HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 4 (2024)

Deep Learning for Comic Book Emotion Analysis

Walter Hsieh[1] and Eric Sakk[#]

[1]Taipei Fuhsing Private School, Taiwan
[#]Advisor

ABSTRACT

Deep learning techniques have been applied successfully in a number of fields, including computer vision and image processing. This study presents a comprehensive approach to analyzing the sentiment conveyed in comic book panels through the emotions depicted on characters' faces. We hypothesize that it is possible to accurately interpret emotional content using deep learning techniques even without a pre-existing sentiment dataset. Using optical character recognition and pretrained sentiment analysis models, the basis of an NLP model was formed to comprehend emotional context from characters' dialogues and thoughts. A neural network then categorizes emotions exhibited by characters' facial expressions. Our findings confirm that sentiment analysis can indeed be performed on comic book data, with tests on the Digital Comic Museum dataset demonstrating sentiment analysis efficacy of 89%. Then, the optimal configuration of convolutional neural networks was identified with 7x7 filters, 200 neurons, and 64 filters per layer, achieving an accuracy of 86%. This research advances the capability of facial recognition technology, expanding its application from humans to fictional characters in comic books. It also sets the groundwork for future research to generate datasets when specialized data is not available, demonstrating the practicality of performing sentiment analysis on comic book faces.

Introduction

In recent years, research in computer vision and image processing has advanced the analysis of comic books, especially in the realm of digitized comics, where art and text within originally on-paper comics are stored digitally to enable easier information access both by humans and computers.1 The development of image analysis techniques to encode comic book images into text files allows for computer understanding of images within the comic and ability to split comic images into indexable panels.2 Other models, using computer vision alongside natural language processing, are able to analyze emotions in comic documents via visual information like speech balloons, and onomatopoeia.

The 2024 paper, "EmoComicNet: A multi-task model for comic emotion recognition," presents a notable advancement in comic emotion analysis by introducing EmoComicNet. This model employs a multi-task framework integrating image and textual modules to enhance emotion recognition in comics. The image module uses a ResNet-152 based feature extractor, while the textual module utilizes BERT embeddings and a bi-directional Gated Recurrent Unit (BiGRU).3 By combining these features, the EmoComicNet model achieves higher accuracy in predicting emotions, even when dealing with weak or missing modalities. This significant progress in multi-modal analysis paves the way for a more comprehensive understanding of comic narratives.

The primary objective of this research is to construct a model using machine learning capable of predicting the sentiment conveyed in a comic panel by analyzing the emotions depicted on characters' faces. This study investigates the feasibility of sentiment analysis within comic books, focusing on the challenge of interpreting emotional content in a medium that combines textual and visual elements. To achieve this, the study proposes a two-step approach. Initially, Optical Character Recognition (OCR) and pretrained sentiment analysis models will be employed to interpret the sentiments expressed within text boxes present in the comic panels. This textual sentiment analysis will

serve as the foundation for developing a Natural Language Processing (NLP) model that comprehends the emotional context shown through characters' dialogues and thoughts.4,5 Once the NLP model is established, it can be integrated with a neural network designed specifically to categorize the emotions exhibited by characters in the comic book. The information extracted from the text boxes will function as a validation mechanism, ensuring the accuracy of the neural network's ability to recognize and interpret the emotions depicted on the characters' faces.

By combining OCR, pretrained models, NLP, and a dedicated neural network, this research aims to create a sophisticated system that not only understands the sentiments expressed in the textual components of comic books but also accurately identifies and categorizes the emotional states portrayed by characters through facial expressions. Our results indicate that sentiment analysis can be performed on comic book data, showcasing a promising path forward for the application of these technologies in understanding the nuanced emotional expressions of comic characters. This approach utilizes the capabilities of both computer vision and natural language processing, offering a promising avenue for advancing the already-present facial recognition technology from being limited to real humans into being able to recognize emotions in fake characters, starting with comic book characters.

Our investigation into text analysis within comic panels took a significant leap forward after conducting thorough tests on the Digital Comic Museum (DCM) database.6 From this extensive exploration, we successfully identified a handful of models that demonstrated exceptional aptitude in dissecting and interpreting the content within speech bubbles. By meticulously adjusting parameters such as sharpening, blurring, and thresholding values in the OCR models, we achieved remarkable results, attaining text scanning accuracy levels as high as 90%. Concurrently, our efforts in sentiment analysis culminated in the development of a sentiment detection model with an impressive efficacy rate nearing 90%. This success paves the way for a deeper understanding of the emotional undertones embedded in comic narratives.

The integration of these specialized models, combining the power of computer vision, deep learning, and sentiment analysis, holds promise for our future endeavors. Not only do they enhance our ability to decipher text within speech bubbles, but they also lay the foundation for constructing a functional training dataset. This dataset will be used in training models to recognize facial sentiments in comic book characters, providing a more accurate comprehension of the emotions portrayed in the visual storytelling medium. Creating this training and testing datasets using pre-existing sentiment analysis models and text also presents a new way to build training sets for supervised learning applications.

The structure of this paper is organized as follows: The Introduction section provides an overview of the research background, objectives, and significance of sentiment analysis in comic books using deep learning techniques. The Methods section details the datasets used, image preprocessing techniques, OCR processes, NLP models, and the architecture of convolutional neural networks (CNNs) employed for facial emotion recognition. In the Results and Discussion section, we present the findings from various experimental configurations of CNNs, highlighting the optimal performance metrics and accuracy achieved. The Conclusions section summarizes the key contributions of this research, discusses the implications of the results, and suggests potential avenues for future work to enhance the system's capabilities and applications.

Methods

To ultimately be able to construct a convolutional neural network to be able to predict the sentiment on comic characters' faces, we first used DCM Dataset images and some "groundtruth" files located in the dataset in order to extract the faces and speech bubbles of comic characters. Then, optical character recognition was conducted on the speech bubbles for text extraction, which was analyzed based on their sentiment and turned into the values for the training and test sets of our convolutional network. Then, along with the faces extracted at the beginning, we were able to train convolutional neural network models to predict the sentiments of these comic book characters. Figure 1 is a diagram of this entire process, and for more specifics of the process, continue reading subsections of this methods section.

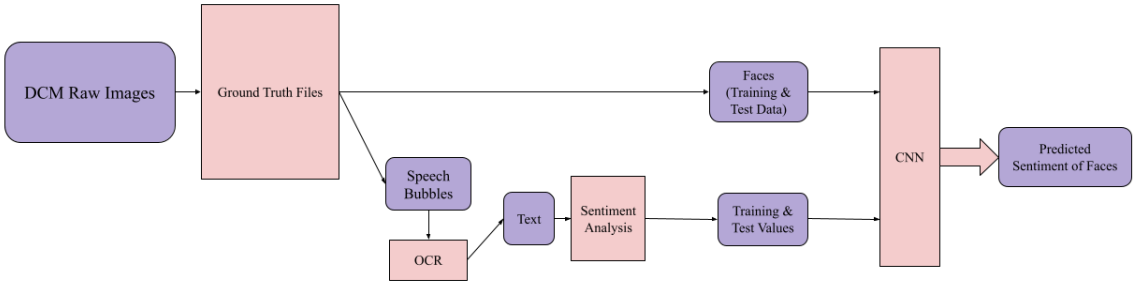

Figure 1. Diagram of Methodology Used to Conduct this Research

DCM Dataset

The DCM dataset utilized in our study is a dataset of 772 images annotated with vector-image files that mark characters and dialogue, with each page being from one of the 27 unique comics collected from the Digital Comics Museum's (DCM) digital comic books.[2,6] This dataset ensures its inclusion of different styles of comic books by including works from different authors. This dataset would be good for my study because a variety of different styles in the training set can prepare the model for different styles that it may encounter in the future. These images in the dataset were annotated with ground-truth bounding boxes covering all panels and characters, giving me a way to locate characters and their dialogue using the coordinates in the ground-truth files.[7]

Characters in the dataset are grouped into four distinct types: human-like, object-like, animal-like, and other supporting role characters. Faces were annotated only for the human-like class, covering features like eyebrows, eyes, nose, mouth, chin, and ears if visible. These will be the characters that we will be focusing on, as computer vision for emotion detection mainly involves the reading of facial emotions based on these facial structures, and it would be impractical to have a neural network train to detect facial emotion based on nothing except for faces.[8]

The dataset repository is organized with image files stored alongside ground-truth annotations. These annotations are provided in text files within the groundtruth folder, using a specific encoding format denoting class identifiers along with corresponding bounding box coordinates.[9] Furthermore, additional annotations, such as links between faces, speech bubbles, and characters, are available in SVG and CSV formats. These vectors and coordinates made it possible for us to track which characters said which of the dialogues.[10] This made it possible to track characters' faces along with the sentiment of their speech to create a training dataset for our comic book character facial emotion detection model. In our study, we use the annotated images to train computer vision algorithms in tasks such as panel detection, character recognition, and sentiment analysis. The bounding box annotations provided for panels, characters, and faces allow us to precisely identify these elements within the comic book images.

Image Pre-Processing

Upon extracting the text, we immediately encountered the challenge of accurately transcribing text from speech bubbles in comic books. Comic books, as a visual medium, exhibit a rich diversity of fonts, styles, and orientations within speech bubbles, making it difficult for standard Optical Character Recognition (OCR) models to decipher them accurately. These variations necessitated a solution for standardizing the text within comic books to effectively analyze their content. Before adjusting images to enhance OCR accuracy, one method we employed to train the OCR model to recognize various comic book styles was by utilizing the DCM dataset mentioned earlier. This approach not only

increased the diversity of content the OCR model was trained on but also expanded the range of comic book word styles the model would be capable of reading in the future.

Certain image preprocessing techniques were required first to improve the clarity of the text within the speech bubbles, at least to jumpstart the making of our first dataset.[11] For convenience purposes to generate a usable dataset without fixating on perfecting it, rather simpler techniques of image gray-scaling, image thresholding, gaussian blur, and sharpening. Image thresholding was used on the grayscale image to first convert it into a binary image. Then, by fine-tuning the OCR model settings and employing the preprocessing techniques of Gaussian blurring and image sharpening, we achieved significant improvements in text recognition accuracy. Our optimal combination involved using a lower thresholding value of 160, a higher thresholding value of 255, and a touch of Gaussian blurring followed by image sharpening. This formula worked wonders, enabling us to extract text from comic book speech bubbles with precision while preserving the integrity of the original images.

In order to facilitate text processing, images were first converted from RGB color images to grayscale images. Figure 2 below is an original image of a text bubble from *48 Famous Americans* in the DCM Dataset. Figure 3 shows a grayscaled version of Figure 2.

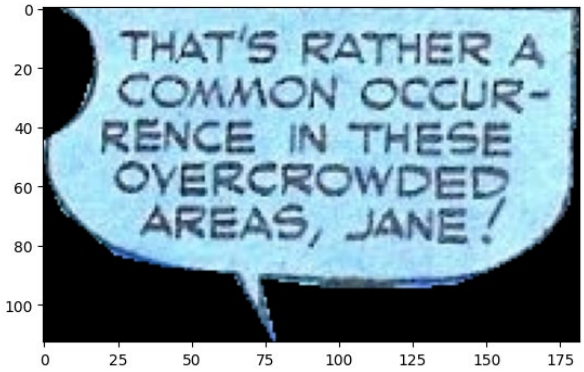

Figure 2. Original image of textbox from DCM Dataset

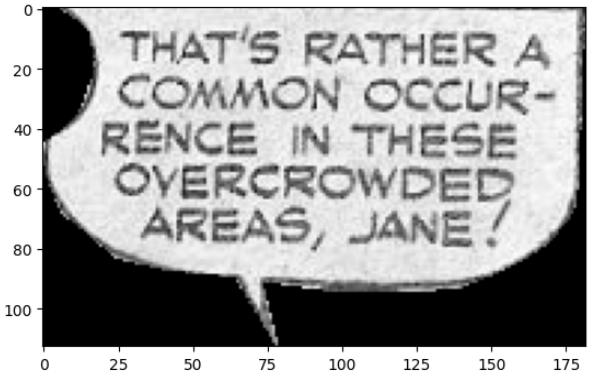

Figure 3. Grayscale version of Figure 2

Image Thresholding

Thresholding is a technique used to simplify images by reducing them to black and white, where pixels with intensity values above a certain threshold are set to white, and those below the threshold are set to black.[12] This process effectively isolates the text within the comic book speech bubbles, making it easier for the OCR model to recognize and extract. Figure 4 depicts a thresholded version of the grayscaled text bubble in Figure 3.

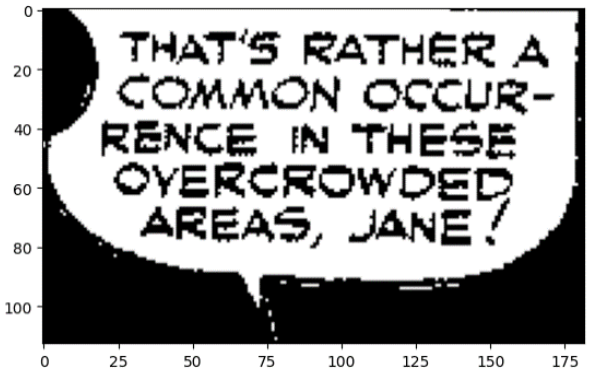

Figure 4. Thresholded version of Figure 3

Gaussian Blur

Gaussian blur is a widely used image-blurring technique that employs a Gaussian function for transforming each pixel in an image.[13] Unlike other blurring methods like circular box blur, Gaussian blur reduces high-frequency components in the image, effectively acting as a low-pass filter, using the Gaussian function as seen below to do so. One of the advantages of Gaussian blur is its circular symmetry, allowing it to be applied to a two-dimensional image as two separate one-dimensional calculations. This property makes Gaussian blur a separable filter, which significantly reduces computational complexity, making it more effective compared to other blurring methods that have non-separable kernels. Equation 1 below is the equation, the Gaussian function, that is used in this blurring mechanism.

$$G(x,y) = \frac{1}{2\pi} e^{-(x^2+y^2)}$$

Equation 1. Gaussian Function

The Gaussian function, being a low-pass filter, smooths the image by averaging pixel values with a Gaussian-weighted average of its neighbors. This process helps in reducing noise and detail, making it particularly useful in preprocessing steps for tasks such as edge detection and image segmentation. Figure 5 below shows the previous example of Figure 4 being blurred using this technique.

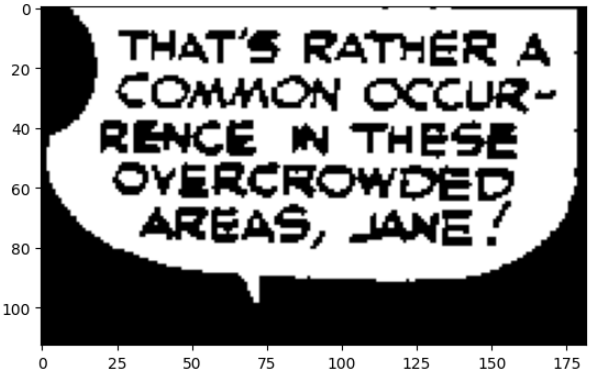

Figure 5. Blurred version of Figure 4 Using Gaussian Blur

Image Sharpening

Image sharpening, or high pass filtering, is a technique used to enhance the edges and details in an image, making them stand out more.[14] In the context of preparing images for OCR, sharpening helps in making the text within speech bubbles stand out more clearly against the background. This is particularly important for comic books where text might be blended with background comic book artwork and colors. Image sharpening works by allowing high-frequency components like edges to pass through while reducing the low-frequency components. This process can be mathematically described by the convolution of the image with a high-pass filter kernel. As it can be observed in Figure 6 below, it is much cleaner compared to Figure 4, before it was blurred and would theoretically produce better results in the OCR process of determining the correct text in these speech bubbles.

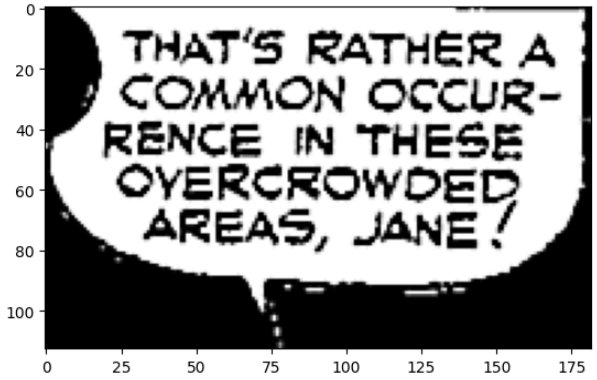

Figure 6. Sharpened version of Figure 5 Using High Pass Filtering

The combination of image preprocessing techniques – grayscale conversion, thresholding, Gaussian blur, and sharpening – along with fine-tuning the OCR model, forms a robust approach to accurately transcribing text from comic book speech bubbles. This multi-step process not only enhances the visual quality of the text but also ensures that the OCR model can handle the diverse and complex styles found in comic books, ultimately leading to more reliable and precise text extraction.

We achieved the highest accuracy (based on human interpretation) of 90% accurate text recognition using the 48 Famous Americans comic book speech bubbles. This high level of accuracy demonstrates the effectiveness of our preprocessing and OCR fine-tuning approach, making it a promising and practical solution for similar applications in other comic book studies lacking a sufficient dataset.

Training Set Construction

The construction of an effective training set is pivotal for developing a robust NLP model capable of accurately interpreting sentiments conveyed in comic book dialogues. After using adjusting speech bubbles using the methods described above, we were able to link each of the speech bubbles to a face of a character. This section describes how the training dataset for our convolutional neural network was built.

Sentiment Analysis

Sentiment analysis in comic books involves interpreting the emotional tone of dialogues within speech bubbles. We utilized pretrained sentiment analysis models, such as BERT, fine-tuned on our annotated dataset to detect the underlying emotions.[15,16]

The fine-tuned models were used to classify the sentiment of each dialogue into the predefined classes of negative, positive, and neutral. This step enabled us to understand the emotional tone of the text within speech bubbles,

which was essential for integrating with the neural network for facial emotion recognition. The accuracy and reliability of our sentiment analysis models were validated through testing and comparison with human interpretations. This ensured that the NLP model could accurately comprehend the emotional context of characters' dialogues. 40 images from the comic *48 Famous Americans* were extracted, and based on human interpretation of the sentiment of each of the texts in the speech bubbles compared to the sentiment obtained from pre-trained sentiment analysis models, the models were able to achieve the greatest accuracy with the model BERTweet, a model used primarily to analyze sentiment of tweets, with an accuracy of 89%.

CNN Architectures

Convolutional Neural Networks (CNNs) are a class of deep learning models specifically designed to handle grid-like data, such as images.[17,18] Unlike traditional neural networks, which fully connect each neuron to all neurons in the subsequent layer, CNNs use convolutional layers where neurons, organized in filters, connect only to local regions of the input. Each neuron within these filters scans the input, capturing local patterns such as edges or textures.[19]
A typical CNN consists of multiple layers, each with specific roles. Convolutional layers apply filters of various sizes (e.g., 3x3 or 5x5) to the input data to extract features. Pooling layers, such as max pooling or average pooling, follow to reduce the spatial dimensions of these feature maps, mitigating the computational complexity and aiding in generalization. The depth of the network, in terms of the number of layers and the increasing number of filters per layer, allows CNNs to learn increasingly abstract representations. Initial layers may capture simple features, while deeper layers combine these into complex patterns.

Using the data that we got from sentiment analysis of the larger data sets, specifically 80% of all of the data, we trained a convolutional neural network to recognize facial emotions on the characters' faces, extracted via additional vectors provided in the DCM files. From the text in the text bubbles, there were around 50 of both positive and negative emotions, while neutral emotions had a clear majority of nearly 280 total pieces of data.

In these experiments conducted for this research the CNNs consisted of two convolutional layers followed by fully connected layers for classification. 3x3 and 5x5 layers were tested, but in the final product, each convolutional layer employed 7x7 filters to capture the facial features of comic book characters. The choice of 7x7 filters was determined through experimentation to maximize accuracy in recognizing these facial expressions. After testing for 16, 32, 48, and 64 filters per layer, the number of filters per layer was optimized at 64, providing an optimized performance for the model. The final product had fully connected layers with 200 neurons that were used to process the extracted features from convolutional layers, culminating in a softmax layer for emotion classification. A depiction of this process is shown in Figure 7 below.

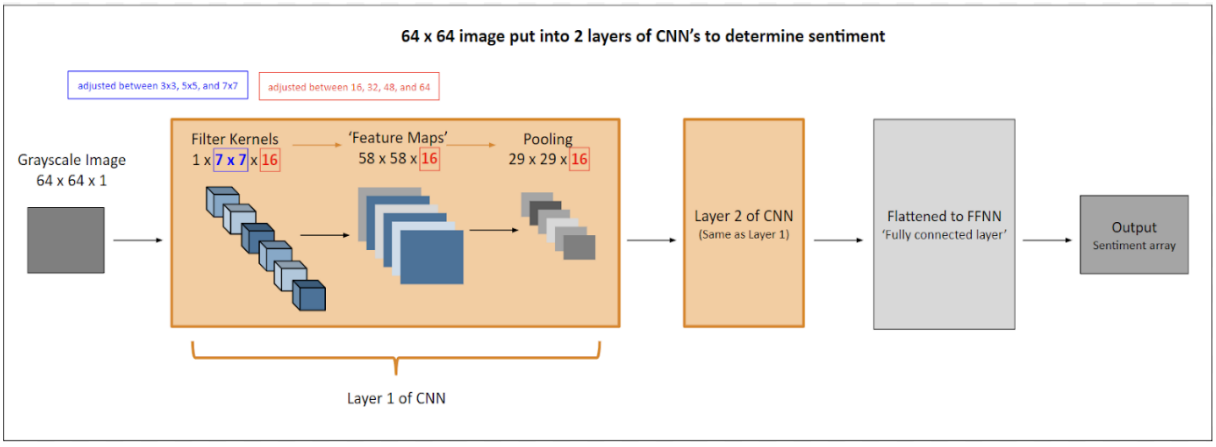

Figure 7. Example diagram of the process of the CNN experiments conducted

Results

The experiments tested various configurations of convolutional neural networks (CNNs) to determine the optimal architecture for recognizing facial emotions in comic book characters. We focused on the number of filters in each convolutional layer and the number of neurons in the fully connected layers.

When comparing different filter sizes, the 7x7 filters consistently outperformed the 3x3 and 5x5 filters across all neuron configurations. The 3x3 and 5x5 filters achieved lower overall accuracies, while the 3x3 filters exhibited a lot more variability in performance. From Table 1 presented below, while fixed at 64 filters per layer, it can be observed that 7x7 filters on each layer performed slightly better in accuracy than the 3x3 and 5x5 filters.

Table 1. Accuracy (%) Of Predictions With 64 Filters Each Layer

Neurons	3x3 filters	5x5 filters	7x7 filters
25	65.278	59.722	66.667
50	65.278	77.778	77.778
75	76.389	77.778	77.778
100	77.778	81.944	81.9444
200	69.444	83.333	86.111
AVG	70.833	76.111	78.056

This was why only the size 7x7 filters were used to compare accuracies for different numbers of neurons. Figure 8 below illustrates the relationship between the number of filters in each convolutional layer and the accuracy of the CNN, with different lines representing different numbers of neurons in the fully connected layers.

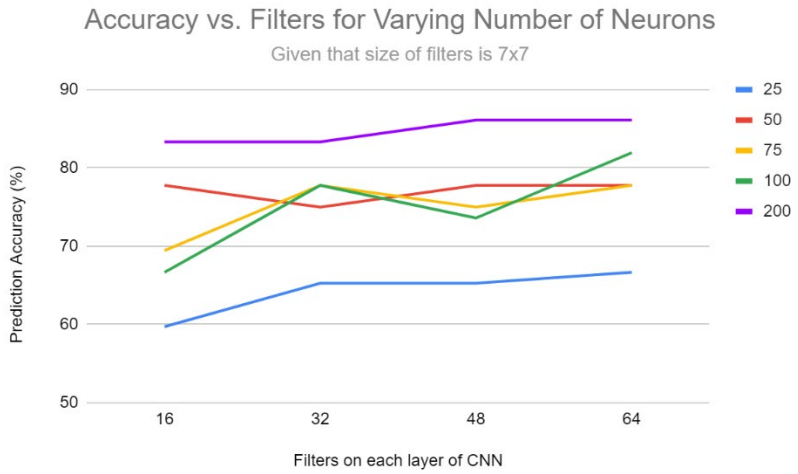

Figure 8. Accuracy vs. Filter Line Graph for Varying Numbers of Neurons

Discussion

The results indicate that the number of neurons in the fully connected layers, the number of filters in each convolutional layer, and the filter size significantly influence the model's accuracy. Models with 25 neurons showed consistent but limited improvement, constrained to a maximum accuracy of 70%. This suggests that while these models can benefit from more filters, their capacity is inherently limited by the smaller number of neurons.

In contrast, models with 50, 75, and 100 neurons showed more variability in performance with fewer filters, yet all achieved peak accuracy at around 80%, with the case of 64 filters on each layer with 100 neurons being the only case crossing that 80% threshold. This indicates that a moderate increase in model complexity, both in terms of neurons and filters, can lead to performance gains.

The final model with 200 neurons consistently achieved an accuracy of around 85%, with a highest accuracy of 86%, across all filter configurations, showing a moderate jump from the tests ran on CNNs with 100 or less neurons. While we did not perform further tests on increasing the number of neurons in the CNN, in each of the trendlines in Figure 8, results don't really relate to the number of filters on each layer, as the performance plateaued as the number of filters increased.

To further analyze the accuracy of the predictions presented from the percentage number, we graphed the actual predicted results of our most accurate neural network onto a confusion matrix, comparing the true values according to our training set to the predicted values generated by the CNN. Such a confusion matrix can be seen below in Figure 9.

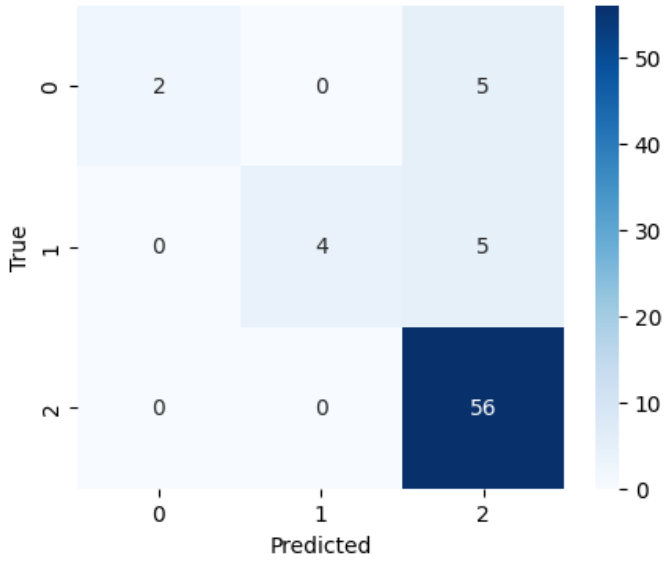

Figure 9. Confusion Matrix of True vs. Predicted Values of Sentiment in Comic Character Faces

The best results yielded in these tests was the last datapoint shown in Table 1, 86.111% with 64 7x7 filters on the two layers in the CNN with 200 total neurons. Figure 6 is a depiction of this 86% accuracy broken down into a confusion matrix. In this confusion matrix, the 0's represent the "negative" sentiment and the 1's represent the "positive" sentiment. While it may seem like 86.111% is quite an impressive accuracy, with the model predicting 62 out of 72 of the test set correctly, all its incorrect predictions were "neutral", and apart from faces that are "neutral", the model predicts less than half of "negative" and "positive" data correctly. This could be due to a lack of a substantial amount of negative and positive training data compared to "neutral" faces linked to "neutral" comic book texts in the database.

Conclusion

This study successfully developed a comprehensive system for analyzing emotions in comic book panels by integrating OCR, pretrained sentiment analysis models, NLP, and CNN for facial emotion recognition. Key findings include achieving 90% text recognition accuracy and 89% sentiment analysis accuracy, demonstrating the effectiveness of the preprocessing and model fine-tuning techniques. The optimal CNN configuration, identified as 64 filters per layer, 200 neurons, and 7x7 filter size, achieved the highest accuracy at 86%, with indications of diminishing returns in increasing model complexity regarding the number of filters in each layer.

Future work will focus on several avenues to enhance the system. First, a major flaw in the model presented in this research is its inability to recognize non-neutral comic book faces due to the lack of nonneutral data presented in the training database. This could be improved by expanding the training dataset to include a larger variety and more balanced set of emotions in the comic characters' dialogue and facial expressions. In the future, including more comic styles and genres will also improve the generalization capability of the models. Exploring more advanced CNN architectures, such as residual networks (ResNets) or attention mechanisms, could potentially further improve accuracy by enhancing the integration between text and visual data.[20,21] Finally, optimizing the models for real-time processing could enable applications such as automated content tagging systems that could sort comic book content based on facial emotion detection.

Beyond comic books sentiment analysis, this study sets a foundation for future studies, creating training datasets when none are readily available. By using the given annotations from the DCM dataset and utilizing sentiment analysis models this research demonstrates the possibility of tailoring datasets to enhance the performance and accuracy of the final model. This approach to creating datasets sets the precedent for future research in domains like this one, where specialized datasets are scarce or non-existent.

Limitations

Despite the 86% success rate of the system, there are several limitations that need to be addressed in future work. A major flaw in the model presented in this research is its inability to recognize non-neutral comic book faces due to the lack of non-neutral data in the training database. This imbalance in the training data significantly impacted the model's ability to accurately predict emotions other than neutral, limiting its overall effectiveness.

Another limitation lies in the scope of the comic styles and genres used in the training dataset. The current model may not generalize well to a broader range of comic styles and genres, potentially affecting its performance on more diverse comic panels. The model's architecture, while effective, may also face limitations in further accuracy improvements. The study observed diminishing returns when increasing model complexity by adding more filters, suggesting that simply scaling up the model may not yield significant gains without exploring more advanced architectures, such as ResNets or attention mechanisms.

Finally, the current system is not optimized for real-time processing, which limits its applicability in dynamic or interactive environments. Future research should address these limitations by expanding the training dataset, exploring advanced CNN architectures, and optimizing the system for real-time applications to enhance its robustness and applicability across different contexts.

Acknowledgments

I would like to thank Dr. Christophe Rigaud for his assistance in accessing the DCM772 dataset. I would also like to express gratitude to my mentor, Dr. Eric Sakk, for his guidance on this project.

References

1. Sagri, M.; Sofos, F.; Mouzaki, D. Digital Storytelling, Comics And New Technologies In Education: Review, Research, And Perspectives; 2018; Vol. 17. https://openjournals.library.sydney.edu.au/index.php/IEJ.
2. Nguyen, N.-V.; Rigaud, C.; Burie, J.-C. Digital Comics Image Indexing Based On Deep Learning; 2018. https://doi.org/10.3390/jimaging4070089.
3. Dutta, A.; Biswas, S.; Das, A. K. EmoComicNet: A Multi-Task Model For Comic Emotion Recognition. Pattern Recognition 2024, 150, 110261. https://doi.org/10.1016/j.patcog.2024.110261.
4. Laubrock, J.; Dunst, A.; Cohn, N.; Magliano, J. P. Computational Approaches To Comics Analysis; 2020; Vol. 12. https://doi.org/10.1111/tops.12476.
5. Chaudhuri et al. Optical Character Recognition Systems For Different Languages With Soft Computing; 1st ed.; Springer Nature, 2017.
6. Rigaud, C. Dataset: Golden Age Comic Book Panels. Digital Comic Museum. http://digitalcomicmuseum.com.
7. Rigaud, C. DCM dataset. GitLab. https://gitlab.univ-lr.fr/crigau02/dcm-dataset.
8. Knowledge-driven understanding of images in comic books. International Journal on Document Analysis and Recognition 2015, 1433–2833. https://doi.org/10.1007/s10032-015-0243-1.
9. Rigaud, C.; Haxaire, A.; Karatzas, D.; Burie, J.-C.; Ogier, J.-M. An Active Contour Model for Speech Balloon Detection in Comics. Int. J. Document Anal. Recognit. 2015, 18 (2), 121-135.
10. Jha, S.; Agarwal, N.; Agarwal, S. Bringing Cartoons to Life: Towards Improved Cartoon Face Detection and Recognition Systems. arXiv 2018, arXiv:1804.01753. https://arxiv.org/abs/1804.01753.
11. Gonzalez, R.; Woods, R. Digital Image Processing Global Edition, 4th ed.; Pearson, 2017.
12. Huang, L.-K.; Wang, M.-J. J. Image thresholding by minimizing the measures of fuzziness. Pattern Recognition 1995, https://doi.org/10.1016/0031-3203(94)e0043-k.
13. D'Haeyer, J. P. F. Gaussian filtering of images: A regularization approach. *Signal Processing* 1989, *18* (2), 169–181. https://doi.org/10.1016/0165-1684(89)90048-0.
14. High Pass Filter Documentation. NV5 Geospatial. https://www.nv5geospatialsoftware.com/docs/HighPassFilter.html.
15. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805.
16. Hugging Face – The AI community building the future. https://huggingface.co/.
17. LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. Nature 2015, 521 (7553), 436-444.
18. Zhang, X.; Zhao, J.; LeCun, Y. Character-level Convolutional Networks for Text Classification. Adv. Neural Inf. Process. Syst. 2015, 28, 649-657.
19. Caceres, P., Introduction to Neural Network Models of Cognition. https://com-cog-book.github.io/com-cog-book/intro (2020).
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. arXiv.org. https://arxiv.org/abs/1512.03385.
21. Soydaner, D. Attention mechanism in neural networks: where it comes and where it goes. Neural Computing & Applications 2022, 34 (16), 13371–13385. https://doi.org/10.1007/s00521-022-07366-3.
22. Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. Adv. Neural Inf. Process. Syst. 2012, 25, 1097-1105.
23. Russakovsky, O.; Deng, J.; Su, H.; et al. ImageNet Large Scale Visual Recognition Challenge. Int. J. Comput. Vision 2015, 115 (3), 211-252.
24. Pal, N. R.; Bhandari, D. Image thresholding: Some new techniques. Signal Processing 1993, 33 (2), 139–158. https://doi.org/10.1016/0165-1684(93)90107-l.

25. Kim, Y. Convolutional Neural Networks for Sentence Classification. Proc. Conf. Empirical Methods Nat. Lang. Process. (EMNLP) 2014, 1746-1751.
26. Tesseract OCR. Tesseract: An Open Source OCR Engine. https://github.com/tesseract-ocr.
27. Sze, V.; Chen, Y.-H.; Yang, T.-J.; Emer, J. S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. Proc. IEEE 2017, 105 (12), 2295-2329.
28. Augereau, O.; Iwata, M.; Kise, K. A Survey of Comics Research in Computer Science; 2018; p 87. https://doi.org/10.3390/jimaging4070087.
29. Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, MA, 2016.