Text Summarization and Small Language Model Fine tuning

Clervilsson Christelle

Project EmotionAI

September 1, 2025

1 Introduction

This work presents an AI system that seeks to contribute to the emotional well-being of orphans with personalized visual narratives. The system begins with detecting a child's facial emotion from an input image through a fine-tuned deep learning model. In an attempt to make a contribution to personalization, it also extracts other attributes such as estimated age, sex, clothing color, and the presence of glasses. These traits, along with the recognized emotion, are translated into a personalized text prompt. A story generation model creates a short, motivational narrative most appropriate to the child's current emotional state, and an image generation model produces a corresponding image.

With emotion recognition, personalization, and visual story-telling, the system offers a nonverbal and interactive way of providing comfort and emotional support. The approach is designed to reach children with limited literacy levels, focusing on inclusion and emotional appeal.

2 Steps

1. Dataset Selection and Pretraitement
2. Emotion Detection Model
3. Story Generation
4. Image Generation
5. Evaluation
6. Conclusion

2.1 Dataset Selection and Preprocessing

The data set chosen for this work is FER2013 (Facial Expression Recognition 2013). It contains approximately 35,000 images, each consisting of 48x48 pixel grayscale faces. The data set includes seven emotions: *anger, disgust, fear, happiness, sadness, surprise,*

and neutral. The training set consists of 28,709 examples, while the public test set contains 3,589 examples. This organization helps reduce overfitting and ensures a balanced evaluation.

The data set is widely used in facial expression recognition research because it is small enough to allow fast experimentation, yet large enough to train deep learning models effectively. Moreover, its diversity in the conditions of the image makes it a good benchmark for testing robustness. For this reason, it is also commonly used for fine-tuning pre-trained models.

During preprocessing, the resizing of the images from 48×48 to 224×224 pixels was tested to match the input requirements of some architectures. However, this significantly increased computation time and was therefore not adopted as the final approach.

2.2 Emotion Detection Model

As shown in previous work [1], emotion detection using deep learning has gained significant attention. In this project, three different models were used to try to detect emotions in the picture

2.2.1 ResNet

ResNet (Residual Network) is a family of convolutional neural network (CNN) architectures specifically designed to overcome the vanishing gradient problem that occurs when training very deep models. This makes ResNet particularly well suited for tasks such as facial emotion recognition, where subtle features must be preserved and propagated through many layers.

ResNet18, the variant used in this project, is lightweight and efficient. Compared to older architectures such as VGG, it requires less memory, is faster to train, and performs well even with relatively small datasets, which fits the constraints of FER2013. The network was trained in ImageNet, which enables transfer learning by reusing the learned low-level visual features.

Pipeline:

- **Input:** Preprocessed face image (normalized, resized, converted to RGB).
- **Convolutional layers + residual blocks:** Extract hierarchical facial features while preserving fine details such as eyebrows and lips.
- **Global Average Pooling:** Reduces feature maps to a single vector representation.
- **Fully connected layer:** Output probabilities for the seven emotion classes.
- **Softmax activation:** Produces the final predicted emotion label.

To adapt ResNet18 to the FER2013 dataset, the first convolutional layer was adjusted to process three-channel RGB images instead of grayscale. Finally, the fully connected output layer was replaced with a new one that contains seven neurons, corresponding to the seven emotion categories. These modifications allowed ResNet18 to take advantage of transfer learning while being specifically tailored for facial emotion recognition.

2.2.2 DenseNet

DenseNet (Densely Connected Convolutional Network) is a convolutional neural network architecture where each layer is directly connected to every subsequent layer in a feed-forward manner. This dense connectivity pattern encourages the reuse of features and facilitates efficient gradient flow during training. As a result, each layer receives information from all preceding layers, which helps preserve fine-grained details that might otherwise be lost in deeper networks.

This property makes DenseNet particularly suitable for facial emotion recognition, since emotions often rely on subtle visual cues such as slight changes in the eyes, eyebrows, or mouth. By promoting feature reuse, the architecture ensures that small but important features are retained and leveraged throughout the network.

DenseNet also performs well on relatively small datasets like FER2013, but compared to ResNet, it tends to be slower during training due to its more complex connectivity structure.

2.2.3 Vision Transformer (ViT)

The Vision Transformer (ViT) adapts the transformer architecture, originally developed for natural language processing, to image classification tasks. Instead of applying convolutional filters, ViT divides an image into fixed-size patches, flattens them, and processes them as a sequence of tokens, similar to how words are processed in a sentence.

ViT is particularly interesting for facial emotion recognition because it can capture long-range dependencies across different regions of the face, allowing the model to analyze both global context and fine-grained local details. However, ViT typically requires very large datasets to achieve strong performance. In addition, ViT is computationally more expensive, both in terms of memory usage and training time, compared to CNN-based models such as ResNet and DenseNet.

In our experiments, we attempted to train a ViT model, but due to limited computational resources and the relatively small size of the FER2013 dataset, the results were not satisfactory. After more than 8 hours of training, we were only able to complete two epochs, reaching an accuracy of 47%. Given the training time, lack of sufficient data, and limited resources, the experiment was stopped. It was therefore not further evaluated, as it was clear that the performance would remain below that of ResNet and DenseNet.

Model	**Key Idea**	**Params (M)**	**Pros**	**Cons**
ResNet18	Skip connections	11.7	Fast, reliable, stable	Misses fine details
DenseNet121	Dense feature reuse	8.0	Efficient, less overfitting, good details	Slower inference
ViT-B/16	Self-attention	86	Strong with big data, multimodal-ready	Heavy compute, needs large data*

Table 1: Compact comparison of ResNet18, DenseNet121, and ViT-B/16 for Emotion Recognition.

2.2.4 Qualitative Example

For the same input image, ResNet18 predicted the emotion as **neutral**, while DenseNet121 and the fine-tuned ResNet18 predicted **angry**. This illustrates how different architectures may cater to distinct features of the same facial expression, leading to divergent predictions.

Model	Training Loss	Training Accuracy
ResNet18	**0.0144**	**99.56%**
ResNet18 (tuned)	0.0213	99.37%
DenseNet121	0.0162	99.51%

Table 2: Training performance comparison between ResNet18 and DenseNet121.

Model	Test Loss	Test Accuracy
ResNet18	3.1125	59.78%
ResNet18 (tuned)	3.2110	60.21%
DenseNet121	**2.3457**	**65.10%**

Table 3: Test performance comparison between ResNet18 and DenseNet121.

2.3 Story Generation

Several strategies were explored to generate therapeutic stories based on detected emotions. The simplest approach was to associate each emotion with a fixed story. However, this method lacked personalization and diversity, as the same story was always produced for a given emotion.

To improve personalization, we incorporated traits extracted directly from the child's image into the story prompt. The attributes considered included estimated age, sex, presence of glasses, and clothing color. Based on estimated age, three categories were defined: *young child*, *teenager*, and *young adult*. The `DeepFace` framework was used for the estimated age and gender, while the YOLO (You Only Look Once) algorithm was used to detect elements such as glasses.

Although this strategy allowed for more detailed prompt construction, the results generated were often unsatisfactory. In many cases, the language model echoed the input prompt rather than producing a coherent continuation. This issue was largely due to excessive prompt length and the limited number of output tokens, which restricted the model's ability to generate consistent narratives.

The most effective approach was to design a set of tailored prompts for each emotion, complemented by a story-based template. Each prompt served as the beginning of a narrative, which the language model (LLM) would then complete. This method ensured variability: even when the same emotion was detected, the generated story differed due to the stochastic nature of text generation.

The trained emotion recognition model produced an integer output between 0 and 6, corresponding to the labels FER2013 data set. Then a mapping was applied to interpret these outputs as one of the seven emotion categories:

0: Angry

1: Disgust

2: Fear

3: Happy

4: Sad

5: Surprise

6: Neutral

2.3.1 Large Language Models

Pythia-410M. Pythia-410M is a lightweight large language model designed primarily for educational and research purposes. It is significantly smaller than modern LLMs, which makes it well suited for efficient inference in constrained environments. However, due to its limited size, the generated stories are often shorter and less coherent than those from larger models.

OPT-1.3B. The Open Pre-trained Transformer (OPT-1.3B) is a decoder-only LLM that is similar in architecture to GPT-3. With 1.3 billion parameters, it represents a mid-sized model. This larger capacity allows it to generate more fluent, coherent, and context-aware stories, especially over longer passages.

Phi-3 Mini. Phi-3 Mini, developed by Microsoft, is a small yet high-performing decoder-only Transformer model. The version used in this work contains 3.8 billion parameters, offering a good trade-off between efficiency and generation quality. Its relatively compact size allows it to run in lightweight environments such as Google Colab while still producing coherent narratives.

Model Test. To illustrate differences between the models, the following experiment was conducted. The detected emotion was *Neutral*, and the prompt provided to the LLMs was:

"A child with big emotions learns that anger transforms into strength when..."

The generated images corresponding to these prompts are shown in (Fig. 1), (Fig. 2), and (Fig. 3).

Model	Parameters	GPU Memory (approx.)	Strengths / Weaknes
Pythia-410M	0.41B	¡1 GB	Fast, light, but less coherent
OPT-1.3B	1.3B	3–4 GB	More fluent than Pythia, but
Phi-3 Mini	3.8B	6–8 GB	Balanced: strong coherence, empathetic

Table 4: Comparison of candidate LLMs for story generation.

2.4 Image Generation

After generating the therapeutic story, image generation models were employed to create illustrations corresponding to the narrative. Most of the models used were based on the Stable Diffusion architecture. To make the images more engaging and relatable for children, we chose to generate illustrations in anime style.

Different hyperparameters were explored to optimize the quality of the generated images. In particular, the number of inference steps was varied in order to compare results and evaluate how this parameter influenced the fidelity of the generated image to the input story prompt.

2.4.1 Image Generation Experiments

Two models were chosen to generate the images:

Figure 1: Pythia-410M : A child with big emotions learns that anger transforms into strength when given. He learns that the only person who can take away your anger is God, whose mind you do not have. When it's time to express your anger, this is a very short process (as we've seen, anger is usually expressed verbally)When it's time to express your anger, this is a very short process (as we've seen, anger is usually expressed verbally).

Figure 2: OPT-1.3B : Alex discovered that underneath anger is always a need for love or understanding. When he gave himself the compassion he was seeking, the anger melted like snow in sunshine. Alex felt heard, valued, and at peace.

Figure 3: Phi-3 Mini: they're calm and in control. Using principles from the document, devise a step-by-step strategy that the child could use to navigate situations where their emotions might escalate. Focus on the steps that integrate self-awareness and self-regulation, and explain the benefits of these techniques for the child's overall emotional well-being.

Counterfeit-V2.5 This is a specialized image generation model derived from Stable Diffusion. It was fine-tuned to generate images in anime style, making it particularly suitable for creating child-friendly and comforting illustrations.

We experimented with four different configurations of `num_inference_steps` and `guidance_scale` using the same input story in order to observe the differences in the generated outputs:

- num_inference_steps = 25, guidance_scale = 7.5 (Fig. 5)
- num_inference_steps = 25, guidance_scale = 15 (Fig. 6)
- num_inference_steps = 50, guidance_scale = 7.5 (Fig. 4)
- num_inference_steps = 70, guidance_scale = 7.5 (Fig. 7)

The results indicate that the number of inference steps should be neither too low nor too high in order to achieve optimal coherence between the generated image and the intended story.

Waifu Diffusion Waifu Diffusion is a latent text-to-image diffusion model fine-tuned on large-scale anime and manga-style datasets. It is derived from Stable Diffusion but adapted to generate high-quality, stylized anime images instead of photorealistic outputs. This specialization makes it particularly suitable for projects involving children, as the generated visuals are colorful, emotionally expressive, and reminiscent of the artistic style found in children's books or animated media.

Compared to Counterfeit-V2.5, Waifu Diffusion produces more consistent results in the anime style, with smoother textures and simplified yet expressive features.

Figure 4: 50 steps, guidance 7.5

Figure 5: 25 steps, guidance 7.5

Figure 6: 25 steps, guidance 15

Figure 7: 70 steps, guidance 7.5

Additional experiments were conducted by varying `num_inference_steps` and `guidance_scale` while keeping the input story fixed, in order to systematically examine the resulting differences in the generated outputs.

- `num_inference_steps = 25`, `guidance_scale = 7.5` (Fig. 8)
- `num_inference_steps = 25`, `guidance_scale = 15` (Fig. 10)
- `num_inference_steps = 50`, `guidance_scale = 7.5` (Fig. 9)

The issue with these two selected models is that they primarily take as input the characteristics you want to see in the image, such as appearance or style, but pay less attention to the narrative of the story itself. The goal was to explore a model capable of focusing first on the story and then generating the corresponding image, which could

Figure 8: Waifu Diffusion result with 25 inference steps and guidance scale 7.5.

Figure 9: Waifu Diffusion result with 50 inference steps and guidance scale 7.5.

Figure 10: Waifu Diffusion result with 25 inference steps and guidance scale 15.

later be transformed into anime style. However, at the moment, such a model requires significantly more computation time, making it less practical for our current workflow.

Negative Prompt Negative prompts were also tested to evaluate their effectiveness. These prompts are used to specify elements that should be avoided in the generated images. Since these models often tend to produce unintended or excessive features, negative prompts help guide the model by explicitly stating what should not appear, improving overall control over the output.

2.4.2 Evaluation

Evaluating the generated stories posed a challenge, as traditional natural language generation metrics such as ROUGE, BLEU, or BERTScore require reference texts for comparison, which were not available in our case. Since our goal was not to reproduce existing

Figure 11: Result of Counterfeit-V2.5 with negative prompt with 25 inference steps and guidance scale 7.5

stories but to generate new therapeutic narratives adapted to emotions, reference-based metrics were not suitable.

To address this, we designed an indirect evaluation strategy. First, we applied an external emotion detection model to the generated stories in order to predict the emotion conveyed in the text. This predicted emotion was then compared to the emotion originally detected by our trained facial emotion recognition model. A high correspondence between the two indicated that the generated story successfully reflected the intended emotional state of the child. While this approach does not capture narrative quality in terms of creativity or linguistic richness, it provides a relevant measure of emotional alignment and coherence.

CLIPScore. In addition to emotion alignment, we employed the CLIPScore metric, which is a reference-free evaluation measure. CLIPScore uses the CLIP model to compute semantic similarity between two modalities (text and image, or text and text). In our context, it was particularly useful to evaluate the consistency between the generated story (text) and the generated therapeutic image (visual). A higher CLIPScore indicates that the visual representation and the narrative are semantically coherent.

This dual evaluation—emotion correspondence for text and CLIPScore for text-image alignment—allowed us to assess not only whether the story matched the intended emotion but also whether the multimodal outputs (story and illustration) were consistent with each other. Although these methods do not fully replace human evaluation (which would be ideal for assessing engagement, readability, and therapeutic effect), they provided a reliable automated framework to validate the system.

3 Conclusion

In this project, the goal was to find a way to comfort Orphan when they are feeling some kind of emotions because sometimes it can be difficult for them . SOME TEST

Figure	num_inference_steps	guidance_scale	Neutral	Happy	Fearful
Fig. 4	50	7.5	0.6172	0.2144	0.1684
Fig. 5	25	7.5	0.5849	0.1970	0.2181
Fig. 7	70	7.5	0.2353	0.4689	0.2958
Fig. 6	25	15	0.3783	0.3140	0.3077

Table 5: Similarity of generated images with emotions and the main corresponding emotion. Dominant emotions: Fig. 4 → **neutral**, Fig. 5 → **neutral**, Fig. 7 → **happy**, Fig. 6 → **neutral**.

was made to find the perfect model to fine-tune to detect the big 7 kind of emotions. After we try to find the best model to generate stories with a lot of imagination because e wanted different stories even tough the emotion was the same we chose to use anime sytle computer vision model to be closer to the child world .

Even though a lot of research was made , this project have a lot of amelioration than can be made . The first is to maybe divide the story so we can generate multiple picture for one story so the child could understand the feeling that we tr to transmit because sometimes it difficult with only one picture. And maybe add sound or animation for this steps because it would have more impact The second thing is to try to build better prompt because even tough we get good results with the ones that we tried , it would be better if we could have take some features from the given picture to generate the story because maybe the child will be more concern if he saw some that is quite similar to him .

References

[1] Yuwei Chen and Jianyu He. "Deep Learning-Based Emotion Detection". In: *Proceedings of the Conference on Emotion AI*. Dublin, Ireland, 2020.