Prediction of Political Violence from Protest Images using Multi-Task Learning

Applications of Deep Learning in High Dimensional Data Analysis and Image Processing

Anastasiia Chernavskaia

Contents

June 21, 2025

1 Motivation

Most event databases and studies of political violence use natural language processing and/or human-coded evaluations to estimate the severity and degree of violence of protests, riots and other forms of civil unrest. For example, GDELT estimates the severity of protests by analyzing news reports, including the reported number of participants, level of violence, and duration of the protest. Another aggregator that provides estimates of severity of political violence - Factal - uses human labeling to determine the severity score for a given event (from 1 to 5).

Possible downsides of both approaches:

- Reliance on news articles means susceptibility to selective reporting, subjective framing, sensationalism, and unreliable estimates (like participant counts).
- Automated categorization using NLP might miss nuanced event interpretations or struggle with subjective definitions of "violence" from text. Automated analysis across multiple languages might miss subtle cultural nuances or misinterpret specific linguistic expressions of protest and violence.
- Human labeling, on the other hand, can lack consistency among annotators, and is labor-intensive.

Additionally, similar event aggregators don't seem to have a way to validate the accuracy and impartiality of their event coding and intensity scores against ground truth data.

Violence is a critical dimension of protest in understanding social mobilization, as violent protests typically generate a much higher level of media and public attention. There might be other cues that one could use to approximate the level of violence in a protest, such as police or government statements or the number of people who have been killed, injured, or arrested. However, this information can be often inaccurate or not provided at all to the public in an official channel. Therefore, the goal of our study is to take advantage of unfiltered stream of data in social media and to assess the level of perceived violence for protest events.

2 Research Question

The goal of this project is to train a multi-task CNN which would be able to detect protest activity and estimate violence intensity from an image. Additionally, the model should be able to detect elements that would be useful for political violence classification, such as the size of the crowd (above 20 and above 100 people), the presence of police, fire, signs, etc.

3 Dataset

I have requested the dataset from the authors of the paper "Improving Computer Vision Interpretability: Transparent Two-Level Classification for Complex Scenes", which was

not publicly available. The data I received contained more than 40,000 images - a mix of protest photos pulled from social media and image banks, and random non-protest images.

Since the dataset had a significant class imbalance (with around 65% of non-protest photos), I undersampled the majority class, reducing the dataset to around 25,000 images.

The data contained a complementary annotation file with the "ground truth" labels: binary flags for protest and the presence of 10 individual attributes: sign, photo, fire, police, children, group of more than 20 people, group of more than 100 people, flag, night, and shouting. The "violence score" was present only for the images depicting protest. According to the authors of the paper, the annotations for the violence intensity score were collected through pairwise comparisons where workers selected the more violent image in randomly sampled pairs from 11,659 protest images. The methodology involved 58,295 image pair evaluations, with each image appearing in 10 different pairs and each pair assessed by 10 annotators to ensure fair comparisons. Using the Bradley-Terry model, these pairwise comparisons were converted into continuous violence scores.

4 Related work

Recent research at the intersection of computer vision and political science has begun to implement automated analysis of protest and political violence using image data. Notably, Chan (2020) developed a deep learning framework for processing protest images from Twitter, identifying characteristics like violence and police presence, and extracting demographic information from faces. Other studies used convolutional neural networks to classify protest images and detect visual attributes. Researchers have also explored the relationship between social media sentiment and offline political violence, trying to forecast violent events from both visual and textual data.

Several studies have explored identification of violent activities in images or videos using static image features or motion cues. Most of them focus on physical violence, including detecting fights among players in sports games, recognizing bloody scenes in films, and identifying aggressive behavior in short video clips. Collectively, these works highlight the growing capacity of deep learning methods to analyze the visuals aspects of political violence, providing new state-of-the-art tools for political science and media studies.

5 Model

The model that I use for the task is a multi-task convolutional neural network (CNN) that performs three related computer vision tasks simultaneously using a shared feature extractor and task-specific heads.

CNN architecture:

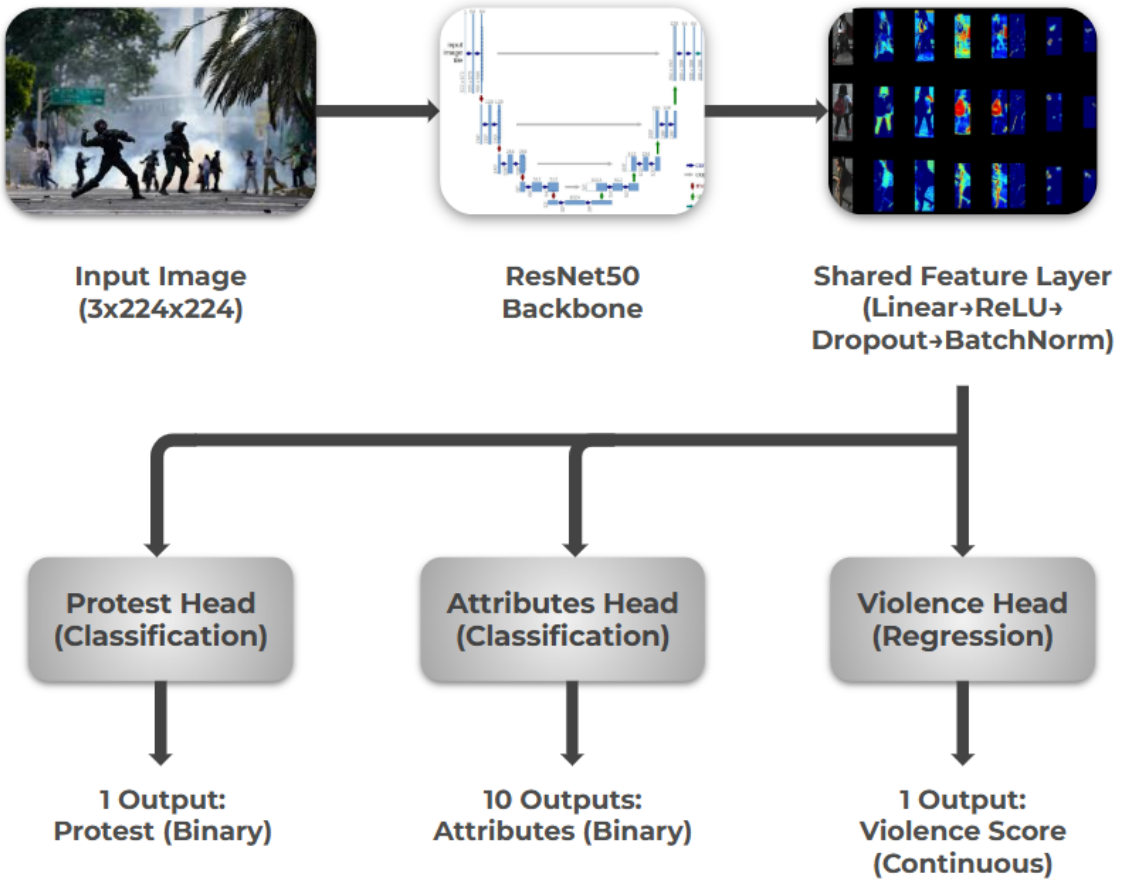

Figure 1: Simplified ResNed Multi-head Architecture

Pre-trained **ResNet50** (ImageNet weights) serves as a shared backbone for feature extraction, outputting 2048-dimensional feature vectors. Then a **feature layer** compresses the shared feature representation from 2048 to 1024 dimensions before passing it to the task-specific heads:

1. **Protest classification head** is a binary classifier (protest vs. non-protest).
2. **Visual attributes head** is a multi-label classifier for 10 attributes. It also has a deeper architecture of 3 linear layers, and uses sigmoid-activated outputs (independent probabilities).
3. **Violence regression head** predicts violence intensity (on a scale from 0 to 1.

Why this architecture?

The model is designed with several key features that enhance its efficiency and effectiveness. First, it employs a single shared backbone to process images only once, enabling multiple task-specific heads to operate in parallel on the extracted shared features. This **multi-task setup reduces computational cost** compared to using

separate models for each task.

Second, the model tackles task-specific complexity by **allocating more layers to the attributes head, which handles classification across 10 different classes**, while keeping the protest and violence heads simpler since they each output a single value.

Third, the model enforces **output constraints by applying sigmoid activation functions** across all outputs. The protest head produces a binary probability indicating the presence or absence of a protest, the attributes head outputs independent probabilities for each of the 10 attributes, and the violence head predicts a normalized score between 0 and 1 representing the level of violence.

Finally, the model incorporates several **regularization techniques to improve generalization and reduce overfitting**. These include a dropout rate, batch normalization layers, and the use of a pre-trained backbone network, which leverages learned features from large datasets to stabilize training.

Additionally, the shared feature layer can be used to extract general-purpose image embeddings, which would be valuable for other downstream tasks or transfer learning scenarios. The modular design also allows for fine-tuning or replacing individual heads to adapt the model to new domains or task combinations.

6 Evaluation of Results

After the first iteration of training, the model showed signs of overfitting across all tasks, particularly in protest classification.
In the second iteration of training, I have made the following adjustments:

- Increased dropout from 0.3 to 0.5.
- Reduced batch size from 32 to 16.
- Increased weight decay from 1e-5 to 1e-4.
- Reduced early stopping patience from 6 to 4.
- Rebalanced task weights: 1 for protest detection, 1 for individual attirbutes, and 0.8 for violence score (since regression converges faster).
- Added learning rate scheduler with patience=2 epochs.
- Performed data augmentation (random crops, rotations, horizontal flip and color jitter).
- Added stronger regularization: changed Adam to AdamW (better for weight decay).
- Added norm-based gradient clipping.

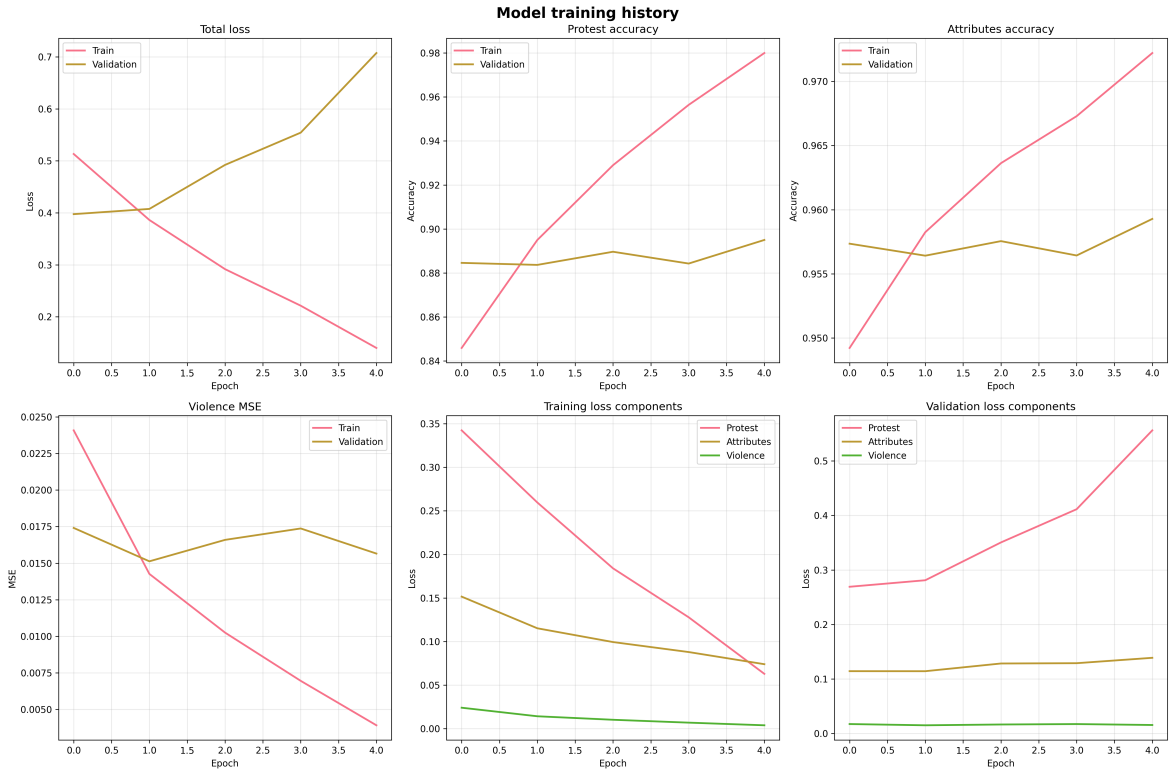

Figure 2: Training history of the first training iteration

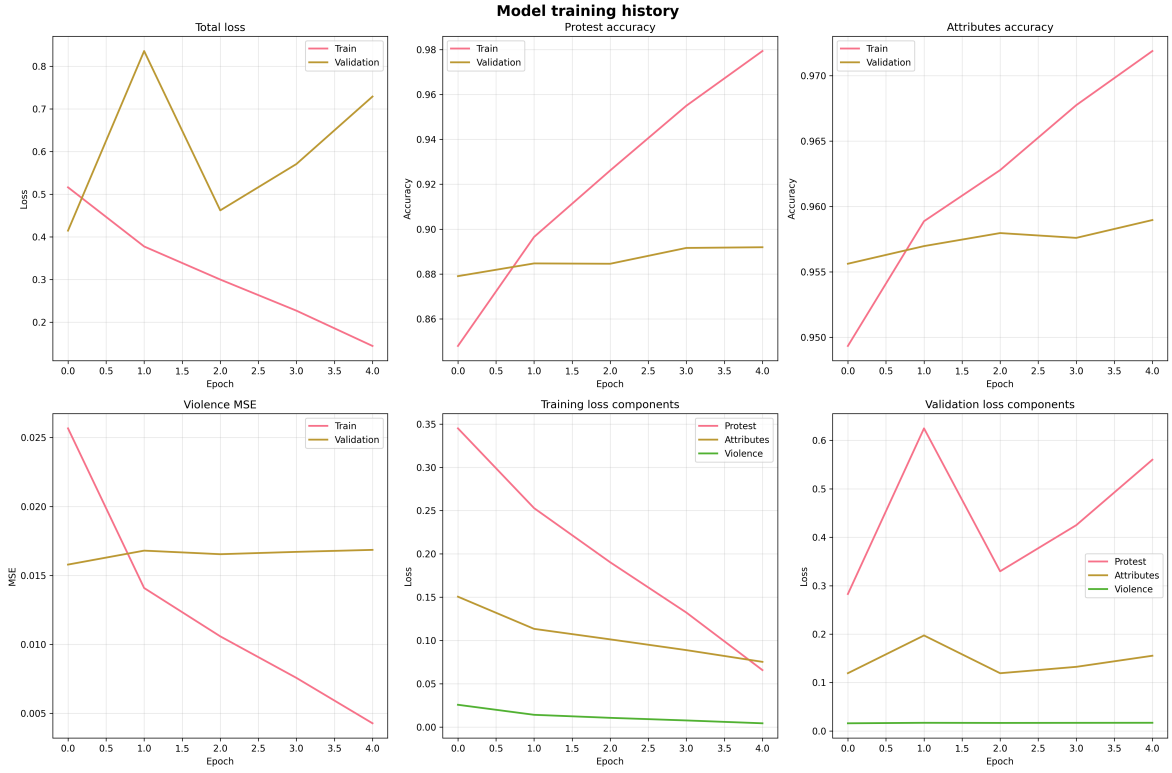

Figure 3: Training history of the second training iteration

The new training history looks almost identical for the training set, but the validation set line now tracks more closely to the training set line. We need to look at some

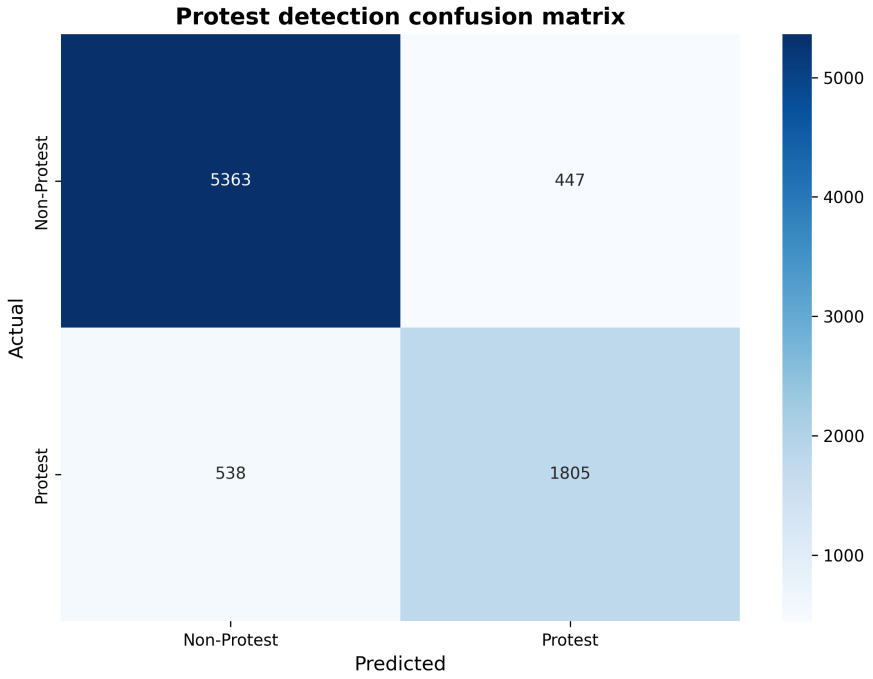

Figure 4: Overall, the model demonstrates strong performance, with most predictions falling along the diagonal, indicating accurate protest detection.

additional metrics to evaluate the model's performance.

It is insightful to look at the examples that were correctly classified and misclassified to understand the direction of potential improvement.

And finally, GradCam analysis might give us a clue about the attention of the model to various elements of an image.

7 Takeaways

The multi-task model was trained to automatically classify the content of images and estimate the violence intensity from the shared visual feature representation. This approach to estimate violence and protest dynamics from social media images has certain advantages over textual analysis, since visual language is more universal than spoken language.

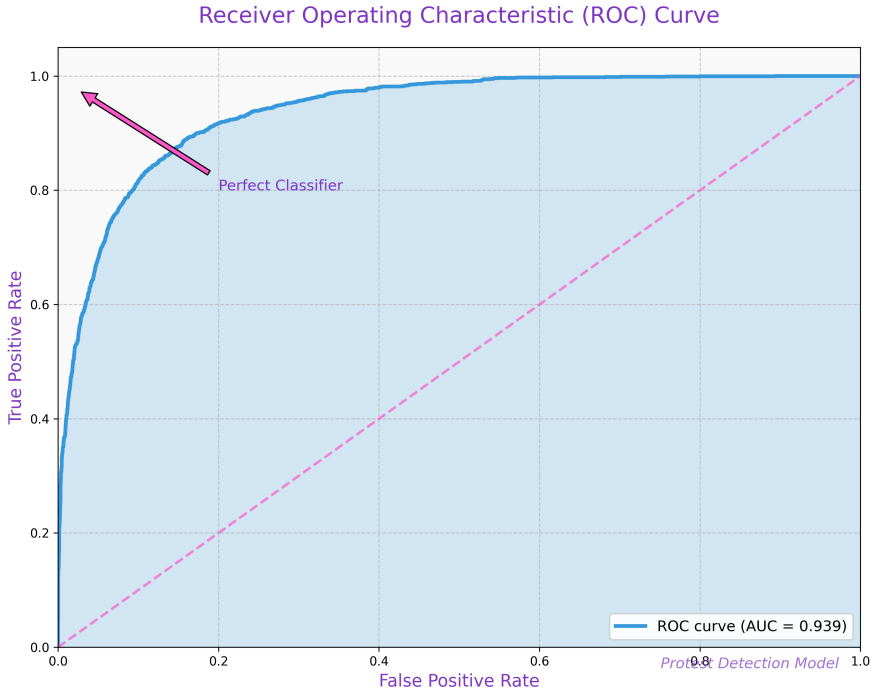

Figure 5: The curve rises steeply toward the top-left corner, indicating a high true positive rate and low false positive rate across thresholds. The area under the curve (AUC) is 0.939, which signifies excellent discriminative ability between protest and non-protest images.

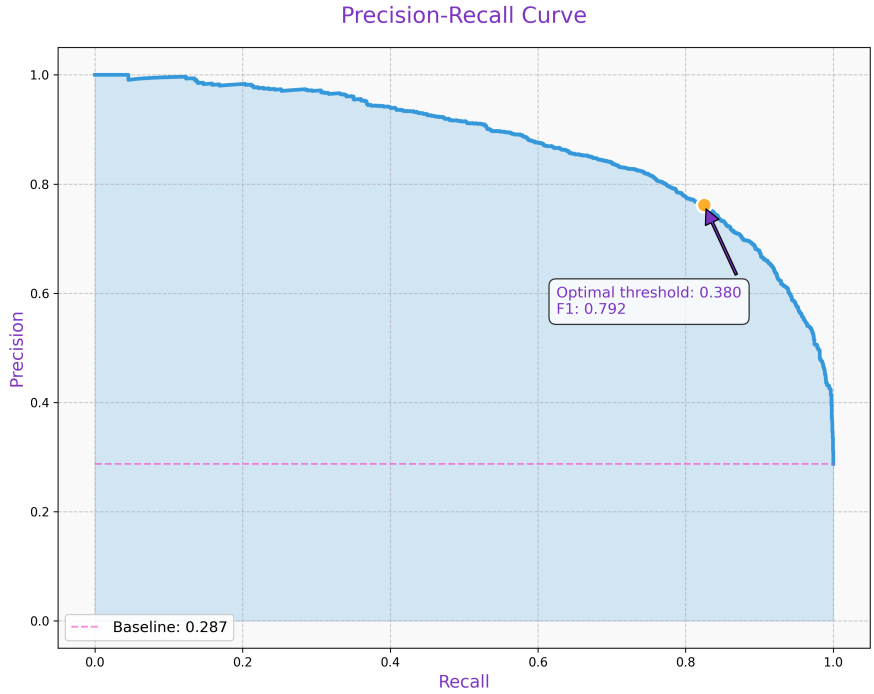

Figure 6: The model achieves a strong balance between precision and recall, with an optimal F1 score of 0.792 at a threshold of 0.38, far above the baseline. This indicates the model is effective at distinguishing protest images from non-protest images, likely due to clear visual cues in protest scenes and effective data augmentation and regularization during training.

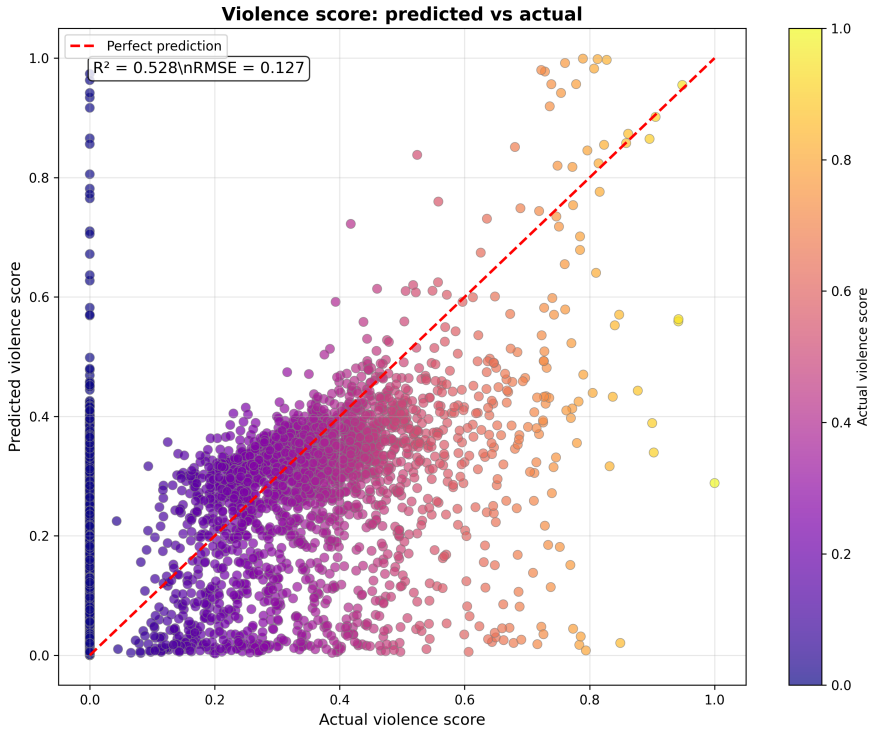

Figure 7: The model captures the general trend (as shown by the clustering along the diagonal), indicating moderate predictive power. However, there is a tendency to underestimate high violence scores and to compress predictions toward the mean, which is typical when the regression task involves subjective or imbalanced labels — here, most images likely have low or moderate violence, making extreme values harder to predict accurately.

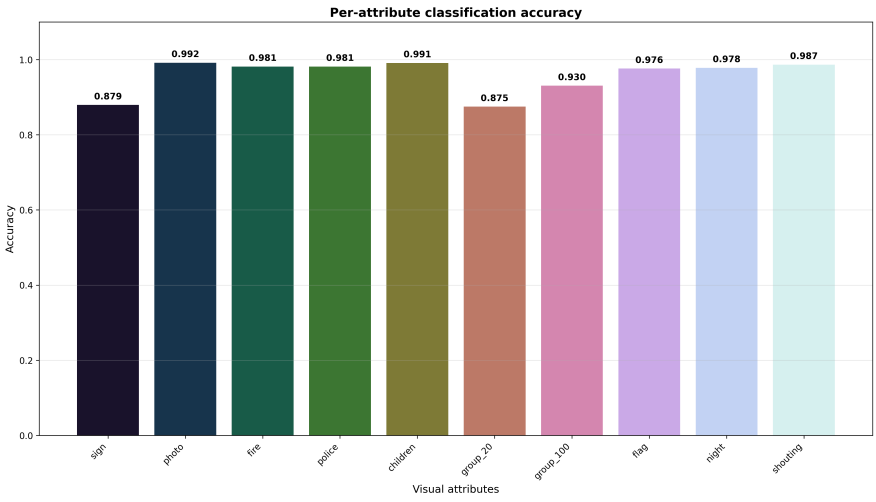

Figure 8: The model achieves very high accuracy on most visual attributes, with values above 0.95. Accuracy is noticeably lower for "sign" (0.879) and "group 20" (0.875). One possible explanation is that the model might default to the majority class ("0"), so the accuracy for attributes that have a higher class imbalance will always be higher. Also, it could be that the attributes that are visually distinctive are easier for the model to learn, while rare or subtle features pose greater challenges.

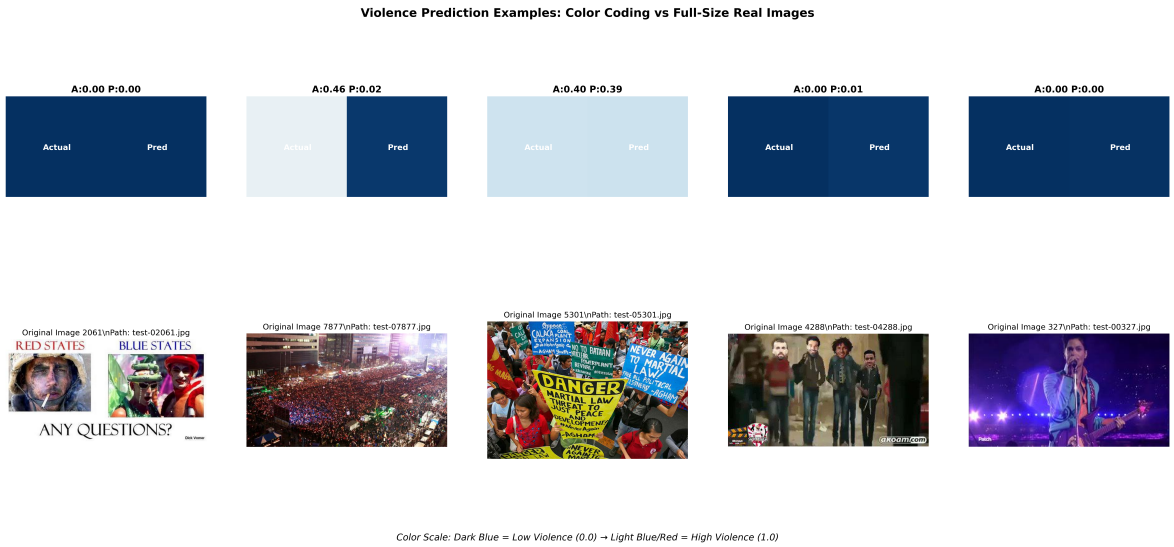

Figure 9: Images with actual vs predicted violence scores. This set of images compares actual and predicted violence scores for a sample of five images. The model performs well on clear-cut cases, accurately assigning very low violence scores (0.00) to peaceful or neutral images, as indicated by the dark blue color. For images with moderate or ambiguous violence, such as large crowds or protest signs, the model produces intermediate predictions (e.g., actual 0.46, predicted 0.02; actual 0.40, predicted 0.39), reflecting some sensitivity to the visual cues but also showing a tendency to underestimate violence in more complex scenes.

Figure 10: Misclassified images. The second set highlights images where the model made mistakes, providing insight into the model's sources of confusion. Some non-protest scenes are misclassified as protests with high confidence, often because they feature crowds or banners — visual elements commonly associated with protests. Conversely, some protest scenes are predicted as non-protests, especially when the visual context is ambiguous, the crowd is sparse, or the protest cues are subtle. The confidence scores further reveal that the model is sometimes quite certain even when it is wrong, which suggests that certain visual features — like smoke, raised hands, or group gatherings — are strong but imperfect signals for protest detection. These errors point to the inherent challenge of distinguishing protests from other large public gatherings and the need for more nuanced feature learning or additional context in future model iterations.

Figure 11: Grad-CAM heatmap for protest detection

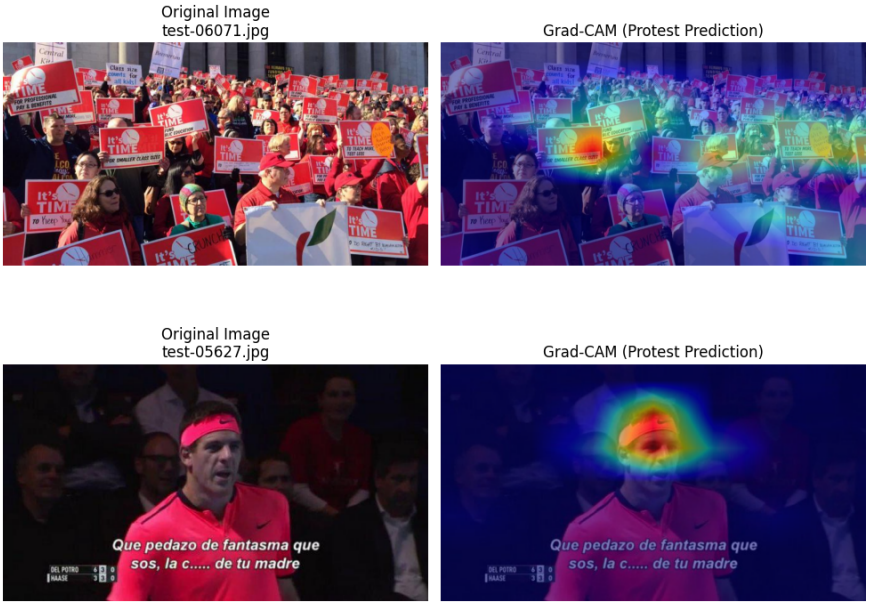

Figure 12: In the first set of images, the Grad-CAM heatmaps clearly highlight the areas containing dense crowds, raised banners, and clusters of people — visual elements that are highly indicative of protest activity. Notably, in the second pair of images in Figure 12, the model focuses on the central figure in a non-protest context, which may explain occasional misclassifications when protest-like visual cues (such as crowds or signage) are absent.

Figure 13: In the "sign" detection examples, the model's attention is strongly concentrated on protest signs, showing that the model is correctly focusing its attention on the relevant textual and symbolic cues when predicting the presence of signs. Overall, these visualizations demonstrate that the model is generally attending to the most informative regions of the image for each task, and zeroing in on specific objects like signs for attribute recognition.

8 References

1. Chhapariya, K., Benoit, A., Buddhiraju, K. M., Kumar, A. (2024). A Multitask Deep Learning Model for Classification and Regression of Hyperspectral Images: Application to the large-scale dataset. *arXiv preprint arXiv:2407.16384.* https://arxiv.org/abs/2407.16384

2. Liang-Hua Chen, Hsi-Wen Hsu, Li-Yun Wang, and Chih-Wen Su. 2011. Violence detection in movies. *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on. IEEE, 119–124.*

3. Markus Brenner and Ebroul Izquierdo. 2012. Social event detection and retrieval in collaborative photo collections. *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval. ACM, 21.*

4. Scholz S, Weidmann NB, Steinert-Threlkeld ZC, Keremoğlu E, Goldlücke B. Improving Computer Vision Interpretability: Transparent Two-Level Classification for Complex Scenes. *Political Analysis. 2025;33(2):107-121. doi:10.1017/pan.2024.18*