Improving Multimodal Sentiment Analysis via Auxiliary Heads & Orthogonality of Modal Representations

Sai Kiran Jagini, Srihith Bharadwaj Burra and **Yasaswini Tiramdas**
Team: Tangled
{sjagini, ytiramdas, sburra}@cs.stonybrook.edu

1 Introduction

Multimodal sentiment analysis, which identifies emotional tones from channels like text, voice, and visuals, is crucial for nuanced human-computer interaction and diverse applications such as feedback analysis and mental well-being monitoring. This project explores Transformer-based representation learning and fusion strategies for this task using the CMU-MOSEI dataset, aiming for a holistic understanding of sentiment by leveraging the unique strengths of each data type.

Our work addresses the core challenge of effectively fusing heterogeneous multimodal information. We investigate sentiment analysis and multimodal learning through Transformer architectures, focusing on learning discriminative and complementary representations from textual, acoustic, and visual data. To this end, we implement and evaluate models incorporating specific techniques such as cross-modal attention, orthogonality constraints between unimodal features, and auxiliary learning via modality-specific prediction heads to enhance the overall predictive performance.

2 Background

This research leverages the Transformer architecture (Vaswani et al., 2017), whose self-attention mechanism has proven effective for sequence modeling across individual modalities like text, audio, and video. Effective multimodal integration often moves beyond simple fusion methods (e.g., early or late concatenation), employing more dynamic, attention-based mechanisms to allow modalities to interact and combine information contextually (Baltrušaitis et al., 2019). Our work focuses on sentiment analysis—classifying emotional polarity on a multi-point scale—within this multimodal learning framework.

To enhance multimodal sentiment classification, we build upon several established concepts. Cross-modal attention allows distinct modalities to query and draw relevant context from one another, leading to more informed representations. We also explore regularization through orthogonality constraints, which aim to encourage learned unimodal representations to be less redundant and more specialized (Bengio et al., 2013). Finally, auxiliary learning, wherein the model is trained on supplementary prediction tasks (in our case, unimodal sentiment prediction), can guide feature learning in the shared layers and often improves primary task performance (Ruder, 2017).

3 Data

For this project, we utilized the CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) dataset (Zadeh et al., 2018), a large and widely used benchmark for multimodal sentiment analysis and emotion recognition. MOSEI consists of monologue video segments collected from YouTube, featuring a diverse range of speakers and topics. This diversity makes it a challenging yet realistic dataset for studying multimodal language.

The dataset provides aligned features for three modalities:

- **Language (Text):** Transcribed speech, typically represented by word embeddings. For our experiments, BERT text features are provided with a dimension of 768.
- **Acoustic (Audio):** Features extracted from the speech signal, such as COVAREP features (Degottex et al., 2014), capturing aspects like pitch, intensity, and voice quality. Our audio features have a dimension of 74.
- **Visual (Vision):** Features extracted from video frames, often focusing on facial expressions and head movements using tools like FACET. Our visual features have a dimension of 35.

Model Card – AuxOrtho Multimodal Sentiment Classifier

Model Details

- **Architecture:** Four late-fusion Transformer variants with unimodal encoders (text, audio, vision), cross-modal attention, and auxiliary losses.
- **Inputs:** Word-aligned features – BERT (768-d), COVAREP (74-d), FACET (35-d) over 50 timesteps.
- **Outputs:** 3-class sentiment prediction – negative, neutral, positive.

Intended Use

- Predict sentiment in single-speaker YouTube monologues.
- Applicable in downstream affective computing and social media monitoring tasks.

Training Data

- CMU-MOSEI dataset: 23,000 labeled video segments from 2,200 speakers.
- Standard train/valid/test splits used.

Evaluation

- Metrics: CrossEntropy Loss, Accuracy and macro-F1.
- Baselines provide lower bounds; improved variants generalize better via auxiliary supervision.

Limitations & Assumptions

- Assumes a fixed 50-step window captures full sentiment expression.
- Sentiment labels are coarse; may miss subtle effect.
- Data is limited to English-speaking YouTubers; domain shift is likely in other contexts.

Ethical Considerations

Trained only on publicly available YouTube content. No personally identifiable information (PII) used. May inherit and amplify biases from online discourse.

Quantitative Analyses

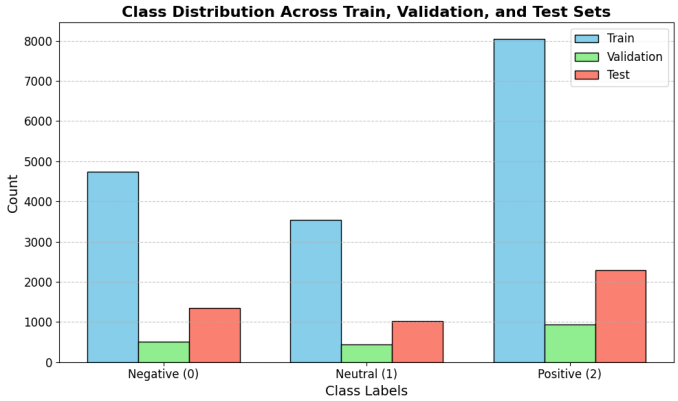

Figure: Class distribution across train, validation, and test sets.

All raw MOSEI files (text, audio, vision, and labels) were first ingested via the CMU Multimodal SDK. High-level feature streams (BERT text, COVAREP audio, FACET visual) are first aligned at the word level and any missing modality values are imputed so that every word has a feature vector for each modality. Finally, the fully aligned data are exported as fixed-length tensors (50 timesteps per modality) for the standard train/validation/test folds. This pipeline ensures synchronicity across modalities and produces clean, uniformly shaped input arrays. Sentiment labels are annotated on a 3-point scale (0 = negative, 1 = neutral, 2 = positive).

Table 1 summarizes the split sizes and basic label statistics.

4 Methods

We implement and compare four distinct Transformer-based models for 3-class sentiment analysis. All models share common hyperparam-

Split	# Samples	Label Mean	Label Std. Dev.
Train	16,326	1.203	0.861
Validation	1,871	1.228	0.847
Test	4,659	1.200	0.860

Table 1: CMU-MOSEI dataset statistics for 3-class sentiment (labels mapped 0-2).

Class Counts	Negative	Neutral	Positive
Train	4738	3540	8048
Validation	506	433	932
Test	1025	1350	2284

Table 2: Class Counts in Train, Val and Test Splits

eters where applicable: a hidden dimension (H) of 128, 4 attention heads for Transformer layers, and a dropout rate of 0.1. The unimodal 'ModalityTransformer' (used by all models) has 2 layers, projecting input features to H, adding positional encoding, passing through a Transformer encoder, and then mean-pooling the sequence to get a fixed-size vector. Figure 1 depicts the Improved 2 ("Auxiliary Heads") architecture.

4.1 Baseline 1: LateFusionTransformer

Baseline 1 first encodes text, audio, and vision independently into $t, a, v \in \mathbb{R}^H$, prepends learnable modality tokens, stacks them, and passes them through a 2-layer fusion Transformer. Its mean-pooled output is fed to a final linear classifier.

4.2 Baseline 2: LateFusionWithCrossModal

Baseline 2 also produces t, a, v separately, then enhances each via a CrossModalAttentionBlock: e.g. t attends to $\{a, v\}$ to yield t_2, and analogously for a_2, v_2. The concatenated $[t_2 \| \ a_2 \| \ v_2]$ vector goes to the classifier.

Hyperparameter	Value
Hidden dim. (H)	128
Attention heads	4
Unimodal Transformer layers	2
Fusion Transformer layers	2 (B1)
Dropout rate	0.1
Learning rate	2×10^{-5}
Batch size	64
λ_{ortho}	1.0
λ_{aux}	0.05

Table 3: Key hyperparameters across models.

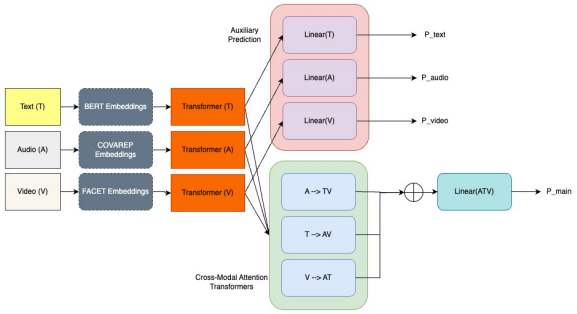

Figure 1: Conceptual diagram of LateFusionWithCrossModalAuxiliaryHeads (Improved 2). Each modality is encoded, cross-attended, and fused before classification.

4.3 Improved Model 1: LateFusionWithCrossModalOrtho

This variant uses the same architecture as Baseline 2 but, during training, returns both the final fused logits and the unimodal embeddings $t, a, v \in \mathbb{R}^H$ from the 'ModalityTransformer's. We apply an orthogonality loss on the L2-normalized embeddings:

$$\mathcal{L}_{\text{ortho}} = \mathbb{E}[\cos^2 \theta_{TA}] + \mathbb{E}[\cos^2 \theta_{AV}] + \mathbb{E}[\cos^2 \theta_{TV}],$$

where θ_{XY} is the angle between embeddings X and Y. The total training objective is

$$\mathcal{L} = \mathcal{L}_{\text{main}} \ + \ \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}}.$$

This encourages each pair of modality encoders to learn decorrelated features.

4.4 Improved Model 2: LateFusionWithCrossModalAuxiliaryHeads

Building on Baseline 2, we attach three small auxiliary heads—each a two-layer MLP (Linear→ReLU→Dropout→Linear)—directly to t, a, and v. In addition to the main loss $\mathcal{L}_{\text{main}}$ on the fused vector, we compute three auxiliary cross-entropy losses $\mathcal{L}^t_{\text{aux}}$, $\mathcal{L}^a_{\text{aux}}$, and $\mathcal{L}^v_{\text{aux}}$. The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{main}} \ + \ \lambda_{\text{aux}} \big(\mathcal{L}^t_{\text{aux}} + \mathcal{L}^a_{\text{aux}} + \mathcal{L}^v_{\text{aux}} \big).$$

This setup encourages each unimodal encoder to learn features that are individually predictive of sentiment.

All models were trained with Adam (learning rate 1×10^{-4}), batch size 64, and CrossEntropy Loss for 40 epochs on NVIDIA A6000 GPU. Key hyperparameters are in Table 3.

5 Evaluation/Results

We evaluated all four models on the CMU-MOSEI 3-class test set, reporting accuracy and macro-F1. Table 4 summarizes the overall performance. A random guess baseline would be 33.3% accuracy; all of our architectures substantially exceed this, demonstrating effective multimodal sentiment learning.

Model	Test Accuracy	F1 Score
Baseline 1	66.6023%	0.635
Baseline 2	66.8598%	0.646
Improved 1	67.7828%	0.662
Improved 2	67.8472%	0.663

Table 4: Three-class sentiment classification performance on the CMU-MOSEI test set.

Introducing cross-modal attention (Baseline 2) yields a +0.5 pp boost over simple late fusion (Baseline 1). Adding the orthogonality loss (Improved 1) and the auxiliary heads (Improved 2) brings further gains. Improved 2 slightly outperforming Improved 1, suggesting that guiding unimodal encoders toward individually predictive features is especially effective.

Figures 2 and 3 show the per-epoch training and validation accuracies for the four models. Both the figures indicate good convergence with minimal overfitting.

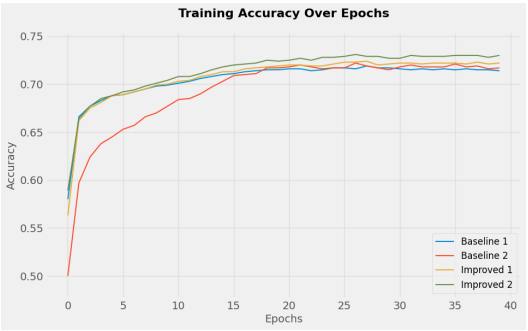

Figure 2: Training accuracy over 40 epochs for all 4 models.

Finally, Figure 4 presents the four models' test-set confusion matrices. We observe that most errors occur between neutral and positive classes, and that both Improved 1 and Improved 2 reduce this confusion compared to the baselines—highlighting their ability to disentangle sentiment nuances.

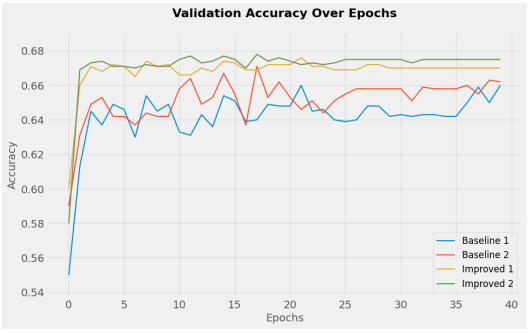

Figure 3: Validation accuracy over 40 epochs for all 4 models.

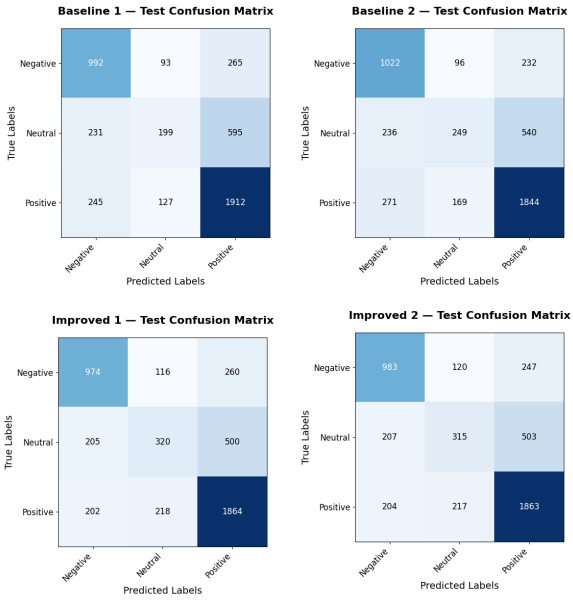

Figure 4: Test confusion matrices for Baseline 1 (top-left), Baseline 2 (top-right), Improved 1 (bottom-left), and Improved 2 (bottom-right).

6 Conclusion

Our findings indicate that incorporating cross-modal attention between unimodal representations provides performance benefits over simpler late fusion. Furthermore, we demonstrated that techniques such as enforcing orthogonality constraints between unimodal features and adding auxiliary modality-specific prediction heads can offer slight but positive improvements to a strong cross-modal baseline. Achieving around 67% 3-class accuracy, these models establish a solid foundation for understanding multimodal interactions, with future work potentially focusing on more sophisticated fusion mechanisms and handling of variable sequence lengths.

References

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(1):49–67. Date of Publication: 07 February 2018. Date of Current Version: 19 December 2018. Issue Date: January 2019. Using key as requested by user, though print year is 2019, online publication was 2018.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep — a collaborative voice analysis repository for speech technologies. pages 960–964.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Pfranger, Esha Vij, Rohit Nevatia, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.