Neural Wave - Duferco
Università della Svizzera italiana

Team Sbirulini

October 27, 2024

De Grandi Alessandro, Gianinazzi Mattia, Karpenko Voloymyr, Lunghi Marzio, Zang Qianbo

1 Project Overview

The project proposed by Duferco aims to develop an AI-based solution for detecting the alignment of steel bars on a production line.

1.1 Dataset Description

Duferco provided a unlabeled dataset of 15,000 images captured under various conditions on their production lines. From the Figure 1, the images differ significantly, with some offering clear visual data for alignment analysis, while others are noisy, making it difficult to discern the steel bars.

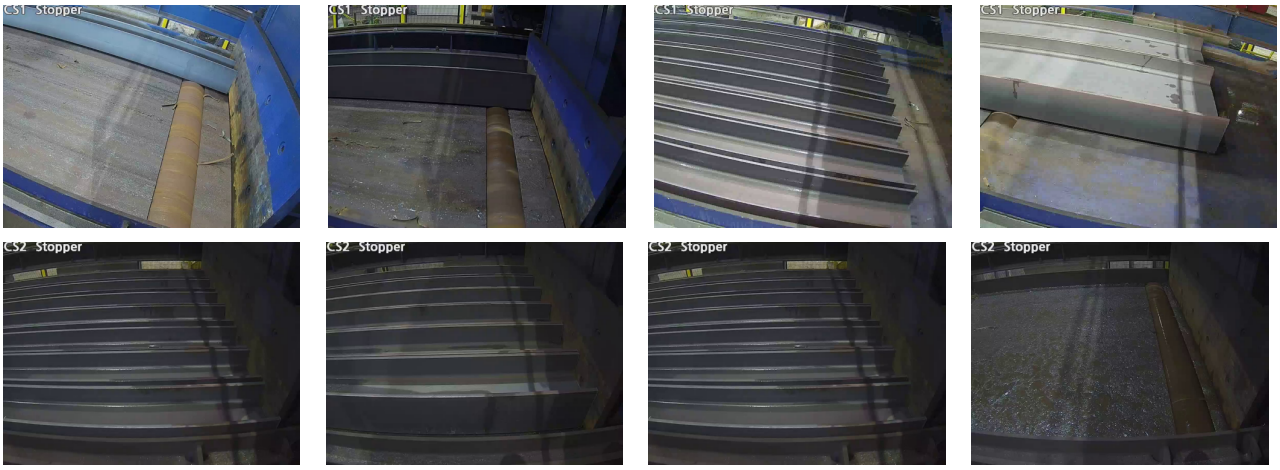


Figure 1: Dataset example

1.2 Problem Definition

The primary objective is to accurately detect the alignment of steel bars across all provided images. More precisely, the model must classify each image into one of the two classes: (0) *not aligned* and (1) *aligned*. The key challenges include:

- **Image Labeling**: Due to the variety and complexity of the images, correctly labeling the dataset for alignment poses a challenge.
- **Model Generalization**: Developing a model that generalizes well to new data is critical, especially given that the dataset originates from two different production lines, each with varying camera positions and lighting conditions.
- **Real-Time Performance**: The model must operate in real time, with an inference time of less than 0.5 seconds per image even without GPU support.

2 Methodology

2.1 Exploratory Data Analysis

2.1.1 Data Engine

Given the discrete quality of several images, we opted for manual labeling. Using our data engine, as shown in Figure 2, each image can be classified according to the presence or absence of an alignment and the scenario. In addition, along with the original image, two others, one corresponding to an denoised image and one with simple edge detection (Canny Edge Detection), serve as support during labeling.

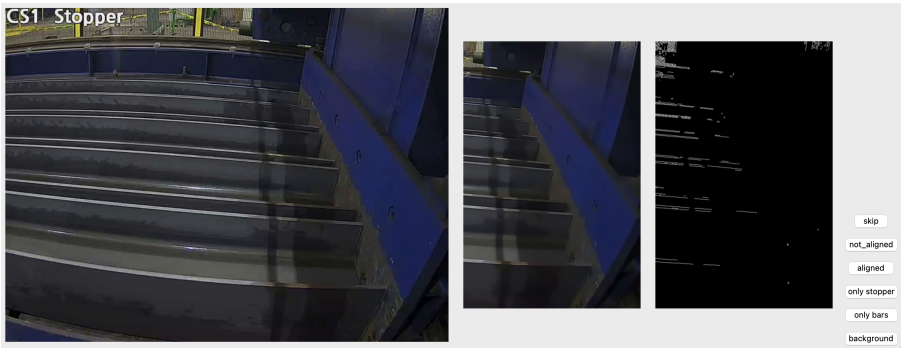


Figure 2: Interface of our data engine

Initially, we manually annotated 8,000 images, subsequently a further we have selected in total, including images under different light conditions (good or bad light), 5572 images which are divided into 55% *aligned* and 45% *not aligned*.

2.1.2 Problem Complexity

To get a better insight of the complexity of the labeling task we first started with a simple approach. We perform clustering algorithms to group the unlabeled datasets leveraging significant differences in the features across different categories. The goal is to explore whether unsupervised learning can effectively substitute for supervised learning in this task.

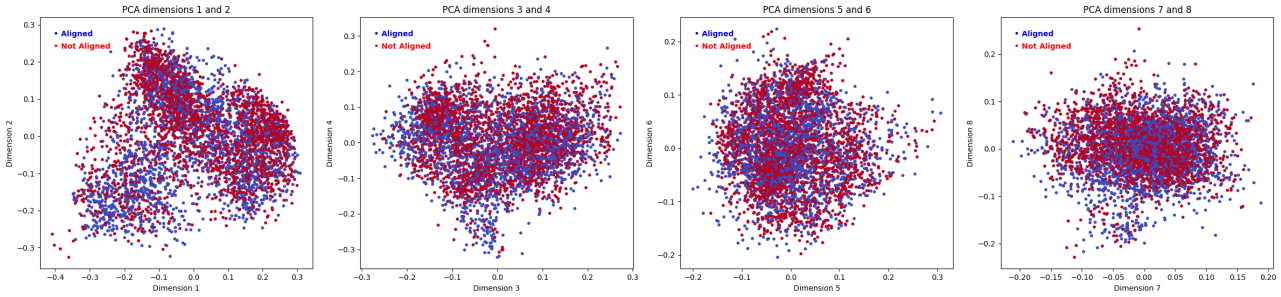


Figure 3: PCA after the image embeddings

We used a Contrastive Language–Image Pre-training (CLIP) [3] model is a multi-media architecture designed to align images and text in a shared embedding space. We first used CLIP's image encoder to generate a fixed-length vector representation (embedding). We conducted Principal Component Analysis (PCA) after the image embeddings to determine the optimal dimensionality. From the Figure 3, our visualization and analysis of multiple principal components revealed no significant differences between the distributions of the two data types across each component. This analysis represents a supervised classifier is necessary.

2.2 Model Building and Validation

To train our classifier we use a Residual Neural Network called ResNet18 [2]. To this model, shown in Figure 4, we attach an additional Linear layer to predict the alignment of steel bars. We separated data which we have labeled to train (80%) and validation datasets (20%).

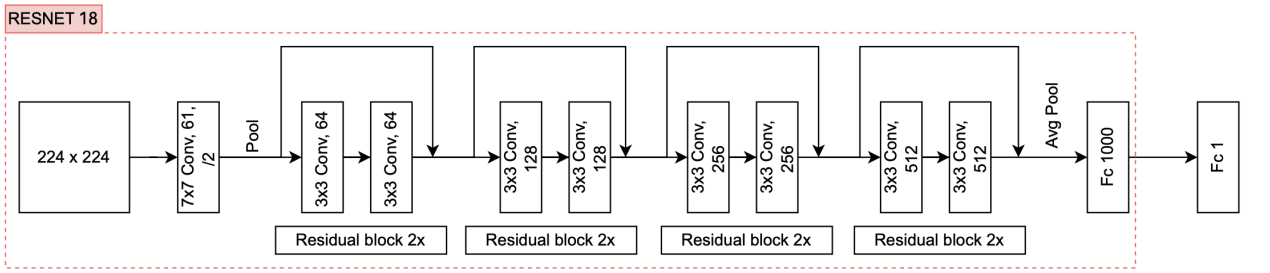


Figure 4: ResNet18 model and binary classification layer

We train this model for a total of 20 epochs using an Adam optimizer and a learning rate of 0.0001. To train the model we employ the `BCEWithLogits` loss and during the validation we keep track of the best model with the f_β-score with $\beta = 0.5$. The full training curve is shown in Figure 5.

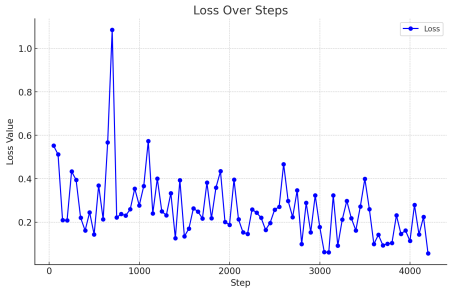


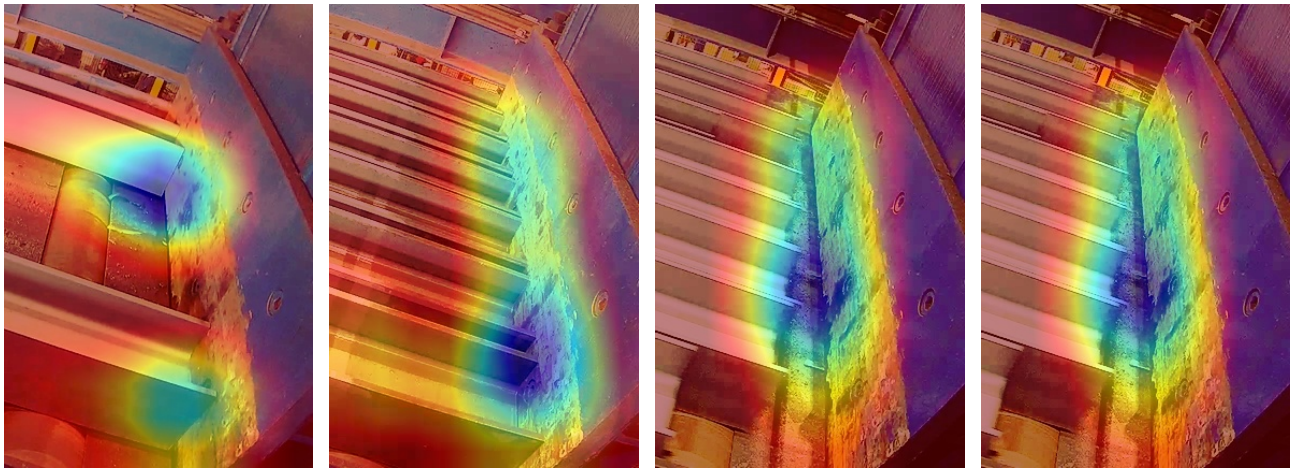
Figure 5: Loss of train dataset

Precision	**Recall**	**f_β-score ($\beta = 0.5$)**	**Accuracy**
0.84	0.78	0.83	0.95

Table 1: Evaluation Metric

2.2.1 Further Analysis

Using PyTorch-Gram-Cam [1], we can inspect the area of interest of the model. As shown in Figure 6 the latter is able to focus on the area where steel bars should touch the stopper. The attention map stretches when there are many steel bars when all of them are touching, suggesting that when all are aligned the model as to confirm that not even a single one is not aligned. When not aligned the model focus just on a single point of interest.



(a) aligned (b) aligned (c) not aligned (d) not aligned

Figure 6: Examples of attention maps on the validation set

References

[1] J. Gildenblat and contributors. Pytorch library for cam methods. `https://github.com/jacobgil/pytorch-grad-cam`, 2021.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.