Vehicle make and model recognition using mixed sample data augmentation techniques

Talha Anwar[1], Seemab Zakir[2]
[1]Center of Chiropractic Research, New Zealand College of Chiropractic, Auckland 1149, New Zealand
[2]Department of Engineering Technology, Foundation University, Rawalpindi, Pakistan

Article Info

Article history:

Received Sep 28, 2021
Revised Jul 7, 2022
Accepted Aug 5, 2022

Keywords:

Deep learning
Mixed data augmentation
Vehicle identification system

ABSTRACT

Vehicle identification based on make and model is an integral part of an intelligent transport system that helps traffic monitoring and crime control. Much research has been performed in this regard, but most of them used manual feature extraction or ensemble convolution neural networks (CNNs) that result in increased execution time during inference. This paper compared three deep learning models and utilized different augmentation techniques to achieve state-of-the-art performance without ensembling or fusing the models. Experimentations are made without any augmentation, with standard augmentation, and by mixed sample data augmentation techniques. Gradient accumulation and stochastic weighted averaging with mixed precision are used to have a large batch size that helped to reduce training time. The dataset comprised 48 vehicles' models running on the road of Pakistan. The highest accuracy and F1 score of 97% and 95% using the FMix augmentation technique with EfficientNetV2-S architecture gave the confidence that the proposed solution can be implemented in production.

Corresponding Author:

Talha Anwar
Center of Chiropractic Research, New Zealand College of Chiropractic
Auckland 1149, New Zealand
Email: chtalhaanwar@gmail.com

1. INTRODUCTION

Vehicle identification system (VIS), an integral component of the intelligent transport system (ITS), brings ease to the traffic management system and helps against criminal activities. VIS is widely used in road violation detection, traffic congestion alarm, and unmanned driving. Millions of vehicles are on the road in big cities, making it challenging to track a particular vehicle. The vehicles' number plate is mostly used to track them [1], but number plates can be changed easily, leading to false identification. VIS also helps automate tax collection at toll plazas based on vehicle type.

With the advent of artificial intelligence (AI), deep learning has been widely used in transportation [2] Some recent studies used traditional imaging techniques such as haar-like features with AdaBoost classifier [3] and pattern descriptors with support vector classifier [4]. The pattern descriptors study used local binary patterns, median binary patterns, directional gradient patterns, and local arc patterns as features. Kiran *et al.* also studied different colour spaces such as red, green and blue (RGB), green (Y), blue (Cb), red (Cr) (YcbCr) and hue, saturation, value (HSV) for descriptor extraction [4] haar-like features-based study first removed shadows using HSV colour space to reduce the chances of false detection. Different single feature methods, such as colour moment, local binary pattern (LBP) features, Hu moment features, angle features, and circularity are also used. Using Adaboost 85.8% accuracy is achieved [3]. Qiu *et al*. [5] compared the performance of haar features along with convolution neural network (CNN). Using haar-like features, 86.72% and 91.86% precision and recall are achieved, which increased by 5.63% and 0.2% with CNN [5]. Gholamalinejad and Khosravi proposed a novel CNN architecture composed of CNN layers with squeeze-and-excitation (SE) modules. Instead of using classic max pooling or average pooling, they used haar wavelet as a pooling layer [6]. The data is composed of 5 classes, including bus, heavy truck, medium truck and pickup. They achieved an accuracy of 95.1% [6]. Ajitha *et al.* proposed a shallow CNN model with traditional augmentation techniques such as flip, rotation, shear, crop and zoom, resulting in an accuracy of 92.3% [7]. Mansor *et al.* [8] achieved an accuracy of 95% with 4 class classification problems. Their work is based on emergency vehicle type classification and had images of fire trucks, police cars, ambulances and standard cars [8]. Hassan *et al*. compared different classifiers with cyclic learning rate and used the MixUp image augmentation technique to achieve an accuracy of 93.96% through ensembling homogeneous models of DenseNet201 [9]. Though the CNN-based model has gained much attention in recent years, manual feature-based classification is still being studied recently. Chen detected multiple features from the vehicle, such as taillight features, shadow area features and other descriptors. Radial basis function (RBF) artificial neural network is further used for classification and achieved 97% accuracy [10]. Another manual feature-based study used histogram-oriented gradients (HOG) and ant colony optimization (ACO) to classify vehicles and achieved an accuracy of 90% [11].

All the existing studies either deal with a few vehicle models, manual features extraction or used ensemble models in which multiple models are tested during inference resulting in increased prediction time. As the VIS is implemented in real-time, it needs to be robust. Keeping in view the limitation, we proposed a single network-based approach that yields the state of the art performance. Three different models and five augmentations techniques are compared. All the experiments are seeded for the purpose of reproducibility. The main contributions of this paper are,

- Different deep learning architectures are compared without using any augmentation technique, with commonly used and mixed sample data augmentation techniques (MSDA).
- Ensemble and fusion of different models increase the inference time, so the approach used a single model that performed better than the existing ensembled models.
- The proposed approach achieved state-of-the-art performance with 97% and 95% accuracy and F1 score, respectively.

The paper is organized: The introduction, motivation, and literature review on vehicle classification are presented in section 1. Section 2 describes the methodology in detail. Section 3 deals with results and discussion. The conclusion is made in section 4. The implementation is publicly available at GitHub [12].

2. METHOD

2.1. Dataset

We used images of common cars running on the road of Pakistan [13]. There are 3,103 and 752 training and test images divided into 48 car models/classes. Figure 1 shows the sample image. Table 1 shows the vehicle name and the number of images available for training for each vehicle.

2.2. Transformation

Transformation is a technique to produce variation in the data. It helps to generalize prediction on test data and avoid over-fitting the model. Albumentation [14] library is used for this purpose. Following the main standard Augmentation used for applied transformations:

- Resize: all images are resized to 256×256
- Center crop: crop all images are centre cropped to 224×224
- Horizontal Flip: fifty per cent of images are horizontally flipped
- Vertical Flip: fifty per cent of images are flipped vertically
- Shift scale rotate: fifty per cent of images are randomly shifted, rotated, and scaled in height and width.
- CLAHE: contrast limited adaptive histogram equalization (CLAHE) is a modified form of adaptive histogram equalization. In histogram equalization, the intensity range of the image is stretched between 0 and 255 to improve the contrast of the image. However, this led to either too dark or too bright picture. Adaptive histogram handled this issue by dividing the image into small patches and applied histogram equalization on each patch. This sometimes led to over-amplification of contrast if the image has noise. CLAHE performed bi-linear interpolation on the edges of patches and reduced this contrast amplification by removing the artificial boundaries.
- Cutout: cutout is one of the ways to handle over-fitting. In this technique, black boxes are introduced in images, making the image classification hard, and reduced the chances of over-fitting.
- Normalization: normalization led to fast convergence and speeds up the training process.

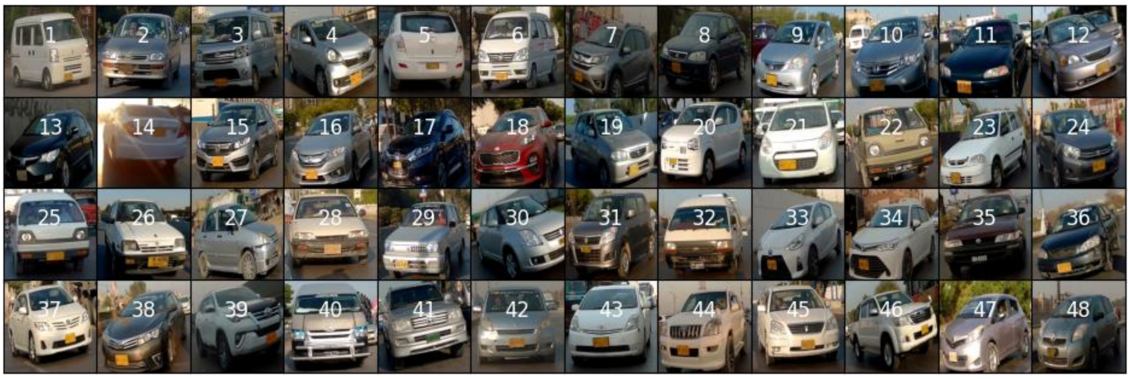

Figure 1. Sample vehicles image from each class label, the number on each image corresponds to the vehicle ID in Table 1

Table 1. Vehicle models and the number of images for that models. ID column is related to Figure 1. No. shows number of training examples for that model

ID	Vehicle model	No
1	Daiatsu Core	80
2	Daiatsu Hijet	44
3	Daiatsu Mira	81
4	FAW V2	29
5	FAW XPV	26
6	Honda BRV	27
7	Honda city 1994	32
8	Honda city 2000	69
9	Honda City aspire	105
10	Honda civic 1994	16
11	Honda civic 2005	34
12	Honda civic 2007	74
13	Honda civic 2015	31
14	Honda civic 2018	82
15	Honda Grace	21
16	Honda Vezell	38
17	KIA Sportage	25
18	Suzuki alto 2007	132
19	Suzuki alto 2019	56
20	Suzuki alto japan 2010	27
21	Suzuki carry	13
22	Suzuki cultus 2018	269
23	Suzuki cultus 2019	108
24	Suzuki Every	20
25	Suzuki highroof	63
26	Suzuki kyber	52
27	Suzuki liana	33
28	Suzuki margala	16
29	Suzuki Mehran	195
30	Suzuki swift	118
31	Suzuki wagonR 2015	112
32	Toyota hiace 2000	23
33	Toyota Aqua	77
34	Toyota axio	20
35	Toyota corolla 2000	39
36	Toyota corolla 2007	82
37	Toyota corolla 2011	127
38	Toyota corolla 2016	270
39	Toyota fortuner	43
40	Toyota Hiace 2012	72
41	Toyota Landcruser	17
42	Toyota Passo	61
43	Toyota pirus	23
44	Toyota Prado	21
45	Toyota premio	18
46	Toyota Vigo	53
47	Toyota Vitz	81
48	Toyota Vitz 2010	48

2.3. Mixed sample data augmentation

Large neural networks are notorious for memorizing data instead of learning it even in strong regularization and fail during inference. Though standard data augmentation helped in generalization, this technique is data-dependent and required domain knowledge. Anwar and Zakir [15] studied that standard augmentation sometimes led to poor results. They explored different image augmentation techniques on electrocardiogram (ECG) graphs and found that the best results are obtained without applying any augmentation. CNN focused on the discriminative part of the image instead of the whole image leading to poor generalization. Regional dropout techniques such as the CutOut helped the CNN to view the bigger image perspective, but this reduced the proportion of informative pixels of training data [16]. Mixed Sample data augmentation (MSDA) techniques are introduced to overcome standard augmentation and generalization issues. MSDA mixed different distributions of data to produce new data from the same distribution of existing data. It is categorized into two policies, interpolation and masking. MixUp is an example of interpolation, whereas CutMix and FMix are an example of masking MSDA.

2.3.1. Mixup

MixUp mixed two images from different classes and linearly interpolated them to produce a new image. It not only interpolated the input images' features but also interpolated the corresponding target [17]. The working principle of MixUp is shown in (1) and (2),

$$\tilde{x} = \lambda x_i + (1 - \lambda)x_j \tag{1}$$

$$\tilde{y} = \lambda y_i + (1 - \lambda)y_j \tag{2}$$

x_i and x_j are raw images in (1) and y_i and y_j are the one-hot encoded labels in (2). λ drawn from β distribution is used to mix two random images. MixUp increased the capability of deep learning architectures to learn from corrupted labels and improved the generalization. Linear interpolation of input images reduced the memorization by large deep learning models [18].

2.3.2. CutMix

Cutout and MixUp inspired CutMix paper. It claimed to resolve the issues in MixUp. Though MixUp improved classification performance, the resulting sample is unnatural. CutMix replaced an image patch with a patch of another random picture from the training data [16]. It is like a cutout where a patch is replaced with zeros and MixUp where two images are mixed.

$$\tilde{x} = Mx_i + (1 - M)x_j \tag{3}$$

Patch mixing in training images is shown in (3). M is a binary mask indicating where the dropout rectangular region should be placed. Then this rectangular dropout region is replaced by a patch of another image. Mixing of one-hot encoded labels is the same as in the MixUp technique. CutMix focused on the less discriminative part of the object, whereas Mixup focused on the entire image but produced unnatural artefacts.

2.3.3. FMix

CutMix reduced overfitting by increasing the observable data points without changing the data distribution. However, CutMix used square patches, which is a limitation and leads to distortion. FMix claimed to resolve the issue in CutMix by using binary masks obtained by applying a threshold to low-frequency images from the Fourier space. The authors first sampled low-frequency grayscaled masks from Fourier space and then converted them to binary masks using a threshold. Once a binary mask is obtained, two images from different classes are overlaid together, such as 0 pixels of binary mask corresponded to one image and pixels with 1 value of binary mask is related to another image from a different class. FMix, unlike CutMix, proposed patches of different shapes which maximize the number of possible masks [19].

Overall, when data is limited and learning from individual examples is easier, MixUp is a good candidate, and FMix is a better choice when data is abundant. In Figure 2, MixUp shows that two images are mixed together in an overlay fashion. CutMix shows that a square patch of another image replaces a square patch. FMix shows that another image from the training data replaced a randomly shaped patch of an image.

2.4. Deep learning architecture

Deep learning is a subset of artificial intelligence that takes the complex raw data as input, automatically extracts valuable features, and performs task-relevant work such as classification or regression.

In image classification, deep learning boomed in 2014 after VGGNet came out. Though before VGG, AlexNet was there, VGG16 outperformed it by 10%. At that time, it was believed that increasing the layer increased the performance of the model, until in December 2015, ResNet paper was released and proved that adding layers helped to some extent and started decreasing the performance beyond that [20]. To date, ResNet or ResNet variants are one of the most used architecture; therefore, we decided to use ResNet as our baseline.

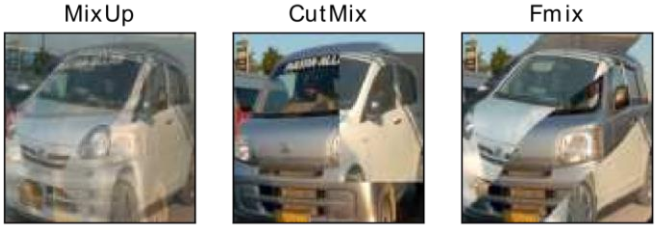

Figure 2. Mixed sample data augmented images of two cars

2.4.1. ResNet

Ideally, a deeper neural network is preferable as it yields better results. Nevertheless, this comes with the cost of vanishing gradient and degradation. By increasing the depth of the neural network, the gradients became very small during back-propagation and reached zero; this phenomenon is known as vanishing gradient. Though this problem can be resolved using the rectified linear units (ReLU) activation function, skip connection also played a role. Skip connection back-propagates the gradient of larger magnitude by skipping some layers in between.

ResNet paper explained that further deepening neural network led to a significant error rate characterized by degradation. Adding layers saturated the model, and the error rate started increasing. It is believed that if a shallow network is working fine, the additional deep layers should work the same though it did not happen, and deep networks start performing poorly. So, an identity function is added from a shallow layer to a deeper layer, and the model started learning that identity function. In ResNet, this identity function ensured that the deep network output should be identical to the shallow network. ResNet paper named this identity function as skip connections that skip some layers and pass information directly to other layers by an identity function. In the worst case, the performance of a deeper network will not be worse than a shallow network, and in the best scenario, it can be better than the shallow network [20]. Multiple ResNet variants are described by network size and the number of layers skipped by the skip connections. We used ResNet-50 as it is neither tiny to underfit nor very large to overfit.

2.4.2. DenseNet

DenseNet was proposed in 2018 by Huang *et al*. [21]. Based on the observation, if there is a shorter connection between input and output layers, the model can be deeper, more accurate, and more efficient to train. DenseNet is based on dense blocks and transition layers. In dense blocks, each coming layer received collective information from all previous layers both directly and indirectly. Similarly, in back-propagation, the error signal collectively flowed to all layers. For each layer, the feature maps of all previous layers are considered output, and the output of that layer is considered as input for all subsequent layers. For the sake of downsampling to reduce network size, a transition layer between two dense blocks is used. This layer is composed of a 1×1 convolution filter preceded and followed by batch normalization and an average pooling layer. We used DenseNet 121 in this study.

2.4.3. EfficientNetV2

Most of the deep learning architecture either scaled the depth such as ResNet by increasing the number of layers or width by adding more neurons/filters in each layer, for example, wide ResNet [22]. Wider networks learn more detailed features and are easier to train because they are usually shallower However, shallower and wider networks have an issue in learning high-level features. Some networks used high-resolution images such as InceptionV3 which used 299×299 image size [23]. Scaling a specific dimension such as depth, width, and resolution increase accuracy up to a limit. EfficientNet in 2019 claimed that its depth, width and resolution should be scaled proportionally to make a deeper network more effective. So the authors proposed a compound scaling method to scale width, depth and resolution proportionally [24].

EfficientNetV2 in June 2021 is one of the latest proposed models and is known for faster training speed [25]. This model is based on training awareness neural architecture search (NAS) and progressive scaling. It is observed that small image sizes require less regularization as compared to large image sizes. So the authors started with small image size and increased the size progressively. They used EfficientNet as their backbone architecture and applied the NAS strategy, though the authors removed unnecessary search options to reduce the search space. This paper used a small kernel size of 3×3 and added more layers to compensate for the reduced receptive field. Other tweaks are applied to reduce the memory access overhead in EfficientNet, such as removing the last stride layer. In our study, EfficientNetV2-S is used.

2.5. Explainability of MSDA techniques

To understand the impact of MSDA techniques, we used gradient-weighted class activation mapping (Grad-CAM) that explained which area of an image is focused by a network to decide the label class. Grad-CAM produced a localization heatmap of the target by utilizing its gradient against the last convolution layers and highlighted the essential regions of the image [26]. To generate Grad-CAM PyTorch library for CAM methods is used [27].

2.6. Additional information

Fifty epochs are trained with a learning rate and batch size of 0.001 and 48, respectively. AdamW optimizer is used instead of Adam as it provides better results [15]. Pytorch Lightning framework is used for implementation. Accuracy, macro F1 score, precision and recall are used for evaluation. Mixed precision, gradient accumulation, and stochastic weight averaging (SWA) techniques are used to speed up the training time. Gradient accumulation is a technique to train the model with larger batch sizes by updating weights after some batches instead of every batch. SWA helps to generalize the model, whereas Mixed precision reduces training time up to 8x [28] by allowing a large batch size.

3. RESULTS AND DISCUSSION

This paper deals with the identification of commonly used vehicles in Pakistan. Table 2 shows the performance of different augmentation techniques with three deep learning architectures. Without using any augmentation technique, an F1 score of 88%,91%, and 90% is achieved using ResNet-50, DenseNet121 and EfficientNetV2-S, respectively. When standard augmentations are applied, the F1 score increased in all three models, which shows the impact of data augmentation. With MixUp augmentation techniques in which two images are mixed together in an overlay fashion, there is not much difference in the F1 score of different deep learning models compared with standard augmentations. When CutMix is applied, there is 1% increment in accuracy obtained using EfficientNet and ResNet. FMix augmentation technique achieved the highest accuracy and F1 score in all deep learning models. EfficientNetV2 with FMix augmented input resulted in accuracy and F1 score of 97% and 95%, respectively. With EfficientNetV2 this is a 2% increment in F1 score compared to MixUp and CutMix augmentation techniques. Without augmentation, the macro F1 score is 90% which increased by 5% with FMix augmentation technique. These MSDA augmentation techniques are applied without standard augmentation to study the impact of MSDA augmentations alone. Figure 3 shows validation loss using five different augmentation techniques. The lowest validation loss is achieved using FMix augmentation technique when EfficientNetV2-S model is used. EfficientNetV2-S also showed the second-lowest curve with the CutMix MSDA technique. CutMix and MixUp produced similar results in standard augmentation, but FMix outperformed them in all three deep learning architectures.

Figure 4 shows the heatmap generated by the Grad-CAM technique. MixUp techniques paid attention to most parts of the car's front, but its focus is diverged. On the other hand, CutMix focused on the right front headlight, but its span of coverage is less. FMix covered both aspects, its heatmap is more focused and spread over the front area. It helped the model visualize and focus broader region while making a decision and providing better results.

The existing studies are either based on manual features extraction [3] or multiple ensemble models [9] resulted in reduced performance during inference. The proposed solution is robust during inference but has some limitations during training. The more the augmentation, the more time a model needs to train itself because an image undergoes a series of transformations before feeding to the neural network. We observed that MSDA augmentation takes time to do the mathematical calculation of image mixing. However, no augmentations are applied during test time, making the model robust during the inference.

The limitation of standard augmented CNN or features-based classifiers is adversarial image attacks. Manipulating certain car parts can make CNN fool, and it would not predict the vehicle. On the other hand, MSDA techniques heavily altered the image by placing other pictures on it; thus, there would be minimal chances of adversarial attacks. FMix resolved the issues of CutMix which is inspired by MixUp, so theoretically, FMix should have better performance [19]. Practically this is proved as FMix augmentation got 1%, 2% and 2% accuracy improvement in EfficientNetV2-S, DenseNet121 and ResNet50 as compared to CutMix, respectively.

Table 2. Model performance using different augmentations techniques

	ResNet-50				DenseNet121				EfficientNet			
Techniques	F1	Prec	Rec	Acc	F1	Pre	Rec	Acc	F1	Prec	Rec	Acc
None	88%	90%	87%	92%	91%	94%	91%	94%	90%	92%	88%	94%
Standard	90%	91%	90%	93%	92%	93%	91%	94%	93%	95%	92%	95%
MixUp	90%	94%	89%	94%	91%	94%	90%	94%	93%	96%	92%	95%
CutMix	91%	94%	90%	95%	91%	94%	90%	95%	93%	96%	92%	96%
FMix	93%	94%	92%	95%	94%	95%	94%	97%	95%	96%	95%	97%

Prec: precision, Rec: recall, F1: f1 score, Acc: accuracy

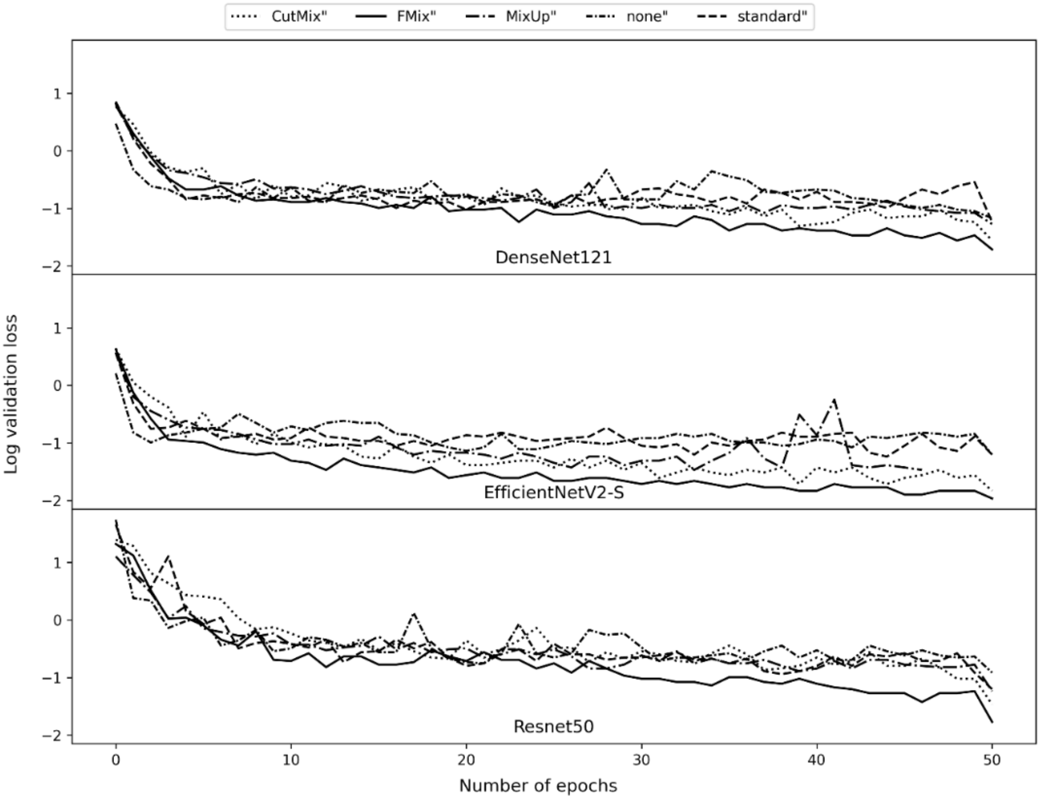

Figure 3. Validation loss using different architectures and augmentation techniques. Three different subplots with a common axis show three deep learning architectures. Five different patterns show five different augmentation methods

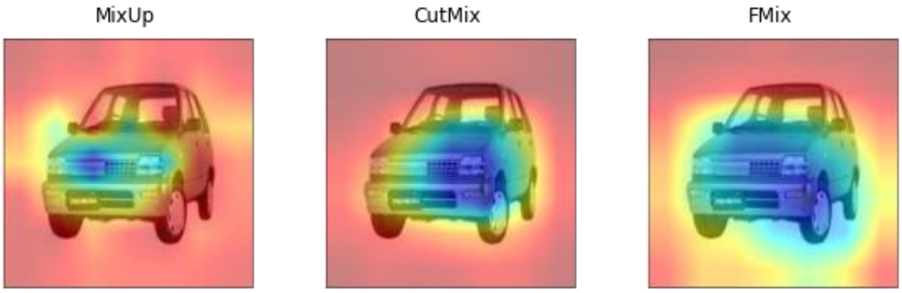

Figure 4. Grad-CAM heatmap for MSDA augmentation techniques

4. CONCLUSION

In this paper, different augmentation techniques are studied to achieve the state of art results. Unlike other studies that used manual feature extraction such as edge detection or haar features, this study used end-to-end CNN to extract and classify features automatically. Ensemble models are not used because they are not feasible for deployment because of time complexity and inference time limitations. Five augmentation scenarios are used, such as no augmentation, standard augmentation, and three mixed sample data augmentation techniques. Three deep learning algorithms such as ResNet, DenseNet and EfficientNet are used. All five augmentation techniques and three CNN architectures are compared. Mixed sample data augmentation techniques helped to achieve state-of-the-art performance using an EfficientNetV2-S model on a dataset comprised of 48 models of vehicles running on the roads of Pakistan. Further, the heatmap of MSDA techniques are compared to understand the learning of deep learning model. FMix image augmentation with EfficientNetV2 resulted in the highest F1 score of 95%, which is 5% better if no augmentation is applied and 2% better if standard commonly used augmentation techniques are used.

REFERENCES

[1] P. N. Huu and C. V. Quoc, “Proposing WPOD-NET combining SVM system for detecting car number plate,” *IAES International Journal of Artificial Intelligence (IJ-AI)*, vol. 10, no. 3, p. 657, Sep. 2021, doi: 10.11591/ijai.v10.i3.pp657-665.

[2] Y. Wang, D. Zhang, Y. Liu, B. Dai, and L. H. Lee, “Enhancing transportation systems via deep learning: a survey,” *Transportation Research Part C: Emerging Technologies*, vol. 99, pp. 144–163, 2019, doi: 10.1016/j.trc.2018.12.004.

[3] L. Zhang, J. Wang, and Z. An, “Vehicle recognition algorithm based on Haar-like features and improved Adaboost classifier,” *Journal of Ambient Intelligence and Humanized Computing*, 2021, doi: 10.1007/s12652-021-03332-4.

[4] V. Keerthi Kiran, S. Dash, and P. Parida, “Vehicle recognition using extensions of pattern descriptors,” in *IOP Conference Series: Materials Science and Engineering*, 2021, vol. 1166, no. 1, p. 12046, doi: 10.1088/1757-899x/1166/1/012046.

[5] L. Qiu, D. Zhang, Y. Tian, and N. Al-Nabhan, “Deep learning-based algorithm for vehicle detection in intelligent transportation systems,” *Journal of Supercomputing*, vol. 77, no. 10, pp. 11083–11098, 2021, doi: 10.1007/s11227-021-03712-9.

[6] H. Gholamalinejad and H. Khosravi, “Vehicle classification using a real-time convolutional structure based on DWT pooling layer and SE blocks,” *Expert Systems with Applications*, vol. 183, 2021, doi: 10.1016/j.eswa.2021.115420.

[7] P. Ajitha, S. Jeyakumar, Y. N. Krishna K, and A. Sivasangari, “Vehicle model classification using deep learning,” in *Proceedings of the 5th International Conference on Trends in Electronics and Informatics, ICOEI 2021*, 2021, pp. 1544–1548, doi: 10.1109/ICOEI51242.2021.9452842.

[8] M. A. Hakim Bin Che Mansor, N. A. Mohamad Kamal, M. H. Bin Baharom, and M. Adib Bin Zainol, “Emergency vehicle type classification using convolutional neural network,” in *2021 IEEE International Conference on Automatic Control and Intelligent Systems, I2CACIS 2021 - Proceedings*, 2021, pp. 126–129, doi: 10.1109/I2CACIS52118.2021.9495899.

[9] A. Hassan, M. Ali, N. M. Durrani, and M. A. Tahir, “An empirical analysis of deep learning architectures for vehicle make and model recognition,” *IEEE Access*, vol. 9, pp. 91487–91499, 2021, doi: 10.1109/ACCESS.2021.3090766.

[10] X. Chen, H. Chen, and H. Xu, “Vehicle detection based on multifeature extraction and recognition adopting RBF neural network on ADAS system,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8842297.

[11] R. S. El-Sayed and M. N. El-Sayed, “Classification of vehicles’ types using histogram oriented gradients: comparative study and modification,” *IAES International Journal of Artificial Intelligence*, vol. 9, no. 4, pp. 700–712, 2020, doi: 10.11591/ijai.v9.i4.pp700-712.

[12] T. Anwar, “Pak vehicle classification,” *GitHub repository*. 2021.

[13] M. Ali, M. A. Tahir, and M. N. Durrani, “Vehicle images dataset for make and model recognition,” *Data in Brief*, vol. 42, p. 108107, Jun. 2022, doi: 10.1016/j.dib.2022.108107.

[14] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: fast and flexible image augmentations,” *Information (Switzerland)*, vol. 11, no. 2, 2020, doi: 10.3390/info11020125.

[15] T. Anwar and S. Zakir, “Effect of image augmentation on ECG image classification using deep learning,” in *2021 International Conference on Artificial Intelligence, ICAI 2021*, 2021, pp. 182–186, doi: 10.1109/ICAI52203.2021.9445258.

[16] S. Yun, D. Han, S. Chun, S. J. Oh, J. Choe, and Y. Yoo, “CutMix: regularization strategy to train strong classifiers with localizable features,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, vol. 2019-Octob, pp. 6022–6031, doi: 10.1109/ICCV.2019.00612.

[17] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, “Mixup: beyond empirical risk minimization,” Apr. 2018, doi: 10.48550/arXiv.1710.09412.

[18] D. Liang, F. Yang, T. Zhang, and P. Yang, “Understanding mixup training methods,” *IEEE Access*, vol. 6, pp. 58774–58783, 2018, doi: 10.1109/ACCESS.2018.2872698.

[19] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prügel-Bennett, and J. Hare, “FMix: Enhancing Mixed Sample Data Augmentation,” *arXiv preprint*, Feb. 2020.

[20] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/cvpr.2016.90.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, pp. 4700–4708, doi: 10.1109/cvpr.2017.243.

[22] S. Zagoruyko and N. Komodakis, “Wide residual networks,” in *British Machine Vision Conference 2016, BMVC 2016*, 2016, vol. 2016-Septe, pp. 87.1--87.12, doi: 10.5244/C.30.87.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016, vol. 2016-Decem, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[24] M. Tan and Q. V Le, “EfficientNet: rethinking model scaling for convolutional neural networks,” *arXiv preprint*, May 2019, doi: 10.48550/arXiv.1905.11946.

[25] M. Tan and Q. V Le, “EfficientNetV2: smaller models and faster training,” *arXiv preprint*, 2021.

[26] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, vol. 2017-Octob, pp. 618–626, doi: 10.1109/ICCV.2017.74.
[27] J. Gildenblat *et al.*, “PyTorch library for CAM methods,” *GitHub*, 2021.
[28] S. Narang *et al.*, “Mixed precision training,” in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018, pp. 1–12.

BIOGRAPHIES OF AUTHORS

Talha Anwar is an AI researcher having a Master's degree in Data Science from FAST, National University, Pakistan. He obtained Bachelor's Degree in Biomedical Engineering from Riphah International University in 2018. His research is in biomedical image analysis, biosignal analysis, particularly in the area of brain-computer interface. He has a special interest in social text analysis in the field of NLP. He is equally interested in machine learning and deep learning and has several publications in this domain. Talha is actively involved in research and working with Centre for Chiropractic Research, New Zealand College of Chiropractic, Auckland 1060, New Zealand. All of his research is available at github.com/talhaanwarch. He can be contacted at email: chtalhaanwar@gmail.com.

Seemab Zakir has Bachelor's and Masters's degrees in biomedical engineering from Riphah International University, Pakistan. She has experience in conducting labs on biomedical engineering subjects, particularly programming, machine learning, and instrumentation. She has also served as a biomedical engineer at Pak-Austria Fachhochschule: Institute of Applied Sciences. She was a lecturer at Foundation University School of Science and Technology, Pakistan. Currently, she is a Ph.D. scholar at Scuola Superiore Sant'Anna Pisa, Italy. Her areas of interest are biomedical instrumentation and artificial intelligence. She can be contacted at email: seemabzakir2@gmail.com.