjan Kogundi
jniranja@usc.edu
University of Southern California
Los Angeles, California, USA

ACM Reference Format:
Peilin Li, Jiaqi Lu, Yihang Yin, and Jayavibhav Niranjan Kogundi. 2025. Multimodal Humor Detection Final Report. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Abstract

The project is using UR-Funny dataset to address multimodel humor dectction problem. Based on Figure 1, the project begins with a filtering stage to improve the data quality. After filtering, features are extracted from each modality using pretrained models, Vit for video, Wav2Vec2.0 for audio, and BERT for text. The initial experiments focused on unimodal and bimodal fusion to analyze the baseline performance. For the final, a modified trimodal architecture, Cross-Modal MoE++ is introduced. It combines bilinear interactions and Mixture of Experts to better capture humor variance. Results show an excellent improvement in humor classification performance, indicating the importance of customized fusion strategies in multimodal humor detection.

2 Problem Description

Humor is a creative and unique part of human communication that plays an important role in social interaction. However, due to the complexity of humor that includes harder to detect types like sarcasm, exaggeration, irony, etc. Correctly detecting humor using machine learning model then will provide much need part of understanding for HCI in general. While certain types of humor are hard to detect or even define, we can always rely on human response to recognize if humor is present in a conversation. For this project we aim to use a multimodal deep learning modal on a human-labled video dataset in hope to capture all kinds of humor and correctly recognize the presence of humor.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

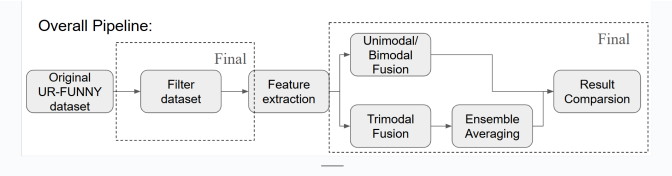

Figure 1: Project pipeline

3 Literature Review

3.1 Multimodal Humor Detection

The UR-FUNNY dataset explicitly addresses multimodal humor detection through providing structured multimodal data consists of text, acoustic and visual components in a context-punchline format. The paper also provided a baseline using a contextualized extention of Memory Fusion Network (2018) that first parse each modality through a Unimodal Context network consists of LSTM layers, then through a Multimodal Context Network with transformers, and finally a MFN is applied [5]. Building on this multimodal setup, the Humor Knowledge Enriched Transformer (HKT) further infuses external humor-specific knowledge into multimodal analysis. HKT employs modality-specific Transformer encoders combined with cross-modal attention layers for capturing the intricate interaction among language, visual, and acoustic signals. HKT also combines semantic ambiguity and sentiment features derived through external knowledge sources such as ConceptNet and the NRC-VAD lexicon, which enables the model to capture the complexity and subtlety of multimodal humorous utterances more effectively [4].

3.2 Data Filtering in Multimodal Systems

Although multimodal technology has made great progress, relatively few works address the impact of noisy or low-quality inputs. In practice, humor datasets often include clips with partial silent audio, static frames, or missing facial features that they will degrade the model's performance. Voice Activity Detection (VAD) has been widely used for speech detection in noisy environments [7]. Face detection methods such as Haar Cascades [11] are crucial in visual signal validation. Prior research [3] has also shown that low edge density and color variance can effectively identify static slide-like content to make a way to do the filtering and discard.

Building on these insights, our work integrates a multi-stage filtering pipeline to ensure high-quality input clips for humor detection.

3.3 Contextual Aware Classification Models

Classification is one of the most researched topic in machine learning. While it has been researched thoroughly through the years as it is also one of the earliest task, how to take in contextual information both in sense of text and in sense of time-series data are still some hot topics today. However, in this paper we utilizes a few proven methods to construct our classification model to take in context information. One of which is LSTM that utilizes a unique architecture using cell states and gates to control the flow of information overtime, thus capturing the contextual and time-series information [6]. Another popular architecture is transformers, with self-attention mechanism and positional encoding it can capture contextual and time series data as well [10].

3.4 Fusion Techniques

During the multimodel learning process, fusion plays an important role to combine the information from different modalities. Flexible use of various fusion strategies will significantly improve the model's performance. Some classic fusions include early fusion and late fusion. Early fusion concatenates raw features from each modality [2], while late fusion combines individual classifiers' outputs [8]. However, such naive fusion methods often fail to model complex interactions across modalities.

Then we consider to use Cross-model attention, such as, Tensor Fusion Network (TFN) [15] to improve interaction modeling, but remain limited in adapting to humor variety.

Finally, we extend the Cross-Model Fusion Network (CFN) [16] by introducing bilinear interactions and a Deep Mixture of Experts (MoE), which allows the model to switch attention to the modality that has more contribution to detect the humor. Our CFN+ leverages ensemble averaging to improve robustness.

4 Dataset and Feature Extraction

4.1 Data Description and Feature Extraction

UR-FUNNY2 dataset is a collection of TED talk clips, thus three modalities are present in it: Text, Audio and Video. The authors of the dataset has provided with pre-extracted features for all three modalities, a text transcription of the videos alongside the raw video clips. However, upon inspecting the dataset manually, we discovered a few problems with it. First for video modality, some of them are very low quality. Due to the dataset being procured from TED talks, some portion of the raw videos have the camera focused on the presentation slides instead of the speaker, making the video lack any distinguishable facial feature for humor recognition. Second for audio modality, some of the audios are meaningless. They either are too short or too long and composed of random noises without any real meaning. These two problems are solved by filtering the dataset based on the information contained within them. For audio modality we filtered out the meaningless ones by checking their text transcription to see if the string contain any meaningful sentence, and for video modality we ran them through a model to see if there are faces present. The first filter give us about 6000 samples and the second filter left us with around 1500 samples from the original around 10000 samples included by the UR-FUNNY2 dataset.

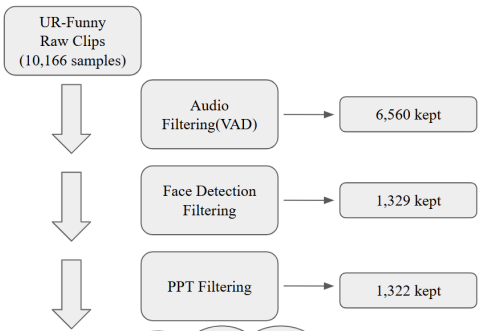

Figure 2: Data Filtering Pipeline

Another problem with the dataset is the pre-extracted features. The models used were considered SOTA during time when the dataset is published, which are considered outdated by current standards. Instead we extracted our own features utilizing the methods we have available today. For text, the transcriptions are accurate enough so we used the transcriptions and use ModernBERT to extract separate features for full sentence as punchline only (as UR-FUNNY2 split the sentence into context and punchline, where punchline is defined as the last sentence in the clip). For audio modality, we used Wav2Vec2.0 to extract two sets of representations. One is the features recognized by the model from the audio before feeding them into the transformer layers, and one is the last hidden states from the transformer layers. For video modality, we used ViT to process the raw videos and extracted their last hidden states as representations.

4.2 Data Filtering

In this section, we will explain in detail how we perform data filtering. To improve the quality of the UR-FUNNY dataset and reduce noise in training, we decide to apply a data filtering pipeline before final model training. You can see the whole filtering pipeline in Figure 2. This was important because,as we mentioned before, the raw UR-FUNNY dataset contains a large number of low-quality clips, including meanless noise, static slides, or frames without faces. These can introduce noise and lead to poor multimodal alignment during feature extraction.

4.2.1 Audio Speech Filtering. We use the method, Voice Activity Detection, to identify and keep clips where clear and active speech is present in at least 85% of the duration of the clips. This ensures that the clips have meaningful audio cues for fusion. Clips with low speech activity are discarded, as they lack useful content for humor recognition while we do audio feature extraction.

4.2.2 Face Detection Filtering. For the visual modality, we decide to sample one frame per second from each video and used OpenCV's Haar Cascade classifier to detect human faces. Clips were kept only if faces were detected in at least 60% of sampled frames. After this filtering, we successfully eliminate videos where the speaker was

not visible or where the camera was stuck on non-human content, such as slides or the audience.

4.2.3 PPT Frame Filtering. There is still a problem that some clips still had faces but they were frozen on slides or cause of video freezing. To further clean the dataset, we design a PPT Frame Filtering to remove static, slide-like clips. In detail, we analyze each sampled frame for edge density and color variance. If more than 30% frames in a video had low edge and color variation, the clip is discarded.

After applying these three stages, the dataset was reduced from over 10,000 raw clips to 1,322 high-quality samples to help our model focus on meaningful multimodal interactions (most content of a clips are meanful and can be used for humor detection) for humor detection.

5 Method

5.1 Unimodal

For Unimodal models, we first tested a preliminary classification method on the different splits of the datasets and then proceeded with other testings. The first model we tested is a BiLSTM classification model that takes all the features extracted with methods mentioned in the previous section. BiLSTM is used in order to take in more context information from all the data as they are all sequential [6].

To test different data splits, we utilized the same model for each modality and tested their performance to see which one performs better. There are three different splits as mentioned in the dataset section: raw, audio-filtered and fully-filtered. For text modality we have 2 sub categories: full sentence and punchline only. Similar categorization happens with audio modality as well, as we have the features and the last hidden states.

For text, The sentence embeddings were extracted using ModernBERT [12], with embedding size 1024. Across all versions of the dataset, the full-sentence setup consistently performed better than the punchline-only features. On the audio-filtered dataset, punchline-only text achieved 64.04% accuracy and 64.10% AUC, while full-sentence text achieved 71.80% accuracy and 76.34% AUC. This result shows that including the context along with the punchline is important for the model to learn valuable cues for humor detection, which makes sense because many jokes rely heavily on buildup and contrast [5].

For audio, we extracted two sets of features: a basic pre-transformer feature vector, and a deeper representation using the last hidden states from Wav2Vec2.0 [1], which gave us a hidden dimension of 512 and 768 respectively. On the unfiltered dataset, the Wav2Vec2.0 hidden states gave the best audio-only results, reaching 71.05% accuracy and 71.08% AUC. The metrics stayed consistent even across the filtered dataset versions. This shows that Wav2Vec2.0 performed well in modeling speech-based humor, possibly because it captures subtle timing, tone, and prosodic variations that signal sarcasm or punchline delivery [4].

For visual, we used ViT-based features extracted from punchline video segments, the feature dimensions was 768 [17]. Out of all three modalities, visual performed the worst. On the unfiltered dataset, visual features only reached 58.05% accuracy and 61.47% AUC. Performance slightly dropped on the audio-filtered dataset and improved a bit on the fully filtered version, but it never came close to text or audio. This shows that it was harder to model visual humor because it is hard to detect without surrounding context or motion [2]. Moreover, many of the visual features in the initial dataest was just still images and did not provide a lot of information, they contained a lot of noise which made it harder to create suitable features for classification. Table 1,2,3 shows model performance with respect to the type of dataset used.

Table 1: Unimodal Performance: Text Only

Dataset	Punchline Only		Full Sentence	
	ACC	AUC	ACC	AUC
Unfiltered	65.44%	65.44%	66.85%	66.83%
Audio Filtered	64.04%	64.10%	71.80%	76.34%
Fully Filtered	60.65%	61.64%	71.36%	75.60%

Table 2: Unimodal Performance: Audio Only

Dataset	Audio Features		Last Hidden States	
	ACC	AUC	ACC	AUC
Unfiltered	67.97%	68.20%	71.05%	71.08%
Audio Filtered	67.32%	66.87%	71.05%	71.12%
Fully Filtered	65.16%	65.82%	68.92%	69.68%

Table 3: Unimodal Performance: Visual Only

Dataset	ACC	AUC
Unfiltered	58.05%	61.47%
Audio Filtered	51.24%	50.94%
Fully Filtered	57.29%	58.01%

5.2 Bimodal fusion

After evaluating Unimodal models, we moved on to testing Bimodal pairs. For this, we used two different types of fusion architectures, Early Fusion and Late Fusion [2].

In the Early Fusion model [2, 15], we concatenated the features from two modalities and passed them through a small MLP with the architecture Linear(128 neurons), ReLU, Linear(2 neurons) and finally softmax to predict the class. This allowed the model to learn interactions between two modalities from the start. The same architecture was used on all combinations (Text + Audio, Audio + Visual, Visual + Text) and the concatenated features size was 128 (64 from each modality) and the output was binary classification.

In the Late Fusion model [8], we first passed we first passed each modality separately through its own projection MLP to get a fixed-size embedding (64 dimensions). Then we concatenated those two projected embeddings and fed them to a final linear layer for classification. This gave the model a chance to learn representations for each modality independently before mixing them.

The best performing model was the combination of Text + Audio (using early fusion), it had an accuracy of 73.87% and AUC of 0.765 [13]. This shows that text and speech cues compliment each other well for detecting humor. It even beats the unimodality score of text only which was the best performing modality for this task. The late fusion version of this pairing still performed decently (accuracy

70.85%, AUC 0.755), but slightly worse than early fusion, suggesting that shared interaction modeling helps when the modalities are semantically aligned [9].

For Text + Visual, the results were relatively balanced across fusion types, with late fusion slightly outperforming early fusion (68.84% vs. 66.83% accuracy, both with 0.761 AUC). This showed that projecting each modality separately before combining may help when working with less semantically aligned inputs [16]. Audio + Visual was the weakest modality pair overall, with early fusion achieving only 55.78% accuracy (AUC 0.580), and late fusion performing slightly worse at 54.65% (AUC 0.572), likely due to the lack of semantic content and weak alignment between the two modalities [15].

Table 4: Bimodal Fusion Results

Modality Pair	Fusion	Accuracy (%)	AUC
Audio + Text	Early	**73.87**	0.765
Audio + Text	Late	70.85	0.755
Text + Visual	Early	66.83	0.761
Text + Visual	Late	**68.84**	0.761
Visual + Audio	Early	**55.78**	0.580
Visual + Audio	Late	54.65	0.572

5.3 Trimodal Fusion

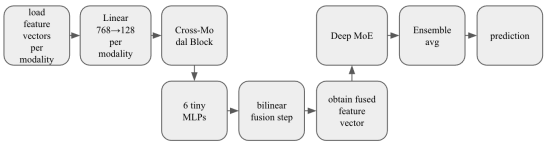

Figure 3: Pipeline of Cross-Modal MoE++ model

For trimodal fusion, to better capture humor from different modalities, our group decide to use the CrossModal Fusion Network (CFN) [14] as the foundation for our trimodal fusion approach. The standard CFN design can be summarized down to three steps: loading the extracted feature vectors of each modality (text, audio, visual), feeding them through a Cross-Modal Block to perform cross-modal fusion, which outputs one single vector for prediction. While this architecture offers a general multimodal framework, we identified multiple limitations when we are trying to apply it to adapt humor detection specifically. CFN uses a single-attention mechanism, then the additive residual interactions in the Cross-Modal block for this figure may overlook multiplicative relations, and finally there are no robustness mechanism to handle the variability of humor expressions.

To solve the limitations we just mentioned, we proposed a improved version of the CFN called Cross-Modal MoE++, which implemented the following changes:

First, we preprocessed the extracted feature vectors for each modality from a 768 dimensional vector to a 128 dimensional vector, as we want to bring all the feature vectors into a shared latent space:

$$\hat{t} = W_t t, \quad \hat{a} = W_a a, \quad \hat{v} = W_v v \quad \text{where } W_t, W_a, W_v \in \mathbb{R}^{128\times 768}$$

This projection step reduces dimensionality and encourage interaction across modalities with a unified feature size.

Next we have replaced the original CFN's attention-residual block for a pair-wise bilinear block. Specifically, six small multilayer perceptrons (MLPs) are used to learn view-specific transformations for every modality, then there is a pairwise bilinear step to capture the multiplicative relations. Here we actually compute 3 element-wise (Hadamard) products, which corresponds to the text-audio, text-visual, and audio-visual relationships.

$$b_{ta} = \hat{t}_{ta} \odot \hat{a}_{ta}, \quad b_{tv} = \hat{t}_{tv} \odot \hat{v}_{tv}, \quad b_{av} = \hat{a}_{av} \odot \hat{v}_{av}$$

After getting the element-wise products, we use them to concatenate with the raw modality feature vectors, to finally get the fused feature vector for Deep Mixture of Experts.

$$h_{\text{fusion}} = [\hat{t}; \hat{a}; \hat{v}; b_{ta}; b_{tv}; b_{av}] \in \mathbb{R}^{768}$$

Deep MoE is the next approach/ improvement we add on top of the standard CFN because it allow the model to specialized subnetworks focus on different humor flavors, without forcing a single classifier to cover everything. For our Cross-Modal MoE++, our MoE layer have six experts, each expert is a tiny 2-layer MLP instead of a single linear layer you see in the classic MoE approach, that extra depth in each expert allows the MoE layer to model richer patterns. The final logit from this Deep MoE is the gate's soft-max weighted sum of the six expert outputs.

Last but not least, we train the entire model five times with different random seeds and simply average their logits at test time. This ensemble averaging approach cuts variance and stabilize our overall accuracy.

Even though the initial projection step uses a simple linear layer, which makes the overall trimodal fusion model looks like it's performing linearly, the overall architecture is highly non-linear due to the use of ReLU activations in the pairwise interaction subnets in the bilinear fusiopn layer, and the deeper MLPs in the MoE experts. So our model still ensures that it has sufficient computation complexity to learn complex, non-linear humor patterns across multiple modalities.

5.3.1 ***Trimodal Fusion's Preliminary Results.***

After training the Cross-Modal MoE++ model on our filtered UR-FUNNY dataset using the train-validation-test split, we got a test accuracy of 74.37% and a ROC AUC score of 77.37%, as well as a ROC-AUC curve shown in Figure 4. The result of the trimodal fusion model outperformed all unimodal and bimodal fusion models, which shows that the model have an advantage in detecting humor by capturing humor across multiple modalities. Moreover, the validation ROC AUC curve in Figure 4 showed consistent improvement over the training epochs, which suggests that the model was not only learning consistently, but also converging well.

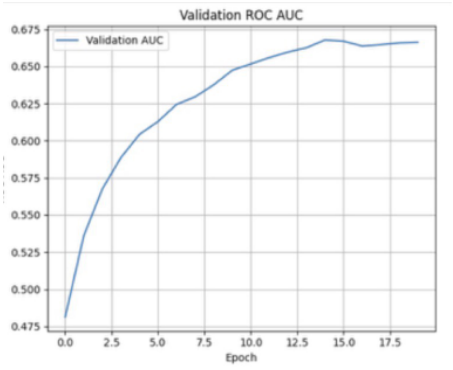

Figure 4: ROC AUC curve of the Cross-Modal MoE++ model

6 Results

We observed that the overall performance steadily improved as we moved from unimodal to bimodal, and finally to our proposed trimodal architecture. This trend clearly shows that combining modalities and modeling their interactions leads to more accurate humor detection.

Among the unimodal models, the full-sentence text setup gave the best results, achieving 71.80% accuracy and 76.34% AUC on the audio-filtered dataset. In the case of bimodal fusion, the combination of audio and text using early fusion gave the strongest results, reaching 73.87% accuracy and 0.765 AUC. This shows that linguistic and acoustic features complement each other well when fused directly. Finally, our best result came from the trimodal fusion using the Cross-Modal MoE++ model, which reached **74.37%** accuracy and **77.37%** AUC. This architecture outperformed all previous configurations, confirming that jointly modeling all three modalities with attention to multiplicative interactions and expert specialization provides a more complete understanding of humor across different forms of expression.

7 Future Plans

For the trimodal fusion, the current result is definitely not the end for our approach, there are still a lot of different implementations that can be added to achieve better results. For instance, we could implement the bilinear fusion more thoughtfully to get a better fused feature vector through calculating different products to capture the multiplicative relations. Moreover, for the Deep MoE layer, we could try for different expert structures to distinguish different humor varieties in more detail. Last but not least, for the overall model, we could finetune a lot of hyperparameter, as well as using a larger dataset for a more stable results.

8 Team contribution

Peilin Li: Designed and implemented the Cross-Modal MoE++ model. Explored on the visual single modality fusion approaches.
Jiaqi Lu: Feature extraction, test extracted features, unimodal classification design.
Yihang Yin: Data filtering pipeline design and implementation. Contributed to the design of the trimodal fusion strategy, including investigating Cross-Modal Fusion limitations,modifying the cross-modal block. Also explored the feasibility of using Deep Mixture of Experts.
Jayavibhav Niranjan Kogundi: Explored, designed and implemented various architectures for Unimodal and Bimodal classification and compared results.

References

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. arXiv:2006.11477 [cs.CL] https://arxiv.org/abs/2006.11477
[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2019), 423–443. doi:10.1109/TPAMI.2018.2798607
[3] Shumeet Baluja and Michele Covell. 2006. Content Detection for Video Summarization using Edge Density and Motion Analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1577–1580. doi:10.1109/ICME.2006.262802
[4] M. Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor Knowledge Enriched Transformer for Understanding Multimodal Humor. In *AAAI Conference on Artificial Intelligence*. https://api.semanticscholar.org/CorpusID:232181726
[5] Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed (Ehsan) Hoque. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics. doi:10.18653/v1/d19-1211
[6] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[7] Google Inc. 2011. WebRTC Voice Activity Detection. https://webrtc.org/. Accessed: 2024-05-05.
[8] Babak Nojavanasghari, Charles Hughes, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction (ICMI)*. ACM, 284–288. doi:10.1145/2993148.2993191
[9] Shraman Pramanick, Dimitar Sharma, Sayan Ranu, and Maunendra Sankar Desarkar. 2022. Multimodal Learning Using Optimal Transport for Sarcasm and Humor Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2740–2749. doi:10.1109/WACV51458.2022.00279
[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. arXiv:1706.03762 [cs.CL] https://arxiv.org/abs/1706.03762
[11] Paul Viola and Michael Jones. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* (2001), 511–518.
[12] Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. 2024. Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference. arXiv:2412.13663 [cs.CL] https://arxiv.org/abs/2412.13663
[13] Haojie Xu, Weifeng Liu, Jingwei Liu, Mingzheng Li, Yu Feng, Yasi Peng, Yunwei Shi, Xiao Sun, and Meng Wang. 2022. Hybrid Multimodal Fusion for Humor Detection. In *Proceedings of the 4th International on Multimodal Sentiment Analysis Workshop and Challenge*. ACM. doi:10.1145/3551876.3554825
[14] Yitian Yuan, Hang Xu, Xiaojun Chang, and Alexander G. Hauptmann. 2021. Cross-Modal Attention with Semantic Consistency for Multimodal Emotion Recognition. https://arxiv.org/abs/2111.02172. arXiv:2111.02172 [cs.CV].
[15] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 1103–1114. doi:10.18653/v1/D17-1115
[16] Tianlin Zhao, Liunian Harold Cui, Yuwei Zhang, Jie Lei, Xinyue Wang, Zhou Yu Wang, and Zhou Yu. 2020. CFN: Cross-modal Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 6362–6372. doi:10.18653/v1/2020.acl-main.568
[17] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. 2022. General Facial Representation Learning in a Visual-Linguistic Manner. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18697–18709. doi:10.1109/CVPR52688.2022.01817