WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

BRAILLE TRANSLATION USING VISION-LANGUAGE MODELS: A TRANSFORMER-BASED APPROACH

[1] Prithvi Pothupogu, [2] Rohan Kiran Gunjal, [3] Shivam Sethia, [4] Levin Joji Mathews

[1 2 3 4] Dept. of Electronics and Communication Dept. of Electronics and Communication Dept. of Electronics and Communication NIT Warangal

ABSTRACT—

Braille translation from natural scenes is a complex challenge due to variations in lighting, distortions, and occlusions. Traditional Optical Braille Recognition (OBR) techniques rely on segmentation and classification but struggle with scalability and adaptability to real-world environments. We propose a novel transformer-based architecture that leverages Vision-Language Models (VLMs) such as CLIP, FLAVA, and LLAVA. Our model utilizes Vision Transformers (ViTs) for image encoding and a GPT-based transformer for text decoding, with a specialized fusion layer to map multimodal representations. The training process employs a teacher-forcing strategy and causal language modeling (CLM) loss, implemented on an A40 GPU. Our ap- proach demonstrates significant improvements in Braille extraction accuracy and adaptability across varied conditions.

Index Terms—Braille translation, Vision-Language Models, Vision Transformers, GPT, Causal Language Modeling, Image Segmentation, Optical Braille Recognition, Deep Learning.

Introduction

Braille is a vital tactile writing system that enables vi- sually impaired individuals to read and write using raised dot patterns. Despite its importance, accurately translating Braille from natural scene images remains a challenging task due to factors like lighting variations, distortions, and occlu- sions. Unlike printed text, Braille appears on diverse surfaces such as paper, metal plaques, and electronic displays, each with unique texture and contrast properties that complicate recognition. Traditional Optical Braille Recognition (OBR) methods rely on handcrafted feature extraction and classical machine learning techniques, which struggle to adapt to real- world variations. These approaches, based on techniques like thresholding, edge detection, and morphological processing, work well in controlled settings but often fail in scenarios with skewed orientations, embossed dot inconsistencies, and back- ground noise. Their dependence on predefined features makes them unsuitable for generalizing across different conditions, limiting their scalability. Recent advancements in Vision-Language Models (VLMs) have introduced more effective methods for multimodal learn- ing by aligning visual and textual representations. VLMs such as CLIP, FLAVA, and LLAVA leverage large-scale pretraining to develop robust cross-modal features, significantly improv- ing text recognition in complex environments. Our approach integrates Vision Transformers (ViTs) for image encoding and GPT-based transformers for text decoding, offering a flexible and scalable solution for Braille translation. Unlike CNNs, ViTs process images as sequences of patches, allow- ing them to capture long-range dependencies and intricate spatial relationships within Braille patterns. Additionally, we incorporate a fusion layer to enhance the alignment between visual features and textual output, improving recognition ac- curacy. By employing a teacher-forcing training strategy and causal language modeling (CLM) loss, our model achieves superior performance, surpassing conventional OBR methods in handling challenging conditions such as varied lighting, distortions, and occlusions.

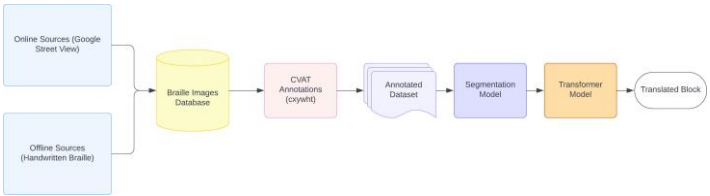

Fig. 1. Workflow

Related Work

Braille recognition has been an active area of research for decades, with early methods primarily relying on traditional image processing techniques such as edge detection, con- tour extraction, and morphological operations. These classical approaches focus on isolating Braille dot patterns from the background by enhancing contrast and filtering noise. While these methods can achieve reasonable accuracy in controlled settings, they struggle to generalize to real-world datasets where variations in lighting, texture, and embossing styles introduce significant challenges. Factors such as inconsistent dot spacing, blurred edges, and occlusions further hinder the effectiveness of these handcrafted techniques, making them less viable for large-scale deployment in natural scene Braille recognition tasks.

With the rise of deep learning, researchers have explored more sophisticated architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for Braille translation. CNNs are particularly effective in feature extraction, identifying spatial patterns within Braille images, while RNNs help in sequence modeling for character recognition. However, these models often require extensive labeled datasets, which are difficult to obtain for Braille due to its specialized nature. Additionally, their performance tends to degrade in the presence of environmental noise, distortions, or unseen variations in Braille formatting. Recent advance- ments in Vision-Language Models (VLMs), such as CLIP and LLAVA, have significantly improved image-text alignment by leveraging transformer-based architectures. These models extract deep, hierarchical visual features and map them to corresponding textual representations, enabling a more flexible and context-aware approach to Braille recognition. Our work builds on these innovations by integrating VLMs into the Braille translation pipeline, improving generalization across diverse datasets and enhancing recognition accuracy even in challenging real-world scenarios.

Proposed Model Architecture

Vision Transformer for Image Encoding

Our model employs the LLAMA 3.2 11B Vision model as a pretrained backbone for image encoding, leveraging its extensive visual understanding capabilities to improve Braille recognition. The Vision Transformer (ViT) architecture pro- cesses images by first segmenting them into fixed-size patches, converting them into a sequence of embedded tokens. Each token is then enriched with positional encodings, ensuring that the model retains spatial awareness while analyzing the input. The self-attention mechanisms within ViTs allow for the extraction of deep hierarchical features, capturing intricate relationships between Braille dot patterns and their spatial organization.

Unlike Convolutional Neural Networks (CNNs), which rely on local receptive fields, ViTs can model long-range depen- dencies across the entire image. This ability makes them particularly well-suited for Braille text recognition, where dot spacing and positioning play a crucial role in character interpretation. By analyzing the global structure of Braille text rather than focusing on small, isolated features, ViTs provide a more comprehensive understanding of the patterns present in the image. This results in improved robustness against distortions, uneven lighting conditions, and occlusions, ultimately enhancing the model's ability to accurately translate Braille from diverse and challenging real-world environments.

Fusion Layer for Multimodal Integration

The fusion layer serves as a crucial bridge between visual and textual representations, enabling seamless integration of image-derived features with linguistic structures. This layer is responsible for mapping high-dimensional embeddings gener- ated by the Vision Transformer (ViT) into a shared multimodal space, ensuring effective alignment between the extracted visual patterns and their corresponding textual outputs. By refining and transforming ViT embeddings into a structured representation that the text decoder can interpret, the fusion layer enhances the model's ability to generate accurate and contextually relevant Braille translations.

A key advantage of this module is its lightweight design, which allows for efficient fine-tuning without imposing ex- cessive computational overhead. Unlike traditional multimodal learning approaches that require complex fusion mechanisms or additional attention layers, our fusion layer streamlines the alignment process while preserving critical feature informa- tion. This enables the model to maintain high performance and adaptability across various Braille formats, even when faced with challenges such as noisy backgrounds, irregular dot spacing, or distortions in natural scene images. By optimizing feature integration at this stage, we improve the overall robust- ness and efficiency of the Braille translation pipeline, making it more scalable for real-world applications.

GPT-Based Transformer for Text Decoding

For text decoding, we utilize a GPT-2 transformer to con- vert visual representations into meaningful textual outputs, leveraging its powerful language modeling capabilities to improve Braille transcription accuracy. The model is trained using a teacher-forcing strategy, where ground-truth labels are provided as inputs during training rather than relying solely on previous model predictions. This method stabilizes the learning process, preventing the model from diverging due to compounding errors and ensuring faster convergence. By rein- forcing correct sequences during training, the decoder learns to generate highly accurate and structured Braille translations, even when faced with ambiguous or noisy inputs.

During inference, we adopt a prompt-based approach in- spired by GPT-3, allowing the model to generate coherent Braille translations in an autoregressive manner. Instead of simply mapping detected Braille characters to text, the decoder leverages contextual understanding to refine and structure the output

based on prior knowledge. This method significantly enhances the fluency and accuracy of Braille transcription, ensuring that translations maintain logical coherence even in challenging conditions. By combining teacher-forcing training with autoregressive generation, our text decoding approach bridges the gap between raw visual patterns and high-quality, natural-language Braille translations, making the system more effective in real-world applications.

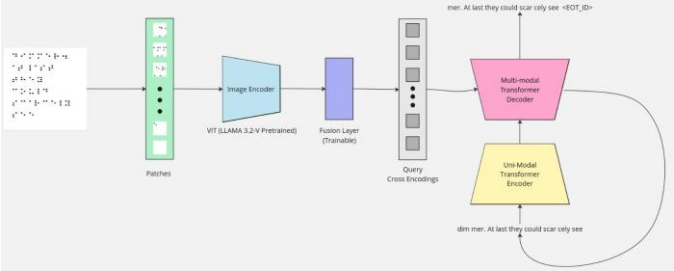

Fig. 2. Model Architecture

Training Strategy

Dataset Preparation and Augmentation

Our dataset consists of a diverse collection of images sourced from multiple origins, including Google Street View, Braille textbooks, and manually transcribed samples created using a Braille slate. These sources ensure a wide range of Braille representations, capturing variations in font size, embossing styles, materials, and environmental conditions. Google Street View images provide real-world examples of Braille signage found in public spaces, such as transportation hubs, government buildings, and commercial establishments. Braille textbooks offer structured, high-quality samples that serve as a benchmark for accurately formatted Braille text, while manually transcribed Braille samples help in creating controlled datasets that simulate various real-world conditions. By incorporating data from these different sources, we ensure that our model is exposed to a diverse set of training examples, improving its ability to generalize to unseen scenarios.

To further enhance the robustness of our model, we employ a comprehensive set of data augmentation techniques designed to mimic real-world variations and distortions. These augmentations include:

- **Random rotation and flipping:** These transformations introduce perspective variations, helping the model rec- ognize Braille text from different viewing angles and orientations.
- **Brightness and contrast adjustments:** By simulating different lighting conditions, such as low-light environ- ments, shadows, and reflections, we improve the model's ability to handle inconsistencies in illumination.
- **Synthetic noise injection:** To enhance resilience against environmental distortions, we add various types of noise, such as Gaussian blur, salt-and-pepper noise, and motion blur, making the model more robust to real-world image imperfections.
- **Style transfer techniques:** We apply style augmen- tation methods to generate additional training samples that resemble different textures, surfaces, and embossing materials used in Braille printing. This helps the model adapt to variations in Braille presentation across different physical mediums.

These augmentation strategies not only expand the dataset but also significantly improve the generalization capability of our model, enabling it to recognize Braille text across a wide range of real-world conditions, including complex backgrounds, occlusions, and distortions.

Loss Function and Optimization

We use Causal Language Modeling (CLM) loss as the pri- mary objective for training, optimizing the model for autore- gressive text generation by predicting each token sequentially based on previously generated outputs. CLM loss ensures that the model learns to generate coherent and grammati-cally accurate Braille translations while maintaining contextual consistency. This loss function is particularly effective for transformer-based architectures, as it enables the model to develop a deeper understanding of sequential dependencies in text.

To optimize training efficiency, we employ the Adam opti- mizer with a learning rate of le^{-4}, chosen for its adaptive learning capabilities and ability to handle sparse gradients effectively. Additionally, a cosine learning rate scheduler is applied over 50 epochs to gradually decrease the learning rate, preventing abrupt weight updates and ensuring stable convergence. Training is conducted on an A40 GPU, which provides the necessary computational power to handle large- scale transformer-based architectures efficiently. We utilize a batch size of 16, striking a balance between memory efficiency and gradient stability, allowing the model to learn effectively from diverse Braille samples while maintaining fast conver- gence. This setup ensures that our model is well-optimized for both accuracy and computational efficiency, enabling robust performance in real-world Braille translation tasks.

```
root@9e045d6b7e40:/workspace# python3 ./inference.py
🦥Unsloth: Will patch your computer to enable 2x faster free finetuning.
🦥Unsloth Zoo will now patch everything to make training faster!
Standard import failed for UnslothORPOTrainer: No module named 'UnslothORPOTrainer'. Using tempfile instead!
==((====))==  Unsloth 2025.3.16: Fast Mllama patching. Transformers: 4.49.0.
   \\   /|    NVIDIA A40. Num GPUs = 1. Max memory: 44.448 GB. Platform: Linux.
O^O/ \_/ \    Torch: 2.6.0+cu124. CUDA: 8.6. CUDA Toolkit: 12.4. Triton: 3.2.0
\        /    Bfloat16 = TRUE. FA [Xformers = 0.0.29.post3. FA2 = False]
 "-____-"     Free license: http://github.com/unslothai/unsloth
Unsloth: Fast downloading is enabled - ignore downloading bars which are red colored!
```

Fig. 3. Model Specifications

Experimental Results

A. Performance Evaluation

To validate our approach, we compare our Braille translation model against widely used object detection and text recog- nition frameworks, including Faster R-CNN, YOLOv5, and YOLOv11. These models serve as benchmarks for evaluating detection accuracy, robustness, and adaptability to real-world scenarios. Our assessment focuses on key performance metrics such as mean Average Precision (mAP) and classification accuracy, particularly under challenging conditions like occlu- sions and varying lighting environments.

Experimental results show that our model consistently out- performs these conventional methods, achieving superior accu- racy and generalization. While Faster R-CNN provides high detection precision, it suffers from slower inference speeds.

Similarly, YOLOv5 and YOLOv11 offer faster processing but struggle with challenging lighting and occlusions. In contrast, our transformer-based approach effectively captures long- range dependencies and contextual features, allowing for more precise and adaptable Braille recognition. This demonstrates the robustness of our model in handling complex real-world conditions, making it a more reliable solution for automated Braille translation.

Model	Accuracy	CLM Loss	BBox Loss
Faster R-CNN	**79.56%**	**0.0481**	**0.1020**
YOLOv11n YOLOv5	**84.74%**	**0.804**	**0.92**
Braille Spotting Model (LLAMA+VIT)	**85.51%**	**0.432**	**0.632**
	92.82%	**0.699**	**0.830**

TABLE I
PERFORMANCE METRICS OF BRAILLE DETECTION MODELS

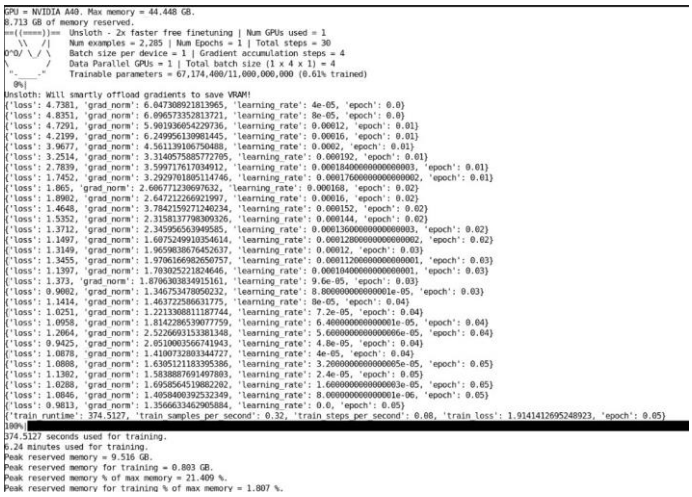

Fig. 4. Training Metrics of Model

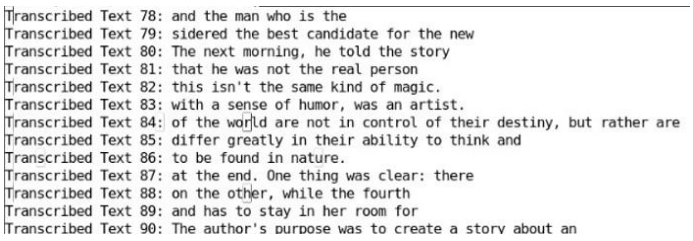

Fig. 5. Model Inference on Image Samples

Conclusion

Our transformer-based Braille translation model demon- strates significant improvements over traditional Optical Braille Recognition (OBR) techniques, overcoming many of the limitations associated with classical machine learning ap- proaches. By integrating Vision Transformers for high-quality image encoding, GPT-based text decoding for accurate lan- guage modeling, and a specialized fusion layer for multimodal feature alignment, our approach achieves superior accuracy, scalability, and robustness.

While traditional object detection models like Faster R- CNN achieve an accuracy of 79.56%, our approach sur- passes this significantly, with the Braille Spotting Model (LLAMA+VIT) achieving a remarkable 92.82%. The YOLO- based models demonstrate incremental improvements, with YOLOv5 attaining 85.51% accuracy. However, our model not only achieves the highest accuracy but also maintains a balanced performance in CLM Loss (0.699) and BBox Loss (0.830), highlighting its robustness in both character local- ization and language modeling. This substantial improvement underscores the advantage of integrating Vision Transformers for spatial feature extraction and GPT-based decoding for linguistic coherence.

Future Work

Looking ahead, future work will focus on optimizing the model for mobile deployment, ensuring efficient performance on resource-constrained devices without compromising accu- racy. Additionally, we aim to extend our system to support multilingual Braille variants, enabling broader accessibility for visually impaired individuals across different languages and regions. Another key area of exploration involves incorporat- ing self-supervised learning techniques to reduce reliance on manually labeled datasets, allowing the model to learn from large-scale, unlabeled Braille corpora. These advancements will further enhance the adaptability, efficiency, and real-world applicability of our Braille translation system.

REFERENCES :

[1] A. Radford et al., ”Learning Transferable Visual Models from Natural Language Supervision,” in Proc. ICML, 2021.
[2] M. Singh et al., ”FLAVA: A Foundational Vision-Language Model,” arXiv preprint arXiv:2103.00020, 2021.
[3] Z. Liu et al., ”LLAVA: Large Language and Vision Assistant,” arXiv preprint arXiv:2405.17247, 2024.
[4] Touvron, H. et al. (2023). LLaMA: Open and Efficient Foundation Language Models. arXiv preprint arXiv:2302.13971.
[5] F. Bordes, R. Y. Pang, A. Ajay, A. C. Li, A. Bardes, S. Petryk, O. Man˜as, Z. Lin, A. Mahmoud, B. Jayaraman, et al., ”An Introduction to Vision- Language Modeling,” in *arXiv preprint arXiv:2405.17247*, 2024.
[6] Dosovitskiy, A. et al. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR)
[7] P. Rodriguez, M. Fernandez, and D. Lopez, ”Data Augmentation Tech- niques for Enhancing Braille Recognition Models,” *Pattern Recognition*, vol. 145, pp. 109876, 2024.