# MULTIMODAL MULTI-LABEL CLASSIFICATION FOR THEMATIC GROUPING OF ARTWORKS

**Meenu Arasada**
Juanita High School
MehtA+ and Williams University

**Sriya Mandapati**
Hun School of Princeton
MehtA+ and Williams University

**Richard Pan**
West High School
MehtA+ and Williams University

July 27, 2023

**ABSTRACT**

Art galleries and museums house vast collections comprising tens of thousands of paintings, making it challenging for students and researchers from specific majors to find artworks that align with their academic interests. The absence of a standardized classification system further compounds this issue, preventing efficient search. To address this critical need, our research paper presents a novel method that leverages both the artwork and its metadata (consisting of descriptions of the painting) to perform thematic grouping of artworks based on majors. Our custom-built CNN model achieves a training accuracy of 89.11% and a validation accuracy of 67.76% for classifying an artwork with its most comparable major, and extends its capability to categorize it into multiple other similar majors. Through rigorous experimentation and evaluation, we showcase the potential of deep learning architectures in enhancing the museum experience and improving the accessibility of artworks within museum collections, providing a valuable resource for art enthusiasts, researchers, and museum curators.

## 1 Introduction

Art galleries and museums house extensive and diverse collections of artworks, constituting tens of thousands of paintings from various periods, genres, and artistic styles. These repositories serve as cultural hubs, preserving our heritage and fostering appreciation for the arts.

However, navigating through such massive collections to find artworks that align with specific interests, especially for students and researchers from particular majors, proves to be a tedious challenge due to the absence of a classification system.

The importance of effective artwork classification cannot be understated. A well-organized and accessible collection benefits a plethora of stakeholders, from art enthusiasts and researchers to museum curators and educators. A well-designed classification system can enhance the museum experience, allowing visitors to explore artworks through thematic lenses. Moreover, such a system can facilitate targeted research, enabling scholars to delve into specific themes within the art world. In this paper, we present a novel method that combines both visual and textual features to perform thematic grouping of artworks based on their related majors.

## 2 Related Work

The realm of art classification and curation has seen significant interest from researchers, driven by advancements in deep learning techniques[1, 2]. One common approach is the utilization of image-based techniques such as CNNs for visual feature extraction [3]. Various CNN architectures have been used to identify artistic styles and genres. However, while these methods excel at image-based classification, they often neglect contextual information.

On the other hand, some studies focus on textual analysis and natural language processing to extract themes and concepts from artwork descriptions[3]. These approaches use vectorizers such as TF-IDF, Bag-of-Words, or LDR to

reveal underlying patterns in the text. However, by solely relying on textual information, these models do not fully capture the visual complexities present in artworks.

To bridge the gap between image and text analysis, we explored a multimodal approach that combines both visual and textual features for artwork classification. Our approach offers a holistic solution by utilizing both visual and textual data to achieve a more accurate categorization of artworks.

## 3 Dataset

Our model utilized data containing essential information related to the Williams College Museum of Art (WCMA) events and exhibitions. The WCMA events dataset was an extensive compilation of over 16,000 rows, detailing classes that visited the specific artworks. Additionally, the WCMA exhibitions dataset comprised over 10,000 rows, documenting exhibitions hosted by the Williams College, featuring artworks from the museum's collection. We also had access to WCMA's collection images dataset, which consisted of metadata for all the artworks housed in the museum.

### 3.1 Preprocessing

To create a cohesive dataset, the exhibition and events datasets were mapped to the collection images dataset using the TMS ObjectID column through an inner-join operation. By using this approach, only the rows with matching keys in both datasets were retained, allowing for the identification of corresponding images through their unique IDs. This alignment ensured that each entry in the dataset was associated with its respective artwork, laying the foundation for subsequent refinement.

To streamline the classification process, we chose 12 distinct majors, representative of the diverse academic offerings at Williams College. We then deployed a BERT model to assign each entry in the exhibitions and events dataset to one of the chosen majors. This BERT model served as a baseline, providing an initial classification based on the available information.

Mappings from the BERT model were scrutinized and updated as necessary to ensure the highest possible accuracy. This was a crucial step aimed to enhance the reliability of the dataset, ultimately contributing to the accuracy of our models. Figures 4 and 5 summarize the preprocessed datasets.

## 4 Model

Throughout the research, we experimented with various models for the Convolutional Neural Network (CNN), including VGG-16 and ResNet-50. Additionally, we explored the impact of employing both unimodal and multimodal architectures with the same base CNN model. Further details and results can be found in the Appendix section of this paper.

### 4.1 CNN Model

After rigorous experimentation, we identified a custom-built, multimodal CNN model that emerged as the most accurate in our thematic grouping task. This model combines both image features and textual data, and operates as follows:

Input: Resized images of dimensions 224x224 pixels are fed into the CNN layers.

CNN Layers: The image data undergoes several convolutional layers, designed to extract relevant visual features from the input images.

Flattening: The outputs from the CNN layers are then flattened to create a feature vector.

Vectorized Additional Features: The feature vector is concatenated with the vectorized additional features, which were transformed into Bag of Words representations.

Fully Connected Layers: The concatenated feature vector is passed through two fully connected layers. This process allows the model to learn and capture complex relationships between image features and textual data.

Output Layer: The final fully connected layer contains twelve output nodes, each representing a different major. The CNN model predicts the thematic category for each artwork based on these output nodes.

Figure 1 is the internal architecture of the mutlimodal CNN model.

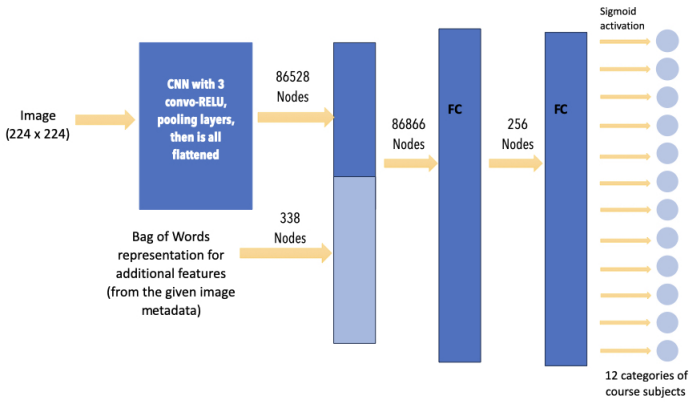


Figure 1: The internal architecture of the CNN model

### 4.2 Loss Function

Loss function defined below:

$$L = \frac{-1}{output\ size} \sum_{i=1}^{output\ size} y_i \cdot \log \widehat{y}_i + (1 - y_i) \cdot \log(1 - \widehat{y}_i) \tag{1}$$

Binary Cross Entropy is an ideal loss function for our multi-label thematic grouping model due to its suitability for handling independent labels. Additionally, the function's ability to handle class imbalances ensures fair treatment of less frequent majors, contributing to a more balanced model training. Its smooth and convex nature further aids efficient optimization during training, making it a fitting choice for our model's multi-label classification objectives.

### 4.3 Activation Function

Activation function defined below:

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2}$$

The benefit of using the Sigmoid activation function in our multi-label thematic grouping model is due to its ability to produce probabilistic outputs. The Sigmoid function maps the model's raw output to a range between 0 and 1, representing the probability that each major is present in the artwork. This allows us to interpret the model's predictions as the likelihood of each label's presence independently. By using Sigmoid activation, our model can handle multi-label classification efficiently, as it assigns different probabilities to each major without being constrained by the presence of other labels.

## 5 Results

The CNN yielded a validation accuracy of 67.76% and a training accuracy of 89.11%.

| Model | Global Accuracy | Global Loss | # of epochs trained |
|---|---|---|---|
| CNN (additional features) | Train: 89.11% / Val: 67.76% | Train: 0.0494/ Val: 0.1610 | 50 |

Figure 2: Training and Validation Accuracy of multimodal CNN model

### 5.1 Model Evaluation Metric

To assess the effectiveness of our models, we defined accuracy as the percentage of predictions that matched the true label for each artwork. We focused on the top 1 prediction for the model, rather than considering multiple labels that could be predicted.

## 6 Criticism

| | |
|---|---|
| Africana Studies | 1 |
| American Studies | 2 |
| Arabic Studies | 3 |
| Classics (Greek and Latin) | 4 |
| East Asian Languages and Culture | 5 |
| Environmental Studies | 6 |
| History | 7 |
| Religion | 8 |
| Science and Technology Studies | 9 |
| Sociology | 10 |
| Women's, Gender, and Sexuality Studies | 11 |

Figure 3: Legend used for Figures 4 and 5

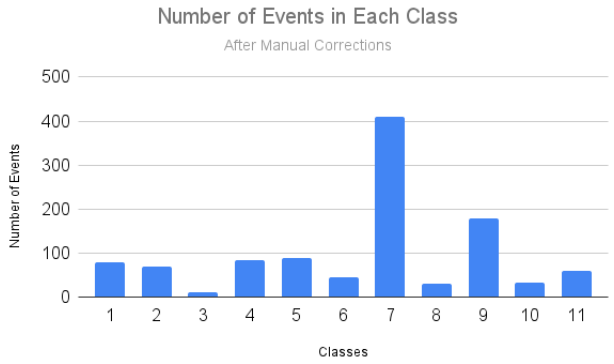


Figure 4: Displays the distribution of themes in the events dataset after manual editions

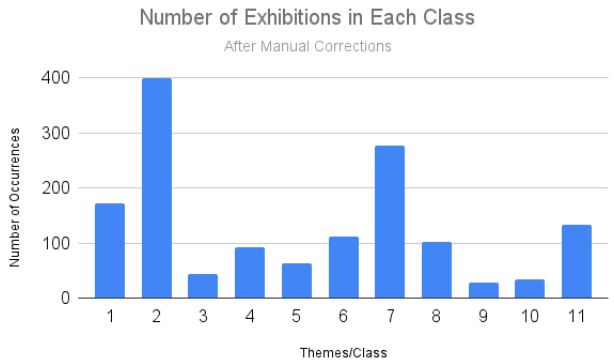


Figure 5: Displays the distribution of themes in the exhibits dataset after manual editions

One concern was the imbalanced nature of the dataset shown in figures 4 and 5. As the number of artworks per major varied, some labels were more prevalent than others, potentially leading to biased predictions and unequal representation. While the Binary Cross Entropy Loss function handled the class imbalance to some extent, techniques such as class weighting could have been applied to mitigate the impact of imbalanced data.

Another limitation was the restriction imposed by computing power. The complexity of deep learning models and the large dataset size demanded substantial computational resources. Due to these limitations, we had to make compromises in model architecture and training duration, which might have affected the model's performance and generalization.

Moreover, while traditional accuracy is a commonly used metric, it may not be the most suitable evaluation measure for our multi-label classification task. Since our model can predict multiple majors for each artwork, the true accuracy is likely higher, as the model can make accurate predictions for other majors even if the top prediction does not exactly match. Alternative evaluation metrics, such as the F1 score or Jaccard index, could offer a more comprehensive assessment of our model's performance.

# 7 Conclusion

In conclusion, our research paper presents a novel approach for thematic grouping of artworks based on multi-label classification, leveraging both visual and textual data. Our contributions hold valuable implications for art enthusiasts, researchers, and museum curators, and transforms the way they explore and appreciate the painting present in art galleries and museums.

# 8 Future Work

There are several avenues for future work and improvement- by handling class imbalance, incorporating more data, and using a better accuray metric, we can enhance the existing model. We can also explore the model's applications beyond art, such as thematic grouping of historical documents, literature, and other domains.

# 9 Appendix

## 9.1 Other Attempts

Figure 6 displays the themes that all of our models produced for the given image.

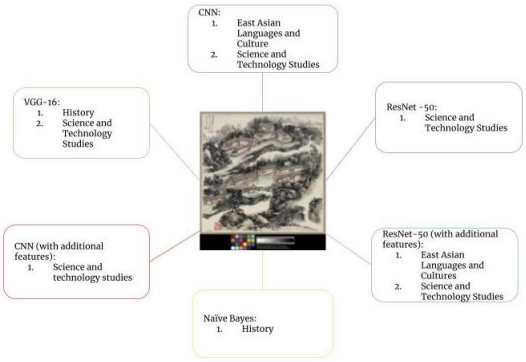


Figure 6: Resulting classes given this specific artwork

### 9.1.1 CNN Unimodal Model

Our first CNN model was constructed using the TensorFlow Keras sequential model. Comprising three convolutional-ReLU and max-pooling layers, this unimodal model solely processed the images. The flattened output from the CNN was directly fed into the dense layers for label prediction.

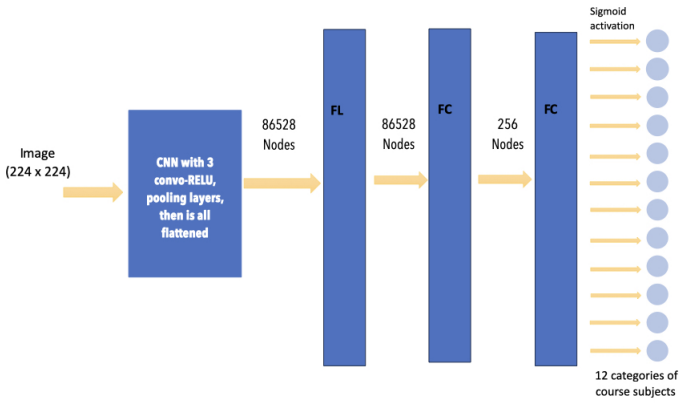


Figure 7: The internal architecture of the unimodal CNN model

| Model | Global Accuracy | Global Loss | # of epochs trained |
|---|---|---|---|
| CNN | Train: 67.49% / Val: 64.47% | Train: 0.1146/ Val: 0.1610 | 32 |

Figure 8: Training and Validation Accuracy of Unimodal CNN model

### 9.1.2 VGG-16 Models

The VGG-16 model served as another unimodal approach, processing the images through its layers to effectively recognize features and categorize the artworks.

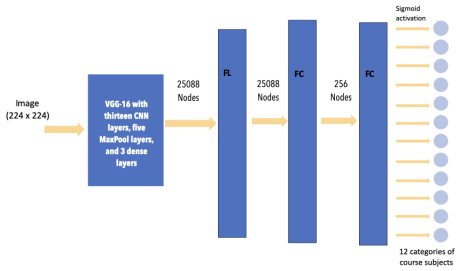


Figure 9: The internal architecture of the VGG-16 model

| Model | Global Accuracy | Global Loss | # of epochs trained |
|---|---|---|---|
| VGG-16 | Train: 66.34% / Val: 62.50% | Train: 0.1430/ Val: 0.1829 | 42 |

Figure 10: Training and Validation Accuracy of VGG-16 Model

### 9.1.3 ResNet-50 Models

Experimentation with the ResNet-50 model involved fine-tuning it on the provided images and data. Two versions of ResNet-50 were explored: a unimodal and a multimodal approach. Similar to the CNN, the unimodal ResNet-50 processed the images through its layers, and the flattened output underwent direct processing through the dense layers. In contrast, the multimodal version concatenated the flattened output from ResNet-50 with Bag of Words vectors representing the textual data before passing through the dense layers.

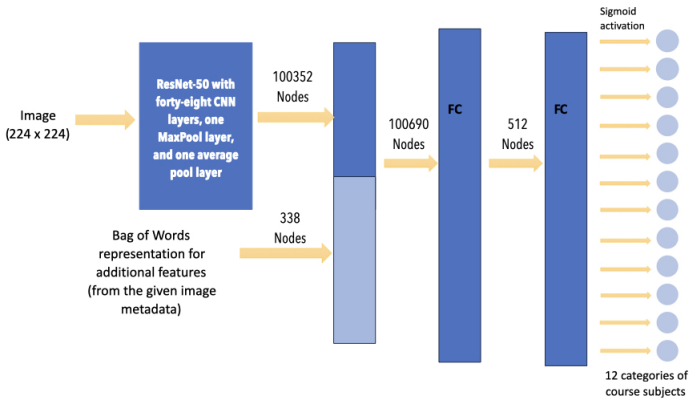


Figure 11: The internal architecture of the mutlimodal ResNet-50 model

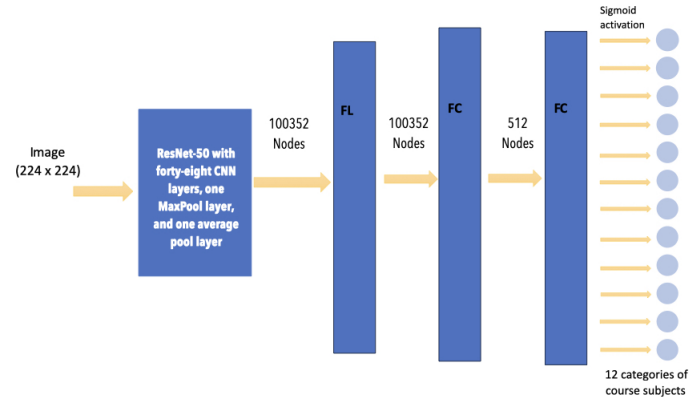


Figure 12: The internal architecture of the unimodal ResNet-50 model

| Model | Global Accuracy | Global Loss | # of epochs trained |
|---|---|---|---|
| Res-Net 50 | Train: 67.49% / Val: 65.79% | Train: 0.1130/ Val: 0.1557 | 39 |
| Res-Net 50 (additional features) | Train: 67.00% / Val: 64.47% | Train: 0.1114/ Val: 0.1863 | 31 |

Figure 13: Training and Validation Accuracy of ResNet-50 Models

### 9.1.4 Naive Bayes Model

In addition to deep learning models, we also employed a Naive Bayes model on the provided data with images. The model leveraged information from the 'Title' and 'Subject' columns in the mapped dataframes to determine the probability of each major.

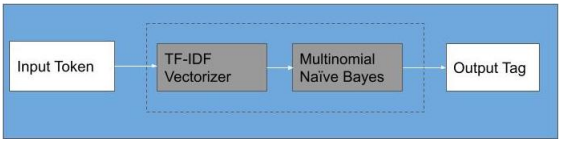


Figure 14: The internal architecture of the Naive Bayes Model

| Model | Global Accuracy | Global Loss | # of epochs trained |
|---|---|---|---|
| Naïve Bayes | Train: 68.34%/ Val: 61.88% | N/A | N/A |

Figure 15: Training and Validation Accuracy of the Naïve Bayes Model

Bayes theorem equation:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3}$$

### 9.2 Division of Labor

All members contributed equally to each part of the project. Special acknowledgment is made to Richard for training our categorization models; Meenu for developing an image captioning model and working on most of the paper; and Sriya for working on most of the poster and visuals.

### 9.3 Acknowledgements

We would like to sincerely thank our instructors, Ms. Haripriya, Mr. Bhagirath, and Ms. Minnie, for guiding and supporting us. We would also like to give our gratitude to Dr. Beth Fischer from Williams College for giving us ideas and structure through this project.

## References

[1] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[3] Moses Soh. Learning cnn-lstm architectures for image caption generation. *Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep*, 1, 2016.