A Late Fusion Approach for Multimodal Image-Text Data

Shreyes Balaji

***Abstract*—As the number of users on social media platforms continues to rise, an increasing number of people are expressing their emotions and opinions. Before the popularity of multimodal machine learning, previous works included identifying texts and images separately to determine sentiment. However, these studies often overlooked features that other modalities could capture that are crucial to emotion. Recent advances in multimodal machine learning are now making essential and precise judgments in image recognition [1]. In this project, we explore a specific type of fusion called late fusion, demonstrating its ability to combine different modalities. In particular, we highlight two late fusion methods that combine the predictions of the image and text models. Additionally, we investigate a contrastive learning approach, which will help improve the visual embeddings, making them more discriminative in feature space. We compare the strengths and limitations of late fusion against other fusion approaches. Our findings are based on experiments conducted with the publicly available MVSA-Single dataset.**

I. Introduction

Social media platforms such as Instagram and Twitter have become the primary sources for users to express their sentiments. Consequently, with the increasing use of smartphones, people are more likely to post multimodal data, such as a tweet with a corresponding image to express their emotions. Accurately detecting sentiment is critical for various applications, including market analysis, public opinion monitoring, and mental health assessment.

Single-modality sentiment analyses typically use Convolutional Neural Networks (CNNs) for image feature extraction. For instance, Xu et al. [2] utilized pre-trained CNNs for object recognition and adapted them for sentiment analysis. However, focusing solely on text or images can miss crucial contextual information provided by the other modality. These methods often fail to capture the full spectrum of human emotions in social media posts. For example, if the text contains a negative, sarcastic tone that is not noticeable, and the image depicts a person smiling, focusing on the text loses critical information from the image.

When discussing multimodal sentiment analysis, the earliest image-text models were feature-based, which included manually selecting attributes to make predictions. Borth et al. [3] proposed a feature selection model using SentiBank to extract 1200 adjective-noun pairs (ANP) as mid-level features in an image and employed SentiStrength to capture the sentiment strength of text tweets. The results of the image and text were combined. Later, deep learning methods showed better results. Cai et al. [4] introduced a method where CNNs are trained to extract textual and visual features, which are then combined and passed into another CNN. Xu et al. [5] proposed a method where text, object, and scene features represent the multimodal tweet. They developed MultiSentiNet, a visual-guided attention mechanism to extract sentiment-relevant words and integrate them with the object and scene features. Yan et al. [6] proposed a Multi-view Attention Network (MVAN), which uses memory networks to highlight the interaction between different modalities. To jointly learn representations between visual and textual content, the Co-Mem network [7] iteratively models and dynamically updates the image and text interactions for multimodal sentiment analysis. The Multi-channel Graph Neural Network (MGNNS) [8] leverages graph neural networks to learn multimodal representations of the global features in the dataset. A scene and object ResNet is used to encode image representations, with the final fusion of the image features and text features achieved through a multi-head interaction mechanism. Due to the success of contrastive learning on downstream tasks, Li et al. [9] proposed CLMLF, which uses supervised contrastive learning to align and fuse token-level features, enhancing the understanding of standard features related to multimodal sentiment analysis. Wang et al. [10] introduced the MLFC module, which uses a CNN-connected Transformer fusion. The authors also introduce SCSupConLoss, which combines cross-entropy loss and supervised contrastive loss.

In most deep learning methods, data fusion is used to combine the modalities in a certain way to enhance understanding of the sentiment. Data fusion involves integrating information from multiple sources to create a unified and accurate system representation. Different multimodal data methods exist in early, intermediate, and late fusion. Early fusion involves combining the different modalities before the feature extraction process. Intermediate fusion involves extracting features from each modality and fusing these marginal representations later inside the network. In late fusion, the predictions from each modality are combined to produce a single final prediction [11].

In this project, we highlight:

- A late fusion method that weights the output probabilities of the sub-models. The sub-model prone to more errors than the other sub-model contributed less to the final prediction. The final multimodal model could more accurately balance the confidence scores of both sub-models. We compare the effectiveness and simplicity of this approach with other fusion methods.

- A supervised contrastive learning approach to improve

the generalizability of our Image model. Contrastive learning encourages feature vectors of the same/'positive' class to be close together in latent space, pushing the dissimilar/'negative' class feature vectors farther apart.

II. Dataset

The dataset used in this study was the MVSA Single dataset, published by the Multimedia Communications Research Laboratory [11]. The dataset contains 4869 samples and is collected from Twitter with each text-image pair comparison labeled by a single sentiment. We process the dataset in a way that all contradicting sentiments are removed. For example, if the text sentiment is positive but the image sentiment is negative, the sample is removed. However, if one label is positive (or negative) and the other is neutral, we take the sentiment of this pair to be positive (or negative). The resulting dataset contains 4511 samples.

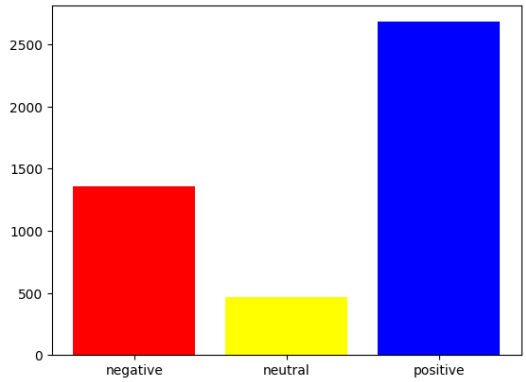

Fig. 1: Distribution of sentiments in the MVSA-Single dataset

III. Implementation Details

In our experiments, the data is split into training, validation, and testing sets randomly into the ratio 8:1:1. For the VGG16 Image Model (5.1.1), the learning rate is set to 1e-3 and Adam is used as the optimizer. In the ResNet50 model, the learning rate is set to 2e-5 and the AdamW optimizer is used. A weight regularizer is used to apply a penalty to the Dense layer's kernel during optimization. The BERT model uses the same learning rate and optimizer as the ResNet50 model. Due to the limited data, the batch size is varied under 20 across all models. We primarily evaluate model performance using accuracy and F1 score, which are defined as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

TP stands for True Positive, FP stands for False Positive, FN stands for False Negative, TN stands for True Negative. All of the experiments are done on one NVIDIA GPU P-100 in a Kaggle Notebook.

IV. Methodology/Methods

A. Visual Feature Detectors

Sentiments are closely correlated with objects in an image. For instance, a victory parade might evoke positive sentiments, while a natural disaster might evoke negative sentiments. To extract the visual context from each image, every image I in the dataset is resized into a 224×224 size image and converted into RGB color. In both of the image models below, ImageDataGenerator is used to create new training examples by augmenting the original images, which enhances the model's generalization ability.

1) VGG16: To detect visual scene features in an image, we first use a VGG16 model, which is composed of 16 layers—5 convolutional blocks and 3 dense layers for processing information incrementally [13]. The model's weights are pretrained on the Places365 dataset, which consists of over 10 million images across 400 unique scene categories [14], [15]. To evaluate the model's performance comprehensively, we use KFold cross-validation with four splits.

2) ResNet50: The ResNet50, with weights pretrained from ImageNet [16], is used as the primary Image Encoder. Previous studies [8], [9], [10] extract visual features with a ResNet50. The popularity of ResNets is due to its skip connections, which allow for information to flow from earlier layers to later ones, enabling the network to learn complex high-level features without the issue of vanishing gradients. EarlyStopping is applied to monitor the validation loss, and stop the training of the model when the validation loss increases. To consider an improvement in the model's performance after each epoch, the mindelta parameter is set to $1e^{-4}$.

B. Textual Feature Detectors

To understand textual context, we use the RoBERTa-base model for textual feature extraction. Unlike static masking in the BERT model [17], RoBERTa [18] uses dynamic masking as a data augmentation technique, where tokens are masked differently for each training epoch, and there is no fixed number of masked tokens. RoBERTa utilizes a larger Byte-Pair Encoding (BPE) vocabulary containing 50K subunits. BPE iteratively merges the most frequent pair of consecutive bytes until a predefined vocabulary size is reached. We use a RoBERTa-base model, which was trained on 58 million tweets, based on the original RoBERTa-base checkpoint [19]. We fine-tune this model on our dataset of tweets. Each tweet is preprocessed using standard techniques with the Natural Language Toolkit (NLTK), including converting all characters to lowercase and removing nonessential characters that do not affect sentiment polarity. For text enhancement, a data augmentation method called back-translation is used [20].

In back-translation, an input text t_1 of a language E is translated into another language G, and then retranslated back into the language E as shown Figure 2. This creates

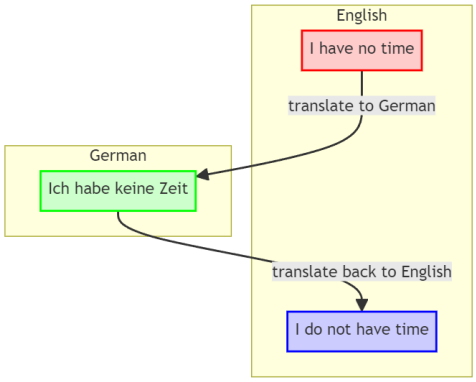

Fig. 2: Backtranslation between English and German

a new output text t_2, which serves as the augmented text for the original input text. Due to limited computing power, we randomly sample 15% of the training dataset for back-translation between English and German. After the tweets in the dataset are tokenized using BPE, they are passed in the model to retrieve the prediction probabilities. To lower the overfitting of the finetuned RoBERTa on the dataset, we employed strategies such as adding a weight decay penalty and using EarlyStopping.

C. Contrastive Learning

Recent advancements in contrastive learning have made it effective across domains like natural language processing and computer vision. Contrastive learning aims to learn relevant features of data and group similar samples together while pushing dissimilar samples away in latent space [21]. Self-supervised contrastive learning methods like SimCLR [22] are popular because of the massive amounts of unlabelled data, so training encoders are relatively accessible for image-representative tasks.

A critical objective in contrastive learning is examining the loss function's ability to capture relevant features from the data [23]. During training, the neural network aims to minimize the loss function to adjust the weights and biases accordingly. In the absence of contrastive learning in the supervised setting, the encoder and classifier parts of the network are trained together. The standard loss function used for multi-class classification is categorical cross-entropy.

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{j}^{N}\sum_{k}^{C} y_{jk}\log(\hat{y_{jk}})$$

y_{jk} represents the true label indicator for a sample j belonging to class k. $\hat{y_{jk}}$ is the probability of sample j belonging to class k, after the softmax function is applied.

However, when contrastive learning is applied in the supervised setting (supervised contrastive learning), the encoder and classifier network are trained separately in two phases. In the first phase, the encoder produces high-level representations of the input using a contrastive loss function to separate the distinct classes in the embedding space. In the second phase, the projection network is removed and linear classification layer is added on top of the frozen encoder network.

We use N-Pair loss as the contrastive loss for learning embeddings [24]:

$$\mathcal{L}(\{x, x^+, \{x_i\}_{i=1}^{N-1}\}; f) = \log\left(1 + \sum_{i=1}^{N-1}\exp(f^T f_i - f^T f^+)\right)$$
$$= -\log\frac{\exp(f^T f^+)}{\exp(f^T f^+) + \sum_{i=1}^{L-1}\exp(f^T f_i)}$$

The feature vectors f are the result of passing the high-level image representations from the encoder into a projection network, which consists of a single linear dense layer of size 64.

D. Late Fusion Network and Sentiment Classification

Late fusion combines the sub-models trained on each modality to form a prediction. Each sub-model learns $P(y|x_i)$, where x_i is the data from ith modality [25]. The techniques of combining the prediction probabilities of the sub-models in late fusion include averaging, majority voting, weighted voting, and meta-classifiers. Yoo et al. [26] found the mean of the predicted probabilities from two sub-models to form the final predictions. Soto et al. [27] proposed a late-average fusion model where soft voting was used to perform the average probability for each class in the different modalities. The resulting model could outperform other late and intermediate fusion strategies on the same data. However, average-based late fusion assumes that each sub-model holds equal weightage in predicting the target variable. If noise affects one modality, the resulting multimodal model would be severely affected.

Meta-learning approaches involve passing the predicted probabilities from the separate sub-models into a classifier. Usually, the classifiers involve an artificial neural network to learn non-linear relations between the modalities [24]. Reda et al. [28] fed the prediction probabilities as inputs into a constraint sparse autoencoders connected to a classifier for the resulting predictions.

In this project, we implement two weighted average mechanisms to capture the prediction strengths of both models. Wang et al. [29] weighted the prediction probabilities of sub-models of lung cancer survival by its uncertainty such that the sub-model with the lower accuracy contributed less to the final prediction. We evaluate two methods of weighted late fusion on our dataset.

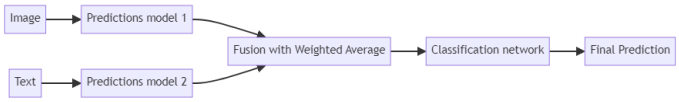

Fig. 3: Rough framework of late fusion

1) Detection Rate Approach: After evaluating the image and text models on the test set, we assign weights to each modality and define the Detection Rate (DR) to be the classifier's ability to identify true positive cases among all predictions [30].

$$DR = \frac{TP}{TP + TN + FN + FP}$$

Give that there are three sentiments, the detection rate involves a one versus all approach, where the sentiment evaluated is considered to be a "positive" finding, and the rest are "negative" findings. The detection rate is calculated for each sentiment across all modalities. We set the weights $W_{ij} = \frac{1-DR_{ij}+\alpha_j}{\sum_j 1-DR_{ij}+\alpha_j}$, where $i \in \{0, 1\}$ and $j \in \{0, 1, 2\}$. W_{ij} represents the weight for each sentiment in a specific modality. We normalize the weights so each sentiment has a somewhat equal contribution to the final prediction. The constant α_j is added to account for the class distribution imbalances. However, its inclusion is optional. To put more emphasis for the minority class, we recommend removing the α_j to set larger weights. In this case, we set α_j to be the ratio of the sentiment evaluated to the total data points. For example, if the negative sentiment of text is analyzed, then α_j is set to the proportion of negative data relative to the dataset size.

After the softmax function is applied to the output for each modality, the resulting prediction probabilities contain three columns corresponding to the three sentiments. The weights of each sentiment are applied to the respective columns of the text and image prediction probabilities. The weighted probabilities for each modality are then aggregated to form the final weighted prediction probability.

2) Minimizing Loss Approach: From each modality, we retrieve the validation accuracy corresponding to the epoch with the least validation loss.

Contrary to the *Detection Rate Approach*, we define two weights W_t and W_i, corresponding to the text and image modalities respectively. The weights are set to the ratio of their validation accuracies to the combined accuracy. We recommend setting the weights as the ratio of each modality's F1 score to the combined F1 score to leverage the strengths of both models in imbalanced datasets. The weighted probabilities are then combined to form a single final weighted probability:

$$P_w = W_t P_t + W_i P_i$$

P_t and P_i are the predicted probabilities after the softmax function is applied.

V. Results and Discussion

The table shows the results of the individual models on the MVSA-Single dataset. Note that these models represent our implementations and may not necessarily reflect the highest accuracy or F1 score possible.

Modality	Model	Accuracy	F1 Score
Image	VGG16	0.588	—
	ResNet50	0.6208	0.6162
Text	RoBERTa	0.7073	0.7113
Multimodal	Detection Rate App.	0.6918	0.6805
	Minimizing Loss App.	**0.7273**	**0.7223**

We can make a couple of observations from the table above. First, the image models perform worse than the text models, which suggest that the images may contain a certain degree of noise, such as irrelevant features, which causes the model to struggle in distinguishing between the high-dimensional features in vector space. However, implementing a contrastive learning approach significantly enhanced the classifier's prediction ability in the minority classes. Compared to the approach without contrastive learning using similar data augmentations with a ResNet50, we observed an approximate 5.3% increase in accuracy and 20% increase in the F1 score.

While we augmented 15% of the training set using back-translation for text features, the optimal augmentation proportion for a dataset remains to be studied in future work. We selected a small proportion to minimize the risk of overfitting, and observed a slight improvement in the text model's accuracy and F1 score.

A. Comparison of the multimodal models

The two multimodal models outperformed the baseline image model's performance. The *Minimizing Loss Approach* (MLA) resulted in a 2.83% increase in the accuracy and 1.55% increase in the F1 score compared to the unimodal RoBERTa text model. We can identify key differences between the two multimodal models:

- In the *Detection Rate Approach*, the emotions are set with a similar weight values, which means that this approach does not adequately account for the imbalance in the dataset. Moreover, because the weights are uniformly distributed across the modalities, this approach tends to favor the modality that shows high confidence on a sentiment for the data. The lower accuracy compared to the RoBERTa text model could be due to some instances of the image modality predicting the wrong class with a huge confidence.
- The *Minimizing Loss Approach* is effective when the dataset is imbalanced. In this approach, each modality is given a different weight value based on the accuracy or F1 score on the test set, while each sentiment within each modality is given the same weight. This strategy is particularly useful as it focuses on the overall performance of each model. Unlike the concern with the *Detection Rate Approach* that the incorrect prediction of the image modality substantially influences the final prediction, the weighting scheme mitigates the false confidence.
  There are some instances where the text model predicts two sentiments with high confidence, while the image model correctly identifies the true sentiment. This situation can occur if the preprocessed text data is "cut short"

and lacks sufficient information for prediction. In such cases, the confidence scores of the image modality become crucial to the final prediction. This observation may explain why Minimizing Loss Approach demonstrates an improvement over the RoBERTa text model.

B. Assessment of Late Fusion and weights

We explore the Minimizing Loss Approach, particularly the effect of setting the weights as the ratio of the model's validation accuracy relative to the total validation accuracy of all modalities. Suppose the image model receives an accuracy of 59.4% but a 44% F1 score, indicating that the model is predicting the majority class, while struggling with the minority class. In other words, the model is "guessing" the sentiment instead of making accurate predictions. Even though predictions from both modalities are combined, the overall impact on the final multimodal model's accuracy and F1 score is minimal. In this case, the metrics for the multimodal model largely reflect those of the stronger text model. We are not learning anything beneficial from the image model. This situation highlights a potential drawback late fusion approaches as it can rely on solely one model. By setting the weights proportional to the F1 scores, it becomes clearer if one model is underperforming. If there is a lack of improvement in the multimodal model's F1 score compared to the stronger model's F1 score, this indicates that the weaker model is not contributing effectively. If both models have strong F1 scores and complement each other, setting the weights based on these scores can lead to improved metrics in the final multimodal. Strong models are crucial for achieving meaningful feature representations and mitigating bias.

C. Case Study

Table 1 is split into three columns. The first column shows an image and corresponding tweet from the dataset, the second column depicts the incorrect prediction from the RoBERTa text model, and the third column contains the final mulitmodal predictinon using *Minimizing Loss Approach*. In the first example, the text is written in a different language, which makes it difficult for the RoBERTa model to identify the type of emotion. However, the image exhibits a clear negative sentiment, and the predictions are combined, the confidence score from the image model leads to the correct prediction. Similarly, in the third example, the textual content does not clearly indicate the emotion due to the presence of the word "Comedy." The image model demonstrates high confidence in identifying the sentiment as negative, thereby contributing to an accurate prediction.

Image and Caption	RoBERTa Prediction	MLA Prediction
 RT @HergunYeniBilg: Ge?en sene sahilde top oynarken ?srail'in vurdu?u 4 ?ocu?un babas?.Sava?a dair ne varsa yznde.	**Neutral**	**Negative**
 ::Watches as the catering staff serve up some light snacks::	**Neutral**	**Positive**
 #DEVASTATED BREAKING: Jon Stewart is retiring from 'The Daily Show,' Comedy Central says	**Neutral**	**Negative**

TABLE I: Examples misclassified by RoBERTa and correctly classified by the Minimizing Loss Approach (MLA)

VI. Conclusion

In this research project, we developed two late fusion methods to capture the prediction strengths of the image and text modalities. Additionally, we implemented a supervised contrastive method that leverages feature similarity and dissimilarity to enhance the model's ability to differentiate between classes effectively. The experimental results on the MVSA-Single dataset show that the multimodal approaches are competitive with other models for this theme.

However, a limitation is that our models were only tested on the MVSA-Single dataset. We plan to test our models on larger datasets such as MVSA-Multiple [11] and CMU-MOSEI [30].

Since computer memory could be a critical barrier, the proposed late fusion approaches are designed to learn information between the modalities but with less memory usage. The memory associated with late fusion methods is isolated to each modality's processing pipeline because features are combined at the final stage. Late fusion minimizes the need for complex integration during the intermediate processing stages. A downside to late fusion is being unable to learn the rich correlations between different modalities thoroughly.

To address this limitation, the authors [9], [10] propose an intermediate fusion approach that uses a transformer encoder to merge the image and text features earlier to understand fine-grained, course details.

Future research should explore hybrid fusion approaches that combine intermediate and late fusion. Our implementation of weighted late fusion shows effectiveness over baseline methods on imbalanced datasets, and combining this approach with intermediate fusion is necessary for the multimodal model to capture the prediction strengths and cross-modality correlations. Furthermore, since a contrastive learning approach improved the predictions of the minority class, a weighted contrastive learning approach can be explored. In this case, two contrastive loss functions can be combined:

$$L_{combined} = \lambda \cdot L_1 + (1 - \lambda) \cdot L_2$$

A weighted combination of losses can also improve the loss function's convergence by reaching a desired minimum. This approach adjusts the encoder's weights so that the model can capture various features and aspects of the data. We will also incorporate other modalities such as audio in future work.

VII. Acknowledgements

I would like to thank my mentor Varsha Sandadi, Inspirit AI, and my family for guidance and support during this project. A special thanks to user vincemarcs on Kaggle for code to load and preprocess the data.

References

[1] Lark Editorial Team. (2023). Multimodal in Machine Learning. Lark. https://www.larksuite.com/en_us/topics/ai-glossary/multimodal-in-machine-learning.

[2] Xu, C., Cetintas, S., Lee, K., & Li, L. (2014). Visual sentiment prediction with deep convolutional neural networks. arXiv preprint arXiv:1411.5731. http://arxiv.org/abs/1411.5731

[3] Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. Proceedings of the ACM International Conference on Multimedia, 223–232.

[4] Cai, G., & Xia, B. (2015). Convolutional neural networks for multimedia sentiment analysis. In Lecture Notes in Computer Science (Vol. 9380, pp. 159–167). https://doi.org/10.1007/978-3-319-25207-0_14

[5] Xu, N., & Mao, W. (2017). Multisentinet: A deep semantic network for multimodal sentiment analysis. Proceedings of the ACM on Conference on Information and Knowledge Management, 2399–2402.

[6] Yang, X., Feng, S., Wang, D., & Zhang, Y. (2020). Image-text multimodal emotion classification via multi-view attentional network. IEEE Transactions on Multimedia, 23, 4014–4026. https://doi.org/10.1109/TMM.2020.3035277

[7] Xu, N., Mao, W., & Chen, G. (2018). A comemory network for multimodal sentiment analysis. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (pp. 929–932).

[8] Yang, X., Feng, S., Zhang, Y., & Wang, D. (2021). Multimodal sentiment detection based on multi-channel graph neural networks. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 328–339

[9] Li, Z., Xu, B., Zhu, C., & Zhao, T. (2022). CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. arXiv. https://doi.org/10.48550/arXiv.2022.2204.05515

[10] Wang, H., Li, X., Ren, Z., Wang, M., & Ma, C. (2023). Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion. Sensors (Basel, Switzerland), 23(5), 2679. https://doi.org/10.3390/s23052679

[11] Boulahia, S., Amamra, A., & Madi, M. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. In ResearchGate. https://www.researchgate.net/publication/354984828_Early_intermediate_and_late_fusion_strategies_for_robust_deep_learning_based_multimodal_action_recognition

[12] Niu, T., Zhu, S., Pang, L., & El Saddik, A. (2016). Sentiment analysis on multi-view social data. In Proceedings of the International Conference on Multimedia Modeling (pp. 15–27). Springer.

[13] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv. https://arxiv.org/abs/1409.1556

[14] Kalliatakis, G. (2017). Keras-VGG16-Places365 [GitHub repository]. GitHub. https://github.com/GKalliatakis/Keras-VGG16-places365

[15] Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[16] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition (arXiv preprint). https://arxiv.org/abs/1512.03385

[17] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171–4186). Association for Computational Linguistics.

[18] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

[19] Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., & Neves, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. In Findings of the Association for Computational Linguistics: EMNLP 2020 (pp. 1644-1650). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.148

[20] Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 86–96). Berlin, Germany: Association for Computational Linguistics.

[21] Encord. (2023, July 14). A complete guide to contrastive learning. https://encord.com/blog/guide-to-contrastive-learning/

[22] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709. https://arxiv.org/abs/2002.05709

[23] Weng, L. (2021, May 31). Contrastive learning: The state of the art. Lil'Log. https://lilianweng.github.io/posts/2021-05-31-contrastive/

[24] Sohn, K. (2016). Improved deep metric learning with multi-class N-pair loss objective. In Advances in Neural Information Processing Systems (pp. 1857–1865).

[25] Stahlschmidt, S. R., Ulfenborg, B., & Synnergren, J. (2022). Multimodal deep learning for biomedical data fusion: a review. Briefings in bioinformatics, 23(2), bbab569. https://doi.org/10.1093/bib/bbab569

[26] Yoo, Y., Tang, L. Y. W., Li, D. K. B., Metz, L., Kolind, S., Traboulsee, A. L., & Tam, R. C. (2017). Deep learning of brain lesion patterns and user-defined clinical and MRI features for predicting conversion to multiple sclerosis from clinically isolated syndrome. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 7(3), 250–259. https://doi.org/10.1080/21681163.2017.1356750

[27] Torres Soto, J., Hughes, J. W., Sanchez, P. A., Perez, M., Ouyang, D., & Ashley, E. (2021). *Multimodal deep learning enhances diagnostic precision in left ventricular hypertrophy*. *medRxiv*. https://doi.org/10.1101/2021.06.13.21258860

[28] Reda, I., Khalil, A., Elmogy, M., Abou El-Fetouh, A., Shalaby, A., Abou El-Ghar, M., Elmaghraby, A., Ghazal, M., & El-Baz, A. (2018). Deep Learning Role in Early Diagnosis of Prostate Cancer. Technology in cancer research & treatment, 17, 1533034618775530. https://doi.org/10.1177/1533034618775530

[29] Wang, H., Subramanian, V., & Syeda-Mahmood, T. (2021). Modeling uncertainty in multi-modal fusion for lung cancer survival analysis. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)* (pp. 1169–1172). IEEE.

[30] Tsanousa, A., Meditskos, G., Vrochidis, S., & Kompatsiaris, I. (2019). A weighted late fusion framework for recognizing human activity from wearable sensors. *In 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA)* (pp. 1-8). IEEE. https://doi.org/10.1109/IISA.2019.8900725