Threat Detection Using Images Extracted From Video Clips

Shirshajit Sen Gupta, Neeladri Bhuiya, Sumanth Yalamarty, Hardik Narang

School of Computing, National University of Singapore

Abstract

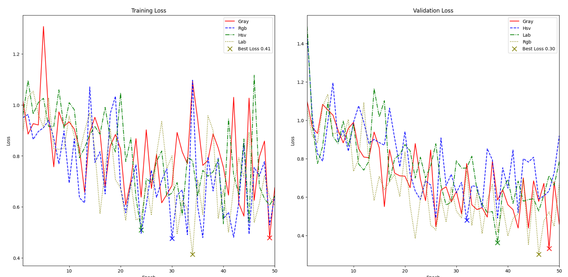
Taking inspiration from Roy et al.[1] and [2], we used a novel **BiFPN**[3] network along with a self-supervised Fourier frequency domain to perform multi-class classification on the threat detection dataset. Training and validation accuracies of **98.46%** and **89.11%** were achieved.

Data Preprocessing

The data was split into 3 sets - a training set consisting of 80% of the data, a validation set, and a test set each holding 10%. On the training set random horizontal flips, random rotations ($\pm 10°$) and normalization with means [0.485, 0.456, 0.406] and standard deviations [0.229, 0.224, 0.225] on the 3 color channels. Images were resized in accordance with the literature, i.e. 224×224 for the ResNet models [2] and 512×512 for the BiFPN [1].

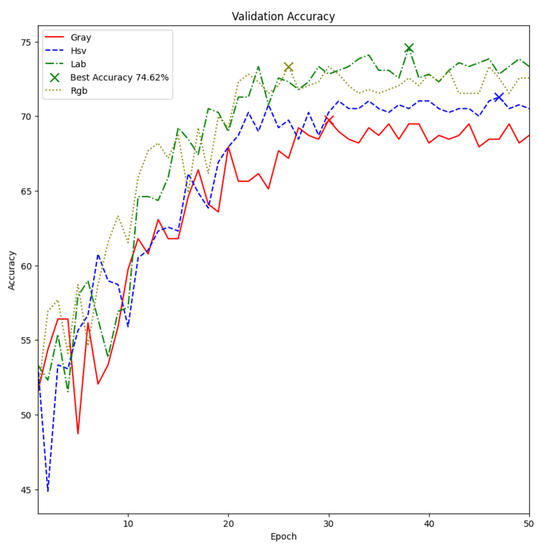
Color Spaces - ResNet18

We ran experiments using the ResNet-18 model on 4 color spaces - RGB, HSV, CIE-LAB, and grayscale. For grayscale images, a single channel was copied thrice to avoid parameter reduction. All models were run for 50 epochs with an initial learning rate of 0.001 and learning rate decay by a factor of 10 every 10 epochs. Adam was used as the optimizer with CrossEntropyLoss.
It should be noted that increasing the depth of the model did not yield much in returns (ResNet-101).



(a) Training Loss

(b) Validation Loss



(c) Validation Accuracy. LAB outperforms the rest by a small margin

Figure: Comparing results across multiple color spaces. In general, LAB performs the best while grayscale is the poorest.

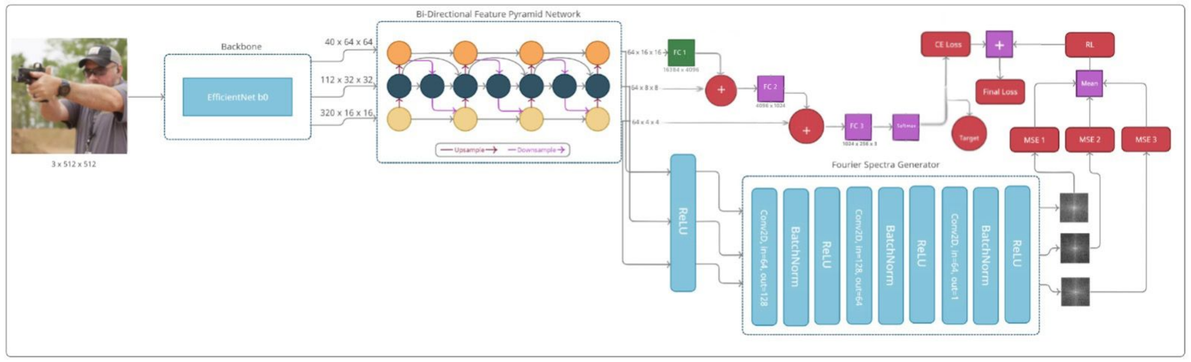


Figure: Model architecture for the improved BiFPN classification model. The additive computations were inspired from [2]. The FC layer outputs are then passed into a softmax

Fourier Spectra - BiFPN

The models are composed of a BiFPN backbone followed by auxiliary feature extraction from the BiFPN feature maps using a self-supervised *Fourier Spectra Generator* model. A **reconstruction loss** is calculated from the 3 resulting Fourier maps and is combined with CrossEntropyLoss. Images of size 512×512 were used with a learning rate of 0.001 and weight decay of 0.00001.

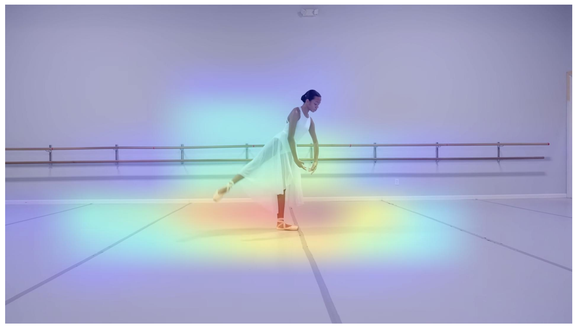


Figure: GradCam[4] attention maps with the BiFPN model. The model successfully differentiates between carrying and threat classes by focusing on the weapon and the person's posture. Above is the attention map for the Normal class

Figure: Attention maps - Carrying

Figure: Attention maps - Threat

Two variants of output probability computations were attempted. Initially, the feature maps were passed through a sigmoid layer and the mean was taken, which was then stacked and passed through a Softmax layer to get a probability distribution.

The second model architecture can be seen above. The fully-connected layers were used to resize the channel weights to the same size. Both losses are weighted equally.

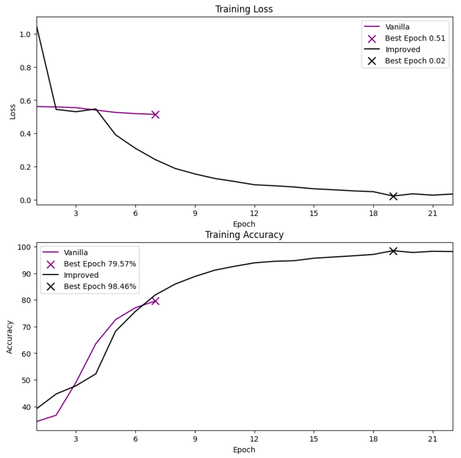


Figure: Training losses and accuracies

Training for the first model was stopped after 7 epochs by early stopping.

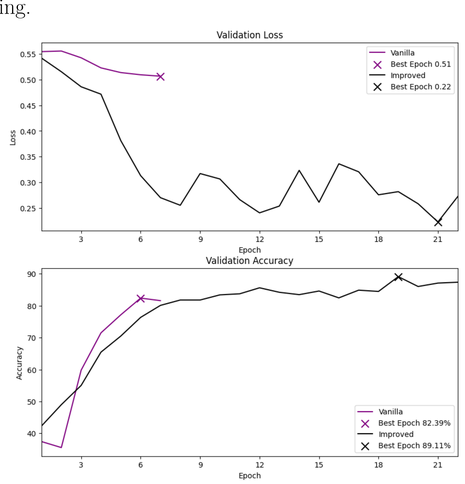


Figure: Validation losses and accuracies

Compared to the ResNet models, the improved BiFPN model reaches accuracies of 89%. This is an almost 20% improvement. Further improvements can be achieved by cleaning the data further and hyperparameter optimization.

Conclusion

In conclusion, it has been demonstrated that Fourier Spectra information can be successfully incorporated into a model to improve classification accuracy. Similar approaches can be tried with Gabor Wavelets and LBP matrices as well.

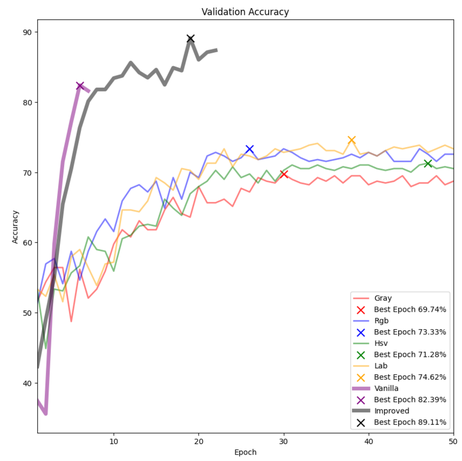


Figure: Accuracies of all models. The BiFPN models significantly outperform the rest.

References

[1] Koushik Roy, Md Hasan, Labiba Rupty, Md Sourave Hossain, Shirshajit Sengupta, Shehzad Noor Taus, and Nabeel Mohammed.
Bi-fpnfas: Bi-directional feature pyramid network for pixel-wise face anti-spoofing by leveraging fourier spectra.
Sensors, 21(8):2799, 2021.

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

[3] Mingxing Tan, Ruoming Pang, and Quoc V Le.
Efficientdet: Scalable and efficient object detection.
In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[4] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra.
Grad-cam: Visual explanations from deep networks via gradient-based localization.
In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[5] Mingxing Tan and Quoc Le.
Efficientnet: Rethinking model scaling for convolutional neural networks.
In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

Acknowledgements

We'd like to acknowledge the main author of [1] Koushik Roy, and Atin Sakkeer Hussain for their help.

Contact Information

- Shirshajit Sen Gupta (Email: shirshajit@u.nus.edu)
- Neeladri Bhuiya (Email: e0589907@u.nus.edu)
- Sumanth Yalamarty (Email: e0638867@u.nus.edu)
- Hardik Narang (Email: e0638906@u.nus.edu)

NUS | Computing
National University of Singapore