CAS2105 Homework 6: Mini AI Pipeline Project 🤗

Minji Jung (2023149028)

1 Introduction

This project provides a simple introduction to designing and evaluating an AI pipeline for a concrete visual classification task. Rather than training large models or relying on extensive model tuning, the goal is to understand the overall workflow of applied AI systems, including problem definition, baseline construction, pipeline design, and evaluation.

We focus on a binary image classification problem: distinguishing between daytime and nighttime scenes. This task is intuitive for humans but presents challenges for simple rule-based systems, especially under visually ambiguous conditions such as bright nighttime scenes or dim indoor daytime images.

To explore these challenges, we first implement a naïve baseline based on image brightness statistics. We then build an improved pipeline using a pre-trained vision–language model (CLIP) in a zero-shot inference setting. By comparing these two approaches quantitatively and qualitatively, this project highlights the limitations of simple heuristics and the benefits of semantic representations provided by modern pre-trained models.

2 Task Definition

- **Task description:** The task is to classify an input image into one of two categories: *daytime* or *nighttime*, based on the visual content of the image.
- **Motivation:** Although the distinction between day and night is trivial for humans, it can be challenging for simple rule-based systems due to variations in lighting conditions, indoor scenes, and artificial illumination. This task serves as a compact and intuitive testbed for examining the limitations of naïve heuristics and the benefits of semantic representations in AI pipelines.
- **Input / Output:** The input to the system is a single RGB image. The output is a binary label indicating whether the image represents a daytime scene or a nighttime scene.
- **Success criteria:** A system is considered successful if it achieves high classification accuracy on the evaluation dataset and consistently outperforms a naïve brightness-based baseline, particularly on visually ambiguous cases.

3 Methods

This section includes both the naïve baseline and the improved AI pipeline.

3.1 Naïve Baseline

- **Method description:** As a simple baseline, we classify images based on their average pixel brightness. Each image is first converted to grayscale, and the mean pixel intensity over all

pixels is computed. If the average brightness exceeds a fixed threshold of 120 (on a 0–255 scale), the image is classified as a daytime scene; otherwise, it is classified as a nighttime scene.

The threshold value of 120 was chosen empirically through preliminary inspection of the dataset. This value provides a rough separation between relatively bright images and darker ones while remaining simple and interpretable. Importantly, no optimization or tuning was performed beyond this manual selection, reinforcing the naïve nature of the baseline.

- **Why naïve:** This approach relies on a single low-level visual statistic and ignores semantic information such as scene context, lighting sources, or object presence. The threshold is fixed and not tuned to the dataset, making the method intentionally simplistic and representative of a naïve rule-based heuristic.
- **Likely failure modes:** The baseline is expected to fail on visually ambiguous cases, such as brightly illuminated nighttime scenes (e.g., urban streets with artificial lighting) and dark daytime images (e.g., indoor scenes or overcast weather). In these cases, brightness alone is insufficient to correctly determine whether an image was captured during the day or night.

3.2 AI Pipeline

- **Models used:** We use a pre-trained CLIP model (Vision–Language Transformer) to perform zero-shot image classification. The model provides aligned image and text embeddings that enable semantic comparison without task-specific training [1].

  The CLIP model was accessed using the Hugging Face Transformers library, which provides a standardized interface for loading and running pre-trained models [2].
- **Pipeline stages:** The pipeline consists of the following stages: (1) *Preprocessing*, where input images are converted to RGB format; (2) *Representation*, where the image and text prompts are encoded into a shared embedding space using CLIP; (3) *Decision*, where similarity scores between the image and candidate text prompts are computed, and the label corresponding to the highest score is selected. No post-processing is applied.
- **Design choices and justification:** CLIP is chosen because it enables inference-only classification through semantic alignment between images and natural language descriptions. This design avoids the need for dataset-specific training or tuning, making the pipeline simple, efficient, and well-suited to small datasets. Using minimal and neutral text prompts further reduces the risk of prompt engineering or overfitting to the evaluation set.

4 Experiments

4.1 Datasets

- **Source:** The dataset was manually collected from public image sources, primarily Google Images and Unsplash. Images were retrieved using diverse search queries related to both daytime and nighttime urban scenes.
- **Total examples:** The dataset consists of 60 images in total, with 30 daytime images and 30 nighttime images, ensuring a balanced class distribution.
- **Train/Test split:** No explicit train–test split is used. Since the AI pipeline operates in a zero-shot inference setting, all images are used exclusively for evaluation.
- **Preprocessing steps:** Images were converted to a unified RGB format and stored as JPG files. During dataset construction, visually unambiguous images (e.g., clear blue skies with visible sunlight or completely dark nighttime scenes) were intentionally excluded. Instead, images with ambiguous lighting conditions—such as overcast daytime streets, rainy urban

scenes, foggy daylight, and brightly illuminated nighttime environments (e.g., street lights or stadium lighting)—were preferentially selected to challenge brightness-based heuristics.

4.2 Metrics

We evaluate both the naïve baseline and the AI pipeline using classification accuracy. Since the task is a balanced binary classification problem (daytime vs. nighttime) with an equal number of examples per class, accuracy provides a clear and sufficient measure of overall performance.

Accuracy directly reflects the proportion of correctly classified images and enables a straightforward comparison between the rule-based baseline and the zero-shot CLIP pipeline. More complex metrics such as precision, recall, or F1-score were not used, as they do not provide additional insight for this balanced and symmetric classification setting.

4.3 Results

Table 1 summarizes the quantitative performance of the naïve baseline and the proposed AI pipeline. The brightness-based baseline achieves an accuracy of 0.617, indicating that simple heuristics struggle with this task under ambiguous lighting conditions. In contrast, the CLIP-based pipeline achieves a substantially higher accuracy of 0.967, demonstrating the effectiveness of semantic representations for distinguishing between daytime and nighttime scenes.

Method	**Accuracy**
Naïve Baseline	0.617
AI Pipeline (CLIP)	0.967

Table 1: Classification accuracy of the naïve baseline and the AI pipeline.

Beyond quantitative results, qualitative inspection further highlights the differences between the two approaches. The naïve baseline frequently misclassifies visually ambiguous scenes, including dark daytime images such as overcast or rainy urban streets, as well as brightly illuminated nighttime scenes with strong artificial lighting. These errors arise from the baseline's reliance on global brightness, which fails to capture higher-level visual context.

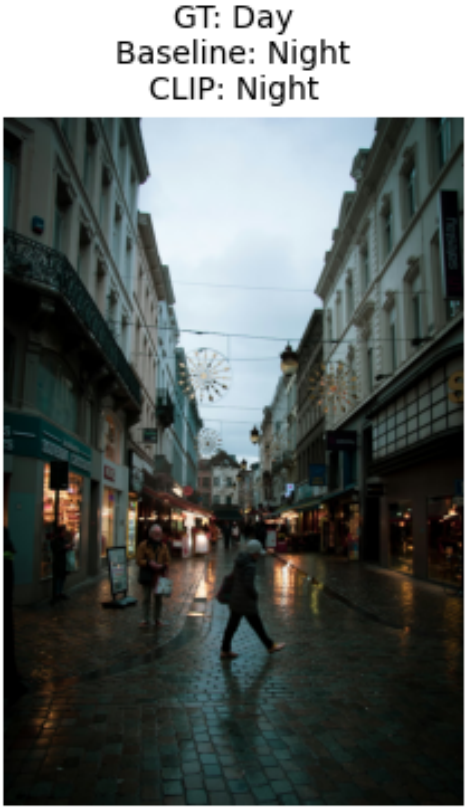
In contrast, the CLIP-based pipeline correctly classifies most of these challenging cases by leveraging semantic and contextual cues beyond overall luminance. Representative qualitative examples are shown in Figure 1, where the baseline fails while the CLIP-based pipeline produces the correct prediction.

However, CLIP is not error-free. Failure cases were observed in ambiguous scenes where specific environmental factors created misleading visual cues. For instance, the model struggled with images dominated by heavy tree shadows where only a small fraction of the sky was visible, as well as overcast scenes characterized by cold color temperatures and low overall luminance. These examples highlight that, despite its strong performance, the CLIP-based pipeline can still misinterpret daytime scenes as night when shadows or weather conditions mimic the visual features of nighttime.

(a) Foggy daytime scene (GT: Day). Naïve baseline failed.

(b) Bright nighttime street scene (GT: Night). Naïve baseline failed.

(c) CLIP failure case (GT: Day), due to highly ambiguous lighting conditions.



(d) Another CLIP failure case (GT: Day), due to low overall luminance.

Figure 1: Qualitative comparison between the naïve brightness-based baseline and the CLIP-based pipeline. Each images is labeled with GT, Prediction from Baseline and CLIP respectively. Top row: representative cases where the baseline fails while CLIP predicts the correct label. Bottom row: failure cases of the CLIP pipeline under highly ambiguous lighting conditions.

5 Reflection and Limitations

The experimental results show that the CLIP-based pipeline performed significantly better than the naïve brightness-based baseline, exceeding initial expectations. In particular, the AI pipeline successfully handled many visually ambiguous cases where simple heuristics failed, such as overcast daytime streets or brightly illuminated nighttime scenes. A total of **21 images** were identified where the baseline produced incorrect predictions while the CLIP pipeline classified them correctly, highlighting the advantage of semantic representations over low-level visual statistics.

Despite its strong performance, the CLIP pipeline was not perfect and failed on a small number of images. The failure cases of the CLIP pipeline, illustrated in Figure 1, highlight that even large pre-trained models can struggle under highly ambiguous lighting conditions.

These failure cases typically involved scenes where the distinction between day and night was inherently ambiguous, such as transitional lighting conditions or strong reflections combined with artificial illumination. This suggests that even large pre-trained models can struggle when visual cues conflict with learned semantic priors.

Classification accuracy proved to be an effective metric for this balanced binary task, clearly capturing the performance gap between the two approaches. However, accuracy alone does not fully describe model confidence or robustness under distribution shifts. With more time or computational resources, future work could explore alternative prompts, additional contextual cues, or larger and more diverse datasets to further analyze the limits of zero-shot inference.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021.

[2] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. https://github.com/huggingface/transformers, 2020.