15.773 Hands-on Deep Learning Final Report

Capturing Wonder:
Bridging Imagery & Text in Children's Books

B11: Sean Huang, Crystal Li, Laura Li, Kevin Sheng

MIT Sloan School of Management
Massachusetts Institute of Technology
March 8, 2024

Massachusetts
Institute of
Technology

1 Problem Description

Image captioning is the process of generating a textual description for an image, which involves both understanding the visual elements of the image and articulating them in a coherent sentence. This task represents an interdisciplinary challenge combining computer vision and natural language processing.

The core of this task lies in the intricacy of accurately interpreting the images, most of which are cartoon illustrations since this is a dataset comprised of children's book covers. The model should be able to recognize objects, their attributes, and actions. Moreover, the system must comprehend the scene's context and the nuanced relationships among its elements. This requires a sophisticated understanding of visual semantics that goes beyond mere object recognition, demanding a deep integration of visual and linguistic processing abilities, which, fortunately, transformers possess.

Specifically, within the realm of children's literature, the application of image captioning to book covers represents an innovative approach to bridging the gap between visual imagery and textual understanding. This technology not only aids in the comprehension of visual information but also enriches the reading experience for young readers.

The significance of implementing image captioning on children's book covers lies in its potential to make literature more inclusive and engaging for all readers, especially those with visual impairments or learning disabilities. By providing descriptive captions for the vibrant and often intricate illustrations found on these covers, we can offer a more welcoming experience for young readers and instill in them a love of learning, as early engagement with literature is instrumental in developing literacy and imagination. In addition, providing standardized descriptions of book covers would help the classification systems in libraries, bookstores, and online websites to generate more accurate labels by providing more context and hints of content beyond book titles.

However, the application of image captioning technology in this context is not without its challenges. The complexity of accurately and creatively describing artistic images in a manner that is both engaging and understandable to children requires sophisticated AI models and a deep understanding of child psychology and language development. Moreover, ensuring the cultural and contextual relevance of captions demands sensitivity and adaptability in the algorithmic approach.

2 Approach

To approach the image captioning problem for children's books based on their cover images, we built a structured pipeline.

2.1 Baseline Model

Our dataset comprises of children's book cover images along with their manually annotated labels (captions). To transform the visual data into textual descriptions, a Vision-Language Model (VLM) was deployed; VLMs are designed to understand and generate content that bridges the gap between visual information and natural language, making them ideal for tasks that require a nuanced understanding of both images and text. This model was trained on large datasets containing paired image and text data, including MSCOCO, SBU Captions, Visual Genome, VQAv2, GQA, and a few internal datasets, which enables the model to learn complex patterns and relationships between visual elements and their linguistic descriptions. By incorporating

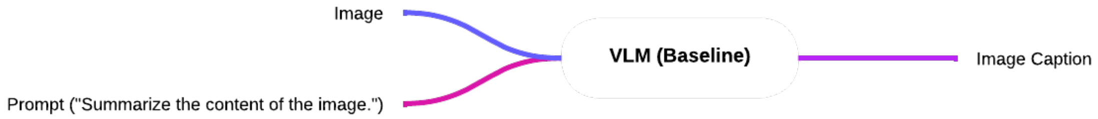

Figure 1: VLM Model Flow

both Convolutional Neural Networks (CNN) for image processing and transformer-based architectures for language understanding and generation, VLMs can effectively interpret visual content, generate descriptive text, and even answer questions regarding the content of the image.

The CNN component acts as a feature extractor for the visual input, transforming an input image into a set of feature vectors that represent different aspects of the image. These feature vectors are then combined with the textual data, which is processed by the Transformer component. Thus, we picked this model to apply to our dataset for captioning children's book covers, which brought together the visual storytelling in the book illustrations with the narrative captured in the titles.

2.1.1 Model Interaction

The processing of each image-text pair was facilitated by the VLMprocessor, which prepared the inputs for the model. This preparation involved converting prompts (intended captions) and images into a format suitable for the model, leveraging PyTorch tensors for efficient computation.

2.1.2 Inference and Decoding

Using PyTorch's inference_mode context manager ensured that computations were performed more efficiently by disabling gradient calculation, which is essential for speeding up the prediction process. The VLM.generate method was invoked with specific parameters tailored to our needs—disabling sampling to ensure consistency in outputs, enabling cache use for efficiency, setting the maximum token limit for generation, and specifying token IDs for end-of-sequence and padding to manage the sequence lengths effectively.

The generated output tokens were then processed to remove the initial prompt portion, ensuring that only the newly generated text—our model-generated captions—was retained.

2.1.3 Output Decoding

Then, the raw output tokens were decoded into English using the VLMprocessor.batch_decode method for interpretation and assessment.

2.1.4 Text Embedding

To evaluate the model output, we aimed to quantitatively assess the relevance and accuracy of the generated captions for children's book covers compared with target labels. To achieve this, we utilized BERT (Bidirectional Encoder Representations from Transformers) embeddings and the computation of cosine similarity between the embeddings. We chose the bert-base-uncased model for its robustness and general applicability to a wide range of text data. The BertTokenizer and BertModel from the transformers library were used to tokenize the input texts and generate embeddings. This setup ensured the text inputs were appropriately processed to reflect the contexts necessary for an accurate evaluation.

For each book cover in our dataset, we prepared a pair of texts: the original caption and the generated caption. These texts were then tokenized and passed through the BERT model to obtain embeddings. The embedding process involved encoding the text into tensors, running these through the BERT model, and taking the mean of the last hidden state outputs to produce a single embedding vector for each text. This method allowed us to capture the semantic meaning of both the original and generated captions in a high-dimensional vector space.

With embeddings for both the original and generated captions in hand, we computed the cosine similarity between them. The cosine similarity metric, ranging from -1 (completely different) to 1 (exactly the same), provides a quantitative measure of the similarity between two vectors in terms of their orientation in the vector space, effectively measuring how closely the generated captions match the semantic content of the original captions.

2.1.5 Cosine Similarity Calculation

To compute the overall similarity for each pair of captions, we concatenated the embeddings and used the cosine_similarity function from sklearn.metrics.pairwise to generate a similarity matrix. This matrix contained the cosine similarities between all pairs of embeddings, enabling us to assess not just the direct similarity between the original and generated captions but also the internal consistency of the embeddings.

2.1.6 Average Similarity Score

To come up with a single, interpretable metric, we first eliminated self-similarities by setting the diagonal of the matrix to 0. This adjustment prevented the inflation of similarity scores by the perfect match of each caption with itself. We then calculated the average similarity across all entries in the matrix, discounting the diagonal. This average provided a concise measure of how closely, on average, the generated caption's semantic content aligned with that of the original caption.

2.2 GPT-4-Vision-Preview

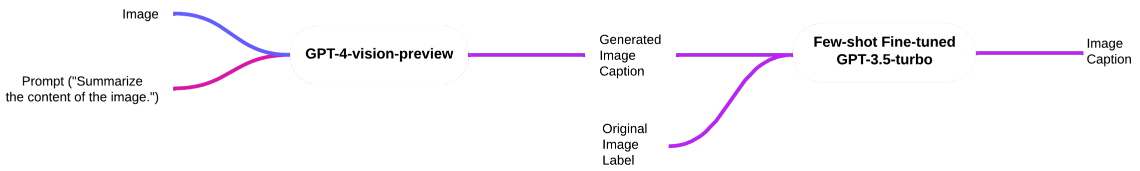

Figure 2: GPT-4-Vision-Preview Model Flow

Next, we leveraged the GPT-4-Vision-Preview model. The GPT-4-Vision-Preview model combines the robust NLP capabilities of GPT-4 with cutting-edge computer vision technology. This integration enables the model to not only understand and generate human-like text but also to interpret and analyze visual content with a high degree of sophistication. While specific details about the datasets used to train GPT-4-Vision-Preview are proprietary to OpenAI, it is known that GPT-4 and similar models are typically trained on diverse and expansive datasets. These datasets include a wide range of internet text and, for vision capabilities, likely comprise vast collections of images and their associated textual descriptions or metadata. This training enables the model to learn complex patterns and relationships between text and visual content.

The GPT-4-Vision-Preview model also operates by processing both textual prompts and visual inputs. When provided with an image, the model can generate descriptive text, answer questions about the visual content, or create text-based content that relates to the image. The model achieves this by employing a transformer-based architecture, with its self-attention mechanism that allows it to weigh the importance of different parts of the input data—whether textual or visual—when generating its output.

2.2.1 Model and API Initialization

We began with the initialization of the OpenAI client, employing a specific API key to authenticate and gain access to the GPT-4-Vision-Preview model.

2.2.2 Image Preparation and Encoding

To prepare the images for analysis, each image within our dataset was encoded into base64 format—a compact, text-based representation of the binary image data. This encoding enabled the integration of visual content into the API requests, allowing for efficient transmission and processing.

2.2.3 Generation of Textual Summaries

To generate summaries, we crafted prompts instructing the model to describe the content of each image for children. These prompts, along with the base64-encoded images, were sent to the GPT-4-Vision-Preview model via the OpenAI client. The model's responses were then collected and stored in a dictionary, mapping each file name to its corresponding generated summary.
Lastly, we used the same evaluation method—cosine similarity score—as we described above for the baseline model.

2.3 Few-shot GPT-4-Vision-Preview

In refining our approach to generate better captions for children's book covers, we used the few-shot fine-tuning with the GPT-3.5-Turbo model.

2.3.1 Training Dataset

We started with the preparation of a training dataset, which comprised of 10 examples, each containing 3 elements: 1) a "persona" for the system 2) a user input, which was the output of the GPT-4-Vision-Preview model, and 3) an assistant content, which was the true label (caption) obtained from the dataset.

2.3.2 Fine-Tuning

With our training data prepared, we proceeded to upload it to the OpenAI platform, specifying its purpose for fine-tuning. A fine-tuning job was then created with the prompt "I have some images of children's book covers. I used a deep-learning model to generate captions for those book covers. Now I am going to give you those captions and I want you to transform them into the style used in your fine-tuning." In return, we received the transformed captions that embodied the desired stylistic qualities. Again, we used the same evaluation method—cosine similarity score—as we described above for the baseline model.

3 Results

Cosine Similarity Score	VLM Baseline	GPT-4-Vision-Preview	Few-shot Fine-tuned GPT
Min Score	0.6113	0.4761	0.4305
Max Score	0.8802	0.8960	0.9434
Average Score	0.7747	0.8067	0.7987

Table 1: Performance metrics of different models

Comparing the performance of the VLM used as the baseline against the GPT-4-Vision-Preview model based on their cosine similarity scores with original captions, the VLM baseline model achieves a minimum cosine similarity score of 0.61, a maximum of 0.88, and an average score of 0.77. On the other hand, the GPT-4-Vision-Preview model shows a broader range of performance with a minimum score of 0.48, a maximum of 0.90, and an improved average score of 0.81. The broader range of scores in the GPT-4-Vision-Preview model, particularly the lower minimum score, suggests that while it may occasionally produce results less aligned with the original captions compared to the baseline. The GPT-4-Vision-Preview model's capacity for generating captions that are highly similar to the originals is greater, as indicated by both the higher maximum and average scores. This enhanced performance could be attributed to several factors inherent in the GPT-4 architecture.

GPT-4 likely has a better understanding of complex visual-text relationships due to its larger training dataset and more sophisticated architecture. The architecture of GPT-4 allows for a more nuanced incorporation of context, which can lead to more accurate and relevant caption generation by effectively synthesizing information from both the visual and textual domains.

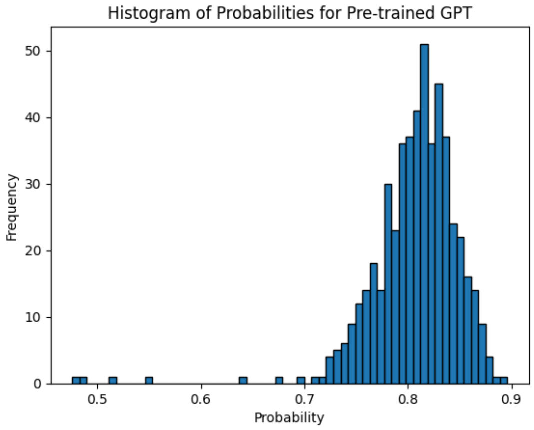

Figure 3: Pre-trained GPT

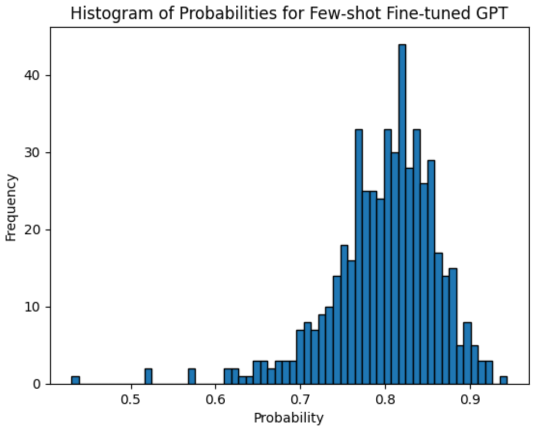

Figure 4: Fine-tuned GPT

Finally, comparing the results from GPT4-Vision-Preview and the fine-tuned GPT3.5-Turbo model, as shown in the table and the histograms, we can see that the fine-tuned model has a more dispersed score distribution. Specifically, the distribution of scores covers a wider range: the highest score becomes even better while the lowest score becomes worse, with the mean staying approximately the same. A possible explanation for this phenomenon is that the output quality tends to go more extreme as one major task of the fine-tuned model is shortening and concentrating the text. This uneven magnification effect highlights an unexpected outcome from the application of more advanced models.

3.1 Examples of Generated Captions

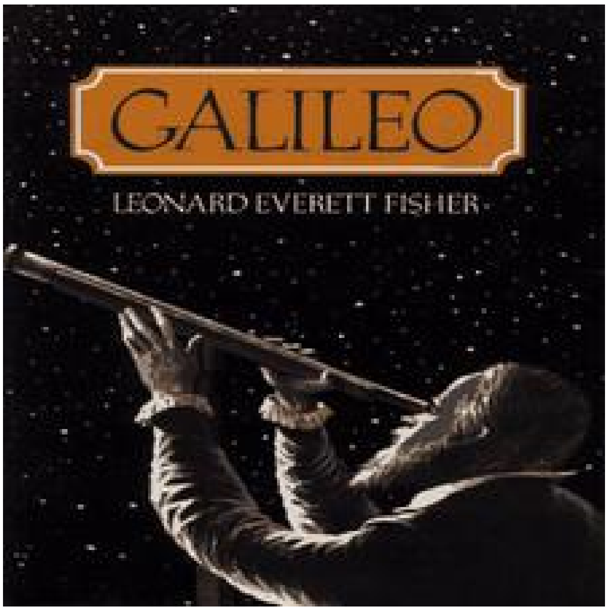

Figure 5: Example 1

The original label from the dataset for the example of children's book cover, presented in Figure 5, is a "drawing of a man with a telescope looking at stars with the title 'GALILEO'."

The generated caption from the VLM baseline model for Figure 5 is "A man sits on a bed, holding a gun and aiming it at a target. He seems to be in a video game, possibly a first-person shooter, as he is dressed in a suit and holding a gun. The setting is a dark room, with the man as the primary subject of the image."

The generated caption from the GPT-4-Vision-Preview model for Figure 5 is "The image shows a person looking up at the stars in the night sky through a long telescope. The background is dark with lots of little white dots that look like stars. Above the person, there's a big word "GALILEO" in gold letters, which is the name of a very famous person who studied the stars a long time ago. It looks like the cover of a book that tells the story of Galileo."

The generated caption from the Few-shot Fine-tuned model is "Whispers of the Cosmos: The Stargazing Tales of Galileo."

A comparison between generated captions from the three models against the original label reveals their differences in interpretation and accuracy.

- VLM Baseline Model: The caption generated by the baseline model does not align with the original context, suggesting a scene unrelated to astronomy or historical figures. It inaccurately describes the scenario and misinterprets the visual cues in the image. This discrepancy could stem from the model's training data, since it was mainly trained on real-life images, as opposed to cartoons. The model's general approach to visual recognition might not be specialized in historical or educational content as well.

- GPT-4-Vision-Preview Model: This model's generated caption closely aligns with the original label, accurately capturing the essence of the image by describing a person observing the stars with a telescope under the name "GALILEO." It demonstrates a nuanced understanding of the historical and educational context, likely benefiting from its advanced training on diverse datasets that include educational material. The detailed description and the mention of "Galileo" indicate a more refined processing of visual and textual information.
- Few-shot Fine-tuned Model: The caption from the fine-tuned model, offers a creative and contextually appropriate label that elegantly encapsulates the theme of the book cover. It suggests that fine-tuning the model with specific examples related to the task at hand can lead to highly relevant and thematic captions. This model is good at generating succinct captions that are both engaging and directly relevant to the content depicted.

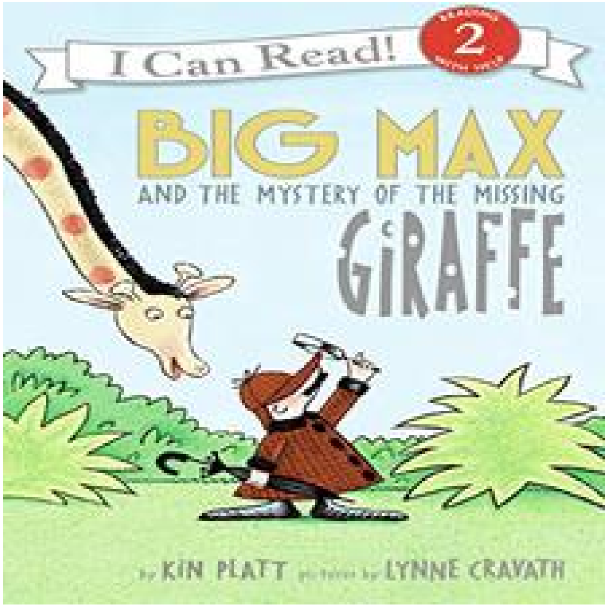
I Can Read!
2
BIG MAX
AND THE MYSTERY OF THE MISSING
GIRAFFE
by KIN PLATT pictures by LYNNE CRAVATH

Figure 6: Example 2

The original label from the dataset for the example of the children's book cover, presented in Figure 6, is a "drawing of a detective in brown clothes holding a magnifying glass in a jungle with a giraffe looming from the left side behind him and the caption 'Big Max and the mystery of the missing giraffe.'"

The generated caption from the VLM baseline model for Figure 6 is "A cartoon depicts a giraffe and a man sitting together in a grassy area. The giraffe is on the left, and the man is on the right. The scene is set in a park, with a potted plant in the background. The giraffe holds a book, possibly a children's book, in its mouth. The man appears to be reading the book, creating a playful and lighthearted atmosphere."

The generated caption from the GPT-4-Vision-Preview model for Figure 6 is "This is a picture of a book cover. It shows a drawing of a detective who looks like he's trying to solve a mystery. The detective, Big Max, has a magnifying glass and seems to be searching for clues. There's a tall giraffe peeping into the picture, which might be part of the mystery he's trying to solve. It looks like a fun story for kids to read about a detective on an adventure to find a missing giraffe."

The generated caption from the Few-shot Fine-tuned model is "Illustration of a giraffe and a detective named Big Max with the title 'Big Max and the Missing Giraffe.'"

The generated labels for Figure 6 once again demonstrate the differences of each model.

- VLM Baseline Model: The caption generated by the VLM baseline model is slightly inaccurate from the original label's narrative and detail. It interprets the scene as a giraffe and a man sitting together in a park, instead of a giraffe looming from the left side behind him. This interpretation misses the detective and mystery elements, reflecting a possible limitation in capturing the full context of the image.

- GPT-4-Vision-Preview Model: This model offers a more accurate interpretation, aligning closely with the original label's narrative. It identifies the key elements of the detective, the magnifying glass, the giraffe, and the mystery theme, packaging these details into a coherent summary that closely resembles the original description. It demonstrates a superior ability to interpret complex scenes and narratives.

- Few-shot Fine-tuned Model: This model generates a succinct caption that, while not as detailed as the original label, accurately captures the essential elements of the book cover: the giraffe, the detective Big Max, and the title. This model's output shows effective learning from the fine-tuning process, demonstrating an ability to generate focused and relevant captions.

4 Lessons Learned

Through this project, we gained valuable insights into the challenges and considerations involved in applying advanced language models to image captioning, particularly in the context of children's literature. One key lesson learned was the importance of domain-specific training data and fine-tuning techniques. While large pre-trained models like GPT-4 demonstrated impressive capabilities on general tasks, fine-tuning on a small set of relevant examples allowed the model to better capture the nuances and stylistic elements desired for engaging and age-appropriate captions.

Another significant lesson was the need for comprehensive evaluation methods that go beyond traditional metrics like cosine similarity. While such quantitative measures could provide a useful baseline, they may fail to capture the subjective aspects of caption quality, such as creativity, coherence, and suitability for the target audience. Incorporating human evaluation and feedback mechanisms could further enhance the assessment process and guide model improvements.

Furthermore, this project highlighted the importance of interdisciplinary collaboration. Combining expertise in computer vision, natural language processing, and child psychology enabled a more holistic approach to the problem, ensuring that the generated captions not only accurately described the visual elements but also aligned with the cognitive and developmental needs of young readers.

Lastly, we recognized the potential ethical implications of this technology. While image captioning for children's books can promote inclusivity and accessibility, it is crucial to consider issues such as cultural sensitivity, bias mitigation, and the impact on children's perception and learning. Responsible development and deployment of these AI systems require ongoing monitoring, evaluation, and stakeholder engagement to ensure positive outcomes for all users.