MobileNet-SA: Lightweight CNN with Self Attention for Sketch Classification

Viet-Tham Huynh[1,2(✉)], Trong-Thuan Nguyen[1,2], Tam V. Nguyen[3], and Minh-Triet Tran[1,2]

[1] Software Engineering Laboratory and Faculty of Information Technology, University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
hvtham@selab.hcmus.edu.vn
[3] Department of Computer Science, University of Dayton, Dayton, USA

Abstract. Sketch classification plays a crucial role across diverse domains, including image retrieval, artistic style analysis, and content-based image retrieval. While CNNs have demonstrated remarkable success in various image-related tasks, the computational complexity of large models poses challenges in resource-constrained environments. To address this concern, we propose MobileNet-SA, a novel lightweight model that seamlessly integrates a self-attention module into the MobileNet architecture, with a specific focus on enhancing sketch classification performance. The MobileNet-SA model leverages the inherent efficiency of lightweight CNN while harnessing the power of self-attention mechanisms to effectively capture spatial dependencies and enrich feature representations within sketch data. In our experiments, MobileNet-SA achieves state-of-the-art results, demonstrating an impressive accuracy of 93.5% on the challenging SketchyCOCO dataset and 96.7% on the GM-Sketch dataset. We thoroughly evaluate the model's performance across diverse sketch classes, confirming its robustness and generalization capabilities, which make it well-suited for real-world applications where input sketches may exhibit significant variations. Our research indicates that MobileNet-SA not only outperforms existing methods but also offers an efficient and interpretable solution for sketch classification tasks.

Keywords: Sketch Classification · Convolutional Neural Network · Self-Attention

1 Introduction

Sketch classification is a vital and challenging task in the field of computer vision, with numerous applications spanning various domains, including image retrieval [7,8], artistic style analysis [18], and content-based image retrieval [6,9]. Despite their simplicity, free-hand sketches present a unique set of challenges

W. Q. Yan et al. (Eds.): PSIVT 2023, LNCS 14403, pp. 110–123, 2024.
https://doi.org/10.1007/978-981-97-0376-0_9

for computer vision models due to their abstract nature and sparse representation compared to traditional full-color images. Nonetheless, humans exhibit remarkable proficiency in recognizing sketches almost as effectively as with highly detailed photographs. This exceptional human capability has increased interest in harnessing the potential of sketch analysis for practical applications, particularly with the proliferation of touchscreen devices, which have fostered sketch-based image retrieval and sketch-to-photo synthesis.

Meanwhile, Convolutional Neural Networks (CNNs) [12] have achieved substantial success in various image-related tasks, fueling hopes for similar performance in sketch classification. However, the computational complexity of large CNN models poses significant challenges, especially in resource-constrained environments such as mobile devices and embedded systems. Addressing this concern and finding the right balance between efficiency and accuracy is crucial to unleashing the full potential of sketch classification in real-world scenarios.

In this paper, we introduce our proposed method dubbed MobileNet-SA, a novel and lightweight model tailored explicitly for enhancing sketch classification performance. The cornerstone of MobileNet-SA's architecture is the seamless integration of a self-attention module into the MobileNet framework. MobileNet, known for its computational efficiency, serves as the foundation, while self-attention mechanisms allow the model to capture spatial dependencies and enrich feature representations effectively. This fusion of lightweight architecture and attention-based mechanisms empowers MobileNet-SA to achieve state-of-the-art results in sketch classification while operating efficiently even in resource-constrained settings. Our main contributions in this research are two-fold:

- Firstly, we introduce MobileNet-SA, which provides a compelling solution to the challenges posed by sketch classification. By leveraging the inherent advantages of lightweight CNNs and harnessing the power of self-attention mechanisms, the model successfully overcomes computational complexity issues while improving classification accuracy significantly.
- Secondly, we conduct extensive experiments to evaluate the performance of MobileNet-SA thoroughly. We employ challenging benchmark datasets, including the SketchyCOCO and GM-Sketch datasets to validate their effectiveness. The results demonstrate an impressive accuracy of 93.5% on the SketchyCOCO dataset and 96.7% on the GM-Sketch dataset, surpassing the performance of existing state-of-the-art methods in sketch classification tasks.

Furthermore, we analyze MobileNet-SA's robustness and generalization capabilities, ensuring that it can adapt to diverse sketch classes and handle variations commonly encountered in real-world applications. Additionally, the model's interpretability is a key aspect, allowing users to gain insight into its decision-making process, thereby building trust in its output.

The remainder of this paper is organized as follows. Section 2 provides an overview of related works in sketch classification, highlighting the state-of-the-art approaches and the challenges they address. Section 3 details the architecture and implementation of MobileNet-SA, elucidating the design choices made to enhance its efficiency and classification performance. Section 4 presents the

experimental results and performance analysis of MobileNet-SA on benchmark datasets, demonstrating its superiority over existing methods. Section 5 discusses the implications of our findings, potential applications, and future research directions in sketch analysis and classification.

2 Related Work

In computer vision, sketch-based representations have garnered significant interest and research due to their unique challenges. Yang et al. [19] have extensively studied the abstract nature of free-hand sketches, composed of only a few coarse strokes, yet remarkably recognizable by humans almost as well as full-color photos. This intrinsic human ability has led to the exploration of practical applications, particularly in sketch-based image retrieval and sketch-to-photo synthesis, propelling the development of purpose-made sketch representations that can accommodate their abstract characteristics for various downstream tasks in computer vision.

In the field of fine-grained sketch-based image retrieval (FG-SBIR), Sain et al. [14] have made noteworthy advancements by addressing critical issues faced by the FG-SBIR community. They propose a robust baseline that outperforms existing methods by approximately 11%, achieved through a simple modification to the standard triplet loss. This modification enforces the separation of photo and sketch instances, ensuring a holistic latent space geometry. At the same time, a novel knowledge distillation module leverages photo data during training to overcome the scarcity of sketch data. The proposed approach demonstrates promising results, enhancing the potential of sketches in image retrieval tasks. Chowdhury et al. [2] emphasize the expressive power of human sketches, historically used to convey and record ideas. They introduce a sketch-enabled object detection framework that identifies objects based on corresponding sketches, a novel approach to the fundamental vision task of object detection. Bhunia et al. [1] explore using sketches to detect salient objects in images. Their novel method generates sequential sketch coordinates through a 2D attention mechanism, harnessing attention information inherently embedded in sketches to learn image saliency. Both studies extend the applications of sketches beyond traditional image retrieval and classification, showcasing their versatility in computer vision.

Innovative efforts to transform sketch representations have been made by Lin et al. [10] and Ribeiro et al. [13]. Lin et al. propose the Sketch Gestalt task, considering sketches as sequential data rather than traditional 2D images, aiming to recover masked parts of sketches and complete their shapes. They introduce Sketch-BERT, a generalized version of BERT [3] for the sketch domain. It is trained using self-supervised learning [11] of sketch gestalt, contributing to improved performance in various downstream tasks. Ribeiro et al. introduce Sketchformer, the first-ever Transformer-based network designed to learn deep representations for free-hand sketches, evaluating its efficacy across everyday

sketch interpretation tasks, including Sketch Classification, Generative Sketch Model, and Sketch-based Image Retrieval (SBIR).

Furthermore, Tripathi et al. [16] present the intriguing problem of sketch-guided object localization, where hand-drawn sketches serve as queries to locate instances of objects in natural images. They propose a cross-modal attention scheme, guiding the region proposal network to generate object proposals relevant to the sketch query. The proposed method demonstrates effectiveness with as little as a single sketch query. It generalizes well to object categories not seen during training, showcasing the potential of sketch-guided object localization in practical applications.

In literature, the related work in sketch-based representations illustrates the diverse efforts to explore and utilize sketches in various computer vision applications. The studies discussed above contribute valuable insights, techniques, and methodologies, propelling sketch-based representations to the forefront of computer vision research and opening new avenues for future advancements and practical applications. We present our proposed method in Sect. 3.

3 Proposed Method

In this paper, we introduce a novel approach termed MobileNet-SA, specifically tailored for the classification of sketches, as exemplified in Fig. 1. This approach involves integrating attention modules into the ***MobileNetV2*** architecture to enhance feature representation and object recognition. These ***Self-Attention*** modules are deliberately inserted at specific layers to strategically capture both local and global dependencies within feature maps. This coordinated attention enables the model to effectively discern intricate details and overarching spatial relationships. By dynamically modulating the significance of spatial positions based on contextual relevance, the attention mechanism enhances feature representation, reduces noise, and amplifies visual cues.

3.1 Base Architecture

MobileNetV2 [15] leverages depthwise separable convolutions and inverted residual blocks, offering computational efficiency without compromising accuracy. We briefly introduce its components as below.

Depthwise Separable Convolution consist of a depthwise convolution followed by a pointwise convolution. This factorized convolution significantly reduces the computational cost by decoupling spatial filtering and channel mixing.

Inverted Residual Blocks consist of three main layers: an expansion convolution to increase the number of channels, a depthwise separable convolution to capture spatial features, and a projection convolution to reduce the number of channels back to the original size. The residual connection is used to aid gradient flow during training.

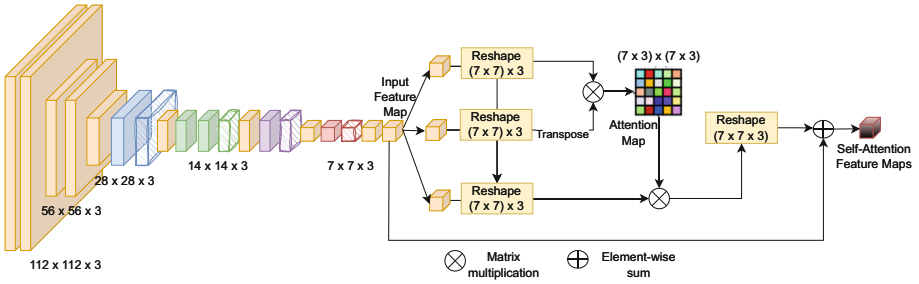

Fig. 1. Overview of the proposed system: attention modules integrated into MobileNetV2 and the self-attention mechanism.

Global Average Pooling spatially downsampled feature maps, resulting in a fixed-size feature vector regardless of the input image size. This pooling operation further reduces the model's computational complexity.

3.2 Self-Attention Module

Self-Attention mechanism draws inspiration from the Transformer's [17] attention mechanism, widely used in natural language processing tasks. The module comprises three essential components: query convolution, key convolution, and value convolution. These components are formulated as follows:

Query, Key, and Value Feature Maps Generation: The F_{in} is passed through three separate convolutional layers: Query Convolution (Q_{conv}), Key Convolution (K_{conv}), and Value Convolution (V_{conv}). The Query Convolution reduces the number of channels to obtain query feature maps Q, the Key Convolution produces key feature maps K with the same reduced channel size, and the Value Convolution retains the original number of channels to obtain value feature maps V.

$$Q = Conv(F_{in}) \in R^{C_q \times H \times W} \tag{1}$$

$$K = Conv(F_{in}) \in R^{C_k \times H \times W} \tag{2}$$

$$V = Conv(F_{in}) \in R^{C_v \times H \times W} \tag{3}$$

Here, C_q, C_k, and C_v represent the reduced number of channels for the query, key, and value feature maps, respectively.

Attention Score Calculation: The attention scores are calculated by performing a matrix multiplication between the query and key feature maps. The resulting matrix contains attention weights that represent the importance of each spatial position in F_{in} with respect to the queries. Additionally, the attention scores are scaled by the square root of the reduced dimension C_q to counteract

vanishing gradient issues. Mathematically, the attention scores (A) are calculated as follows:

$$F_{att} = softmax(\frac{QK^T}{\sqrt{C_q}}) \in R^{H \times W \times H \times W} \tag{4}$$

Attention Application: The attention scores (A) are used to perform element-wise multiplication with the value feature maps (V), resulting in attended feature maps F_{att}. This process allows the model to focus on salient spatial regions in the feature maps.

$$F_{att} = A \odot V \in R^{C_v \times H \times W} \tag{5}$$

Residual Connection: The attended feature maps F_{att} are combined with the original input feature maps F_{in} using a residual connection. The residual connection facilitates gradient flow during training and helps retain essential information from the original feature maps. Mathematically, the output feature maps after applying the self-attention module can be represented as follows:

$$F_{out} = F_{in} + F_{att} \in R^{C_in \times H \times W} \tag{6}$$

3.3 Attention Module Insertion Strategy

We effectively harness the benefits of the self-attention mechanism and attention modules are inserted after specific layers within the MobileNetV2 architecture as illustrated in Fig. 1. The choice of layers for integrating the attention modules is critical and depends on two main considerations:

1. *Desired Level of Spatial Granularity*: Different layers in the MobileNetV2 architecture capture features at varying scales and levels of abstraction. Lower layers tend to capture local details and fine-grained information, while higher layers capture more abstract and global features. By inserting attention modules at different depths in the network, MobileNet-SA can selectively emphasize both local and global spatial relationships within the feature maps.
2. *Computational Budget*: The computational cost of the self-attention module should be taken into account to ensure that the overall model remains efficient and suitable for resource-constrained environments. Placing attention modules after layers that have a manageable number of channels and spatial dimensions helps strike a balance between accuracy and computational efficiency.

In MobileNet-SA, attention modules are thoughtfully positioned to effectively capture both local and global dependencies within the feature maps. Lower-level attention modules are integrated after early layers in the network, concentrating on fine-grained spatial relationships and local contextual information. This strategic placement enables the model to recognize intricate object details and their distinctive features with greater precision. Conversely, higher-level attention modules are inserted after deeper layers in the network to capture

more global dependencies and contextual relationships. These attention modules empower the model to comprehend the overall spatial layout and interactions among different objects in the scene, thus enhancing object recognition within their contextual context. The strategic combination of local and global attention mechanisms results in MobileNet-SA achieving a more comprehensive and robust representation of visual information, leading to improved object recognition performance compared to the baseline MobileNetV2 model.

We strategically leverage attention modules to selectively focus on relevant regions while suppressing noise, which leads to improved feature representation with accentuated visual cues and reduced distractions. The self-attention mechanism dynamically adjusts the importance of spatial positions within the feature maps based on their contextual relevance, allowing the model to learn discriminative features that are highly informative for object recognition. This adaptability enhances the model's robustness to variations in object appearance and background clutter, resulting in higher recognition accuracy. Consequently, MobileNet-SA exhibits promising potential as an architecture for various computer vision tasks, offering an effective solution for tasks that demand enhanced feature representation and improved object recognition performance.

4 Experiments

4.1 Benchmark Datasets

Regarding the evaluation, we employed the SketchyCOCO dataset [4] and GM-Sketch dataset [5] to evaluate and validate our proposed method, MobileNet-SA. The SketchyCOCO dataset and GM-Sketch dataset serve as the foundation for testing the effectiveness and performance of our approach in automatic image generation from freehand sketches.

SketchyCOCO Dataset. To conduct the experiments, we used the MobileNet-SA architecture, which incorporates the Self-Attention Module strategically integrated into the MobileNetV2 architecture. This attention module enables the model to selectively focus on relevant regions and capture both local and global dependencies within the feature maps, leading to improved feature representation and enhanced object recognition performance (Table 1).

By leveraging the SketchyCOCO dataset in combination with our proposed MobileNet-SA method, we aimed to demonstrate its ability to generate high-quality images in response to specific requirements from freehand sketches. The dataset's unique composition of sketch and photo pairs allows us to assess the model's capability to generate visually appealing images based on the user's intuitive freehand sketches.

In this work, in order to effectively learn sketch features, we focused solely on utilizing the data present in the "Sketch" directory for both training and testing purposes. We did not use any data from the "gt" and "edge" directories. Our approach solely relied on the information contained in the "Object" directory,

Table 1. Statistics of subsets used in SketchyCOCO dataset.

Name	dog	giraffe	plane	sheep	motobike	car	cat
Train set	2,902	2,065	1,848	1,095	1,145	1,094	3,917
Test set	80	232	168	111	21	55	7
Name	trafficLights	elephants	bicycle	cow	zebra	horse	fireHydrant
Train set	481	1,395	249	1,694	2,091	1,851	892
Test set	27	104	15	156	189	143	32

including object-specific details and background information, while excluding other data or information not utilized in this research. By concentrating on the "Sketch" directory and the pertinent object-related information, we aimed to develop a robust model for sketch feature learning.

By using the SketchyCOCO dataset alongside our proposed method, we sought to establish the practical applicability and superiority of MobileNet-SA for automatic image generation tasks. The combination of the dataset and the proposed method aimed to contribute significantly to the advancement of image generation techniques, particularly in the context of freehand sketches (Fig. 2).

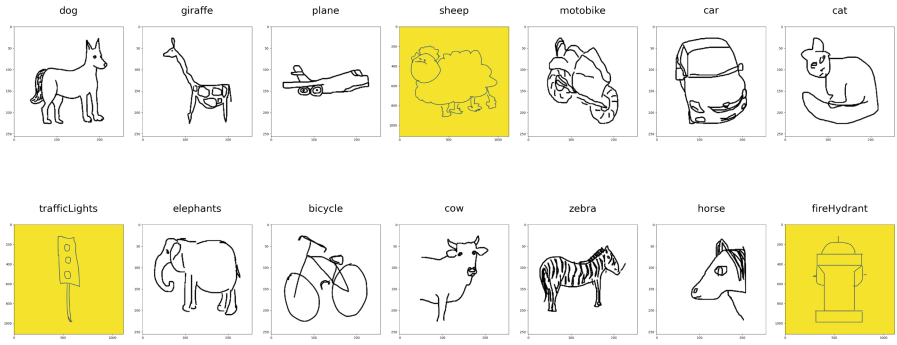

Fig. 2. Samples for each category in the training set of SketchyCOCO dataset.

GM-Sketch Dataset. In addition to testing on SketchyCOCO dataset, we also test our method on GM-Sketch dataset. GM-Sketch dataset is a comprehensive and diverse collection of sketch images carefully curated for artificial intelligence applications. Developed through a meticulous process, this dataset encompasses a wide array of objects sourced from various platforms, including Google, to ensure a rich variety of angles and shapes. The primary focus of GM-Sketch dataset lies in household items, furniture, animals, and vehicles, making it suitable for a multitude of use cases (Table 2).

Table 2. Statistics of subsets used in GM-Sketch dataset.

Name	dog	laptop	banana	clock	bed	car
Train set	150	150	150	150	150	150
Test set	150	150	150	150	150	150
Name	cat	cup	sofa	bicycle	book	bottle
Train set	150	150	150	150	150	150
Test set	150	150	150	150	150	150

To maintain the dataset's high quality, rigorous filtering was applied, eliminating low-quality images and ensuring only the finest representations were included. Additionally, GM-Sketch dataset has been enriched using MidJourney, an advanced AI image generator. Leveraging deep learning, MidJourney empowers users to create unique and captivating images, making it an invaluable resource for designers and creatives seeking visually engaging content.

GM-Sketch dataset comprises 1800 sketch images, thoughtfully organized into 12 categories, each containing 150 images. These categories have been further grouped into Furniture, Personal Belongings, Animals, Vehicles, and Others, facilitating easy access and navigation for various research and application needs.

For utmost accuracy and reliability, the dataset has been thoughtfully split into training and testing sets with an 80:20 ratio, ensuring consistent and unbiased evaluations for AI model development and analysis (Fig. 3).

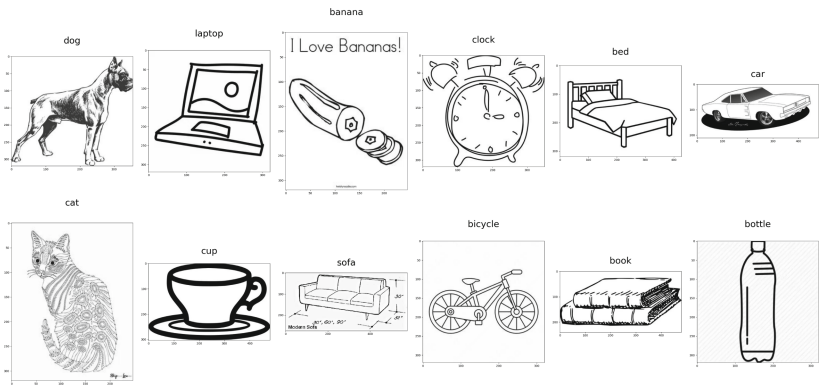

Fig. 3. Samples for each category in trainset of GM-Sketch dataset.

4.2 Implementation Details

We utilize the `PyTorch` library, along with `torchvision`, to build and train the model. `AdamW` and `CrossEntropyLoss` are employed as the optimizer and loss function for model training. During training, the script logs the training loss to `TensorBoard` for visualization and periodically saves checkpoints of the model to track its progress. We also leverage the best-performing model based on the lowest achieved loss on the validation set.

4.3 Evaluation Metrics

Accuracy is a classification performance metric that quantifies a model's capability to correctly predict instances in a dataset. *Precision*, a performance measure, assesses the model's accuracy in identifying positive instances among all instances predicted as positive. *Recall*, also known as sensitivity, evaluates the model's ability to identify all positive instances from the total actual positive instances in the dataset. *F1-score*, a composite metric, takes into consideration both precision and recall, providing a balanced evaluation of the model's effectiveness, especially in datasets with imbalanced class distributions.

4.4 Experimental Results

We dedicate to exploring the utilization of lightweight CNN models, namely ShuffleNetv2, MobileNetV2, ResNet50, and EfficientB0, in conjunction with self-attention mechanisms to enhance the performance of sketch classification. The integration of self-attention is particularly advantageous with lightweight CNN models owing to their inherent parameter efficiency, ability to effectively capture spatial information, and improvements in feature representation. This combination is sought to achieve an optimal equilibrium between model complexity and classification performance.

Table 3. Comparison of lightweight CNN models with Self-Attention (SA) on ***Sketchy-COCO dataset***. The best result is highlighted.

Method	Accuracy	Precision	Recall	F1-score
MobileNet-SA (Ours)	93.5%	91.9%	94.0%	92.1%
EfficientNetB0-SA (Ours)	92.0%	91.5%	94.2%	92.3%
ResNet50-SA (Ours)	91.3%	91.0%	92.1%	90.9%
MobileNetv2 [15]	89.7%	89.2%	92.0%	89.5%
ShuffleNetv2 [20]	90.4%	94.4%	89.2%	90.2%

Our empirical study in Table 3 demonstrates the MobileNet-SA model's superiority over other models in terms of accuracy. This enhanced performance is

primarily attributed to the integration of self-attention within the MobileNetv2 architecture. By incorporating self-attention, the model effectively captures spatial dependencies inherent in sketches, allowing for the identification of intricate patterns and relationships among different sketch elements. This capability proves essential for achieving accurate sketch classification and generating corresponding images. Notably, the parameter efficiency characteristic of the MobileNet architecture remains unaffected despite the inclusion of self-attention, ensuring efficient sketch processing while leveraging the benefits of enhanced feature representation.

MobileNet-SA achieves a notable Accuracy of 93.5%, which signifies a well-balanced trade-off between precision and recall. This balance holds significant importance for tasks like automatic image generation, as it enables the model to accurately identify positive instances with high precision while avoiding the exclusion of critical sketch details with high recall.

Additionally, MobileNet-SA exhibits robustness and generalization capabilities across the 14 diverse sketch classes found in the SketchyCOCO dataset. Its ability to accommodate various sketching styles and object categories is of considerable value for real-world applications, where input sketches may exhibit substantial variations. Moreover, the attention mechanism employed in MobileNet-SA contributes to the model's interpretability and explainability, rendering it highly suitable for tasks that demand transparent decision-making processes.

Similarly, the EfficientNetB0 architecture with self-attention achieves competitive performance, attaining an Accuracy of 92.0%. Notably, it demonstrates high Precision (91.5%) and Recall (94.2%), leading to an impressive F1-score of 92.3%. These results indicate its ability to effectively minimize both false positives and false negatives. Furthermore, the ResNet50-SA model delivers robust classification performance with an accuracy of 91.3%. Although it slightly trails behind MobileNet-SA and EfficientNetB0-SA in terms of accuracy, it consistently exhibits favorable outcomes across various evaluation metrics.

Table 4. Comparison of MobileNet-SA and MobileNetv2 on the ***GM-Sketch dataset***. The best result is highlighted.

Method	Accuracy	Precision	Recall	F1-score
MobileNet-SA (Ours)	96.7%	96.7%	96.7%	96.7%
MobileNetV2[5]	96.1%	96.1%	96.1%	96.1%

In our comparative analysis, we evaluate the MobileNet-SA model with two baseline architectures, MobileNetv2 and ShuffleNetv2. The standard MobileNetv2 architecture achieves an accuracy of 89.7%, displaying stable and commendable results across all evaluation metrics. However, its lack of self-attention may limit its capacity to capture complex spatial dependencies in sketches. On the other hand, the standard ShuffleNetv2 architecture achieves an accuracy of 90.4%, demonstrating notable performance, especially in precision (94.4%), indicating its effectiveness in reducing false positive classifications.

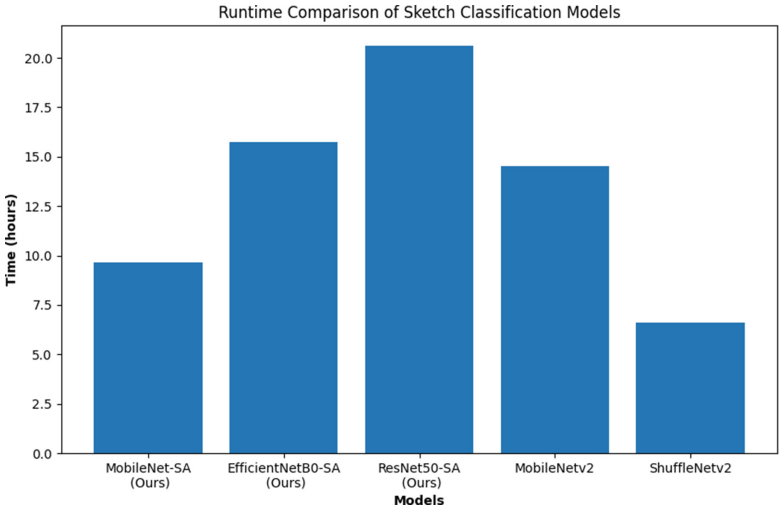

Fig. 4. Runtime Comparison of Sketch Classification Models

From the runtime results, it can be observed that our proposed method - MobileNet-SA, achieves faster execution time compared to the standard MobileNetV2, with a reduction of approximately 33.7%. Specifically, the runtime of MobileNet-SA is 9.635 h compared to 14.54 h for MobileNetV2 (Fig. 4). Despite the significant reduction in runtime, the accuracy of MobileNet-SA is not only unaffected but also shows a notable increase. Specifically, the accuracy improves by 3.8% (Table 3).

Furthermore, MobileNet-SA achieves 96.7% in terms of accuracy on the GM-Sketch dataset, with congruent values for precision, recall, and F1-score (Figure 4). This consistent performance across all evaluation metrics indicates the model's proficiency in accurately classifying sketch images while maintaining a harmonious trade-off between precision (the capacity to correctly identify positive instances) and recall (the capability to avoid overlooking significant instances).

5 Conclusion

In this paper, we have presented MobileNet-SA, a novel lightweight model that seamlessly integrates a self-attention module into the MobileNet architecture to enhance sketch classification performance. Through extensive experiments on challenging benchmark datasets, we have demonstrated the superiority of MobileNet-SA over other models, achieving state-of-the-art results with an impressive accuracy of 93.5% on the SketchyCOCO dataset and 96.7% on the GM-Sketch dataset. The integration of self-attention within the MobileNetv2 architecture has proven crucial in capturing spatial dependencies within sketches,

enabling accurate classification and generating corresponding images. Additionally, MobileNet-SA retains the parameter efficiency characteristic of MobileNet, making it an efficient solution for resource-constrained environments.

Furthermore, MobileNet-SA exhibits robustness and generalization capabilities across diverse sketch classes, showcasing its potential for real-world applications where input sketches may vary substantially. Its efficient yet powerful architecture, leveraging the benefits of self-attention, holds great promise for various practical applications and advances in the field of sketch analysis.

Acknowledgment. This research was funded by Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2019.DA19.

References

1. Bhunia, A.K., et al.: Sketch2Saliency: learning to detect salient objects from human drawings. In: CVPR (2023)
2. Chowdhury, P.N., et al.: What can human sketches do for object detection? In: CVPR (2023)
3. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186 (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
4. Gao, C., et al.: SketchyCOCO: image generation from freehand scene sketches (2020). arXiv: 2003.02683 [cs.CV]
5. Huynh, V.-T., Nguyen, and T.V., Tran, M.-T.: Light-weight sketch recognition with knowledge distillation. In: International Conference on Multimedia Analysis and Pattern Recognition (2023)
6. Kim, H., Yeo, C., Cha, M., Mun, D.: A method of generating depth images for view-based shape retrieval of 3D CAD models from partial point clouds. Multimedia Tools Appl. **80**, 10859–10880 (2021)
7. Le, T.-N., et al.: SketchANIMAR: sketch-based 3D animal fine- grained retrieval. Comput. Graphics (2023)
8. Le, T.-N., et al.: TextANIMAR: text-based 3D animal fine-grained retrieval. Comput. Graph. (2023)
9. Li, Z., Junyu, X., Zhao, Y., Li, W., Nie, W.: MPAN: multipart attention network for point cloud based 3D shape retrieval. IEEE Access **8**, 157322–157332 (2020)
10. Lin, H., Fu, Y., Jiang, Y.-G., Xue, X.: Sketch- BERT: learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt (2020). arXiv: 2005.09159 [cs.CV]
11. Liu, X., et al.: Self-supervised learning: generative or contrastive. en. In: arXiv:2006.08218 [cs, stat] (July 2020). arXiv:2006.08218. http://arxiv.org/abs/2006.08218 (visited on 10/30/2020)
12. O'Shea, K., Nash, R.: An introduction to convolutional neural networks (2015). arXiv: 1511.08458 [cs.NE]
13. Ribeiro, L.S.F., Bui, T., Collomosse, J., Ponti, M.: Sketchformer: transformer-based Representation for Sketched Structure. In: Proceedings of the CVPR (2020)

14. Sain, A., et al.: Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR (2023). arXiv: 2303.13779 [cs.CV]
15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.-C.: MobileNetV2: inverted residuals and linear bottlenecks (2019). arXiv: 1801.04381 [cs.CV]
16. Tripathi, A., Dani, R.R., Mishra, A., Chakraborty, A.: Sketch-guided object localization in natural images. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12351, pp. 532–547. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58539-6_32
17. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
18. Xie, X., et al.: Artistic style discovery with independent components. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19870–19879 (2022)
19. Yang, L., Pang, K., Zhang, H., Song, Y.-Z.: SketchAA: abstract representation for abstract sketches. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10077–10086 (2021). https://doi.org/10.1109/ICCV48922.2021.00994
20. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: an extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856 (2018)