MaskTune: Mitigating Spurious Correlations by Forcing to Explore

Syed Saad Hasan
(DIAG)
Sapienza University of Rome
Rome, Italy
Email: hasan.2106512@uniroma1.it

Simone Scardapane
(DIAG)
Sapienza University of Rome
Rome, Italy
Email: simone.scardapane@uniroma1.it

Abstract—**This article examines techniques for mitigating deceptive patterns in deep learning models, drawing inspiration from a novel approach known as MaskTune. MaskTune is a novel solution to tackle the problem of models having an excessive number of parameters in comparison to the available training data. It aids the models in acquiring significant patterns without excessively fitting to insignificant details. MaskTune employs a masking technique to mitigate the model's over-reliance on a limited number of characteristics. By concealing certain characteristics, the model is compelled to uncover new ones. This masking process is implemented in a singular fine-tuning phase, which is a crucial step in adjusting a pre-trained model to a different task or dataset. This methodology reduces both time and computational resources required, while enhancing task performance in comparison to training a novel model from the beginning. In addition, a job of selective categorization was developed, using MaskTune's capability to promote resilient learning that is less reliant on deceptive characteristics. This enables the model to identify the absence or concealment of crucial characteristics and prevent generating incorrect forecasts. To assess the efficacy of MaskTune in this job of selective categorization, precise metrics were used to measure the accuracy of predictions made by the model and its capacity to abstain from making choices when there was insufficient information for dependable predictions.**

I. Introduction

In the realm of deep learning, a critical challenge to creating highly generalizable models is the presence of spurious correlations within training datasets. These correlations, often stemming from biases in data selection, can lead over-parameterized models to over-rely on irrelevant input features, thus undermining their performance on new data. This project seeks to address this issue through the implementation and adaptation of MaskTune, using the well-known CIFAR-10 and CelebA datasets, alongside two neural network models: the Visual Geometry Group (VGG) network and the Residual Network (ResNet50). MaskTune embodies a state-of-the-art approach designed to enhance the models' ability to identify and prioritize new significant features while reducing their reliance on deceptive ones. Furthermore, this project explores MaskTune's application in the context of selective classification, a technique that allows models to abstain from making predictions in the absence of sufficient data for reliable judgment. This approach ensures decisions are made only when there is concordance between the predictions of the pre-trained model and the fine-tuned model, thereby mitigating the effects of spurious correlations and fostering a more discerning and critical information processing operation by the models.

II. Method

In this context, our efforts focused on evaluating MaskTune's effectiveness in two specific domains: the CIFAR-10 and CelebA datasets, and the application to two distinct neural architectures, a custom VGG and a pre-trained ResNet50. This selection was driven by the necessity to assess MaskTune's impact across scenarios with varying complexity levels and types of spurious correlations.

A. Datasets Used

1) CIFAR-10: A widely recognized dataset comprising 60,000 color images of 32x32 pixels, categorized into 10 classes, with 6,000 images per class. The classes represent airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships, and trucks. It is commonly used to evaluate image classification and recognition algorithms, presenting challenges such as the low resolution of images and the variety of angles and backgrounds within the same classes. Despite its simplicity, CIFAR-10 poses difficulties in achieving high classification accuracy due to the small size of the images and intra-class variability.

2) CelebA (CelebFaces Attributes Dataset): A comprehensive dataset of celebrity face images, containing over 200,000 images, each annotated with 40 attributes (e.g., ”young”, ”glasses”, ”smile”) and five landmark positions (specific predefined points on a face corresponding to notable facial features). It is used for tasks ranging from facial attribute classification and face recognition to face generation and style transfer. The richness of attributes and diversity of images make CelebA ideal for exploring complex facial representations. The main challenges include handling the variety and nuances of facial attributes, as well as training models capable of generalizing across a wide spectrum of facial expressions and configurations.

III. Input Masking

A key aspect of our approach involves employing a masking function, applied once the model has been fully trained. Our objective is to create a new masked dataset by concealing the most discriminative features identified by the model after its complete training. This goal is pursued through attention-based masking, a methodology that uses attention mechanisms to pinpoint the parts of the input deemed most relevant or discriminative by the neural network during training. Attention-based masking follows these steps:

A. Importance Analysis

Initially, the model uses its internal attention mechanisms to assess which parts of the input contribute most to the classification decision. These mechanisms assign relative weights to different input areas or features, identifying those the model perceives as most crucial.

B. Feature Selection for Masking

Based on the importance analysis, the features or areas receiving the highest weights are selected for masking. This process operates on the hypothesis that concealing the most informative parts of the input forces the model to seek and rely on less obvious or previously overlooked features.

C. Application of Masking

The selected features are then effectively masked, making them invisible or altering them to reduce their importance for the model's decisions. This can be achieved through various methods, such as applying black masks, adding noise, or using specific distortion techniques that render these features unrecognizable by the network.

D. Fine-Tuning of the Model

With the new masked dataset, the model undergoes a fine-tuning phase, allowing it to adapt and learn to recognize and utilize new features or parts of the input that were not previously considered critical for classification.

It is important to note that the attention-based masking strategy was adopted in place of the xGrad-CAM (Gradient-weighted Class Activation Mapping) technique initially proposed in the original paper. xGrad-CAM highlights image areas contributing most to the model's classification decision by using gradients of the target concept flowing into the final convolutional layer to produce a coarse localization map. This map visually emphasizes the significant areas for predicting the class label, allowing for an intuitive understanding of the model's decision-making process. By overlaying this heatmap onto the original image, the relevant features or parts of the image become clear. Although xGrad-CAM provides valuable insights, we opted for attention-based masking for its direct and flexible approach in selectively hiding key features. This choice aims to enhance the model's generalization, encouraging it to discover and leverage new dataset attributes, thereby reducing dependence on potentially misleading or overly specific features from the original training set.

E. Adapting MaskTune for Selective Classification

In the context of the project that utilizes MaskTune for selective classification, an approach has been developed that improves the reliability of predictions in deep learning models through a decision-making process that requires agreement between two independent models: one initially trained (model_initial) and the other optimized through fine-tuning (model_finetuned). A prediction is considered valid only if both models agree on the same label and if the prediction of the fine-tuned model exceeds a predetermined confidence threshold. Furthermore, for classes considered more difficult to predict (for example, 'bird', 'cat', 'deer', 'dog' in the CIFAR10 dataset), differentiated weights are applied to the confidence thresholds to increase tolerance in predictions, aiming for an optimal balance between accuracy and rejection rate. Finally, the system's performance is evaluated based on the accuracy of the accepted predictions, the rejection rate, and the agreement rate between the models, providing an overall analysis of its effectiveness and consistency.

IV. Implementation Details and Achieved Results

A. Single-label Classification with Spurious Features

The model employed is a streamlined version of VGG designated as enhancedVGG, featuring an optimized architecture aimed at enhancing feature extraction and classification accuracy. The enhancedVGG model comprises two primary groups of convolutional layers. The first group processes the input from 3 to 32 channels, while the second group further refines it from 32 to 64 channels. Each convolutional layer is followed by a ReLU activation function to introduce non-linearity and improve the model's ability to learn complex patterns. Subsequent to each group of convolutional layers, max-pooling is applied to reduce spatial dimensions and retain crucial features. This pooling step aids in reducing computational complexity and mitigates overfitting by providing spatial invariance. The enhancedVGG model integrates three fully connected layers after the convolutional and pooling layers, progressively reducing the dimensionality to correspond with the 10 classes in the CIFAR-10 dataset. The incorporation of attention mechanisms within the enhancedVGG architecture allows the model to focus on the most relevant parts of the input during training. These attention layers assign varying weights to different features, thereby enhancing the model's capacity to identify and utilize the most informative aspects of the data. This focus on key features not only boosts classification accuracy but also aids in identifying spurious correlations that might mislead the model. In our experiments, the enhancedVGG model exhibited superior performance on the CIFAR-10 dataset compared to the standard VGG architecture. The attention mechanisms enabled the model to achieve higher accuracy by effectively learning to disregard spurious features and concentrate on the most discriminative parts of the images. The results indicate that the enhancedVGG model is robust and efficient in handling datasets with potential spurious correlations, making it a valuable tool for single-label classification tasks.

B. Training Phases

1) Initial Training: The model is trained on the CIFAR-10 dataset without applying masking. This phase establishes a performance baseline for the model.

2) Fine-Tuning: After the initial training, the model undergoes further optimization by applying masking to the input images, compelling it to explore new features beyond those extracted during the initial training phase. The images undergo preprocessing through normalization to ensure that each pixel value is appropriately scaled for effective training. Initially, the Stochastic Gradient Descent (SGD) optimizer is utilized with a learning rate of 0.001, momentum of 0.9, and a batch size of 4, over 12 epochs. This combination of hyperparameters helps efficiently navigate the loss landscape, promoting both speed and stability during the initial training phase. For the fine-tuning phase, the learning rate is reduced to 0.0001. This lower learning rate is crucial as it facilitates a more stable and gradual convergence towards a local optimum, allowing the model to refine its weights without the risk of overshooting minima. This fine-tuning step is essential for enhancing the model's performance, ensuring better generalization on unseen data and improving predictive accuracy. By employing this training regimen, the enhancedVGG model effectively learns and adapts, resulting in robust performance on the CIFAR-10 dataset. The combination of normalization, an initial higher learning rate with SGD, and a reduced learning rate for fine-tuning contribute to the model's ability to achieve high classification accuracy while maintaining stability and convergence throughout the training process. The use of the ”mps” device in the code is intended to leverage Metal Performance Shaders (MPS), a collection of high-performance shaders and image processing functions optimized to fully utilize the GPU of Apple devices. The code was tested on a MacBook Pro with the following specifications:

- **Model:** MacBook Pro (13-inch, 2019)
- **Processor:** 1.4 GHz Quad-Core Intel Core i5
- **Graphics:** Intel Iris Plus Graphics 645 1536 MB
- **Memory:** 8 GB 2133 MHz LPDDR3
- **Operating System:** macOS Sonoma 14.5

Additionally, the development environment used was PyCharm 2022.2.5 (Community Edition), with the following details:

- **Build:** #PC-222.4554.11, built on March 15, 2023
- **Runtime Version:** 17.0.6+7-b469.82 x86_64
- **VM:** OpenJDK 64-Bit Server VM by JetBrains s.r.o.

These hardware and software configurations were chosen to ensure optimal performance and compatibility with the MPS framework, thereby enhancing the efficiency and speed of deep learning model training and inference on macOS.

V. Experimental Results

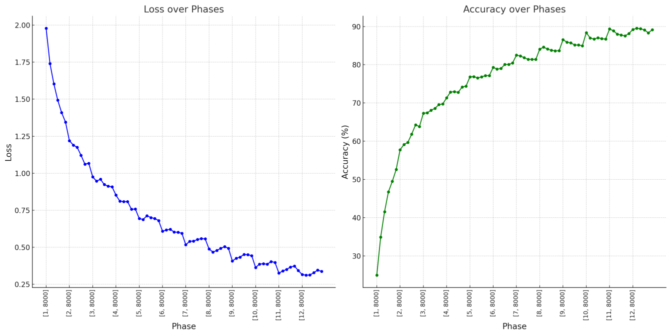

Fig. 1. Training Results

- **Loss over Phases:** The left graph shows a steady decrease in loss from approximately 2.0 to 0.25 over 12 training phases, indicating effective model learning and convergence.
- **Accuracy over Phases:** The right graph demonstrates an increase in accuracy from around 30% to 90% over the same 12 training phases, reflecting the model's improved predictive performance.

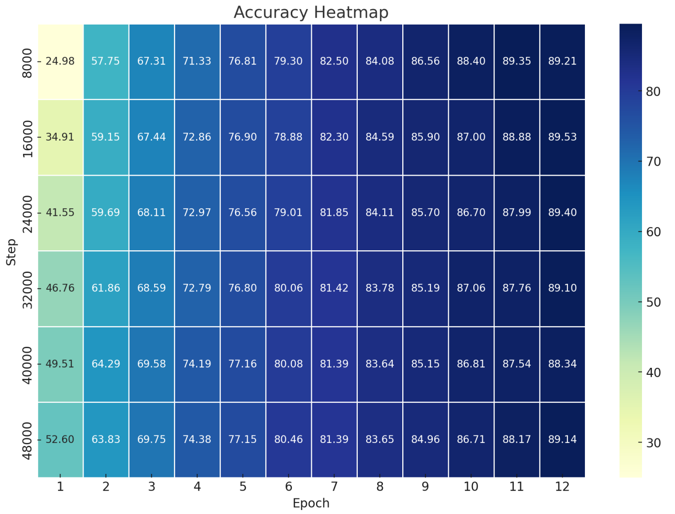

Fig. 2. Detailed Accuracy Heatmap

- The heatmap shows accuracy across epochs (x-axis) and steps (y-axis). Darker blues indicate higher accuracy, while lighter colors indicate lower accuracy. The heatmap quickly visualizes accuracy trends, making it easier to compare performance across different training stages.
- The graph shows the accuracy (blue line, left y-axis) and loss (red line, right y-axis) over various phases during fine-tuning. As the phases progress, accuracy increases from 92.2% to 93.7%, while loss decreases from 0.23 to 0.19, indicating improved model performance and convergence.
- This bar graph compares the accuracy of the model on different classes in the test set with and without masking. Blue bars represent accuracy with masking, while red bars show accuracy without masking. The

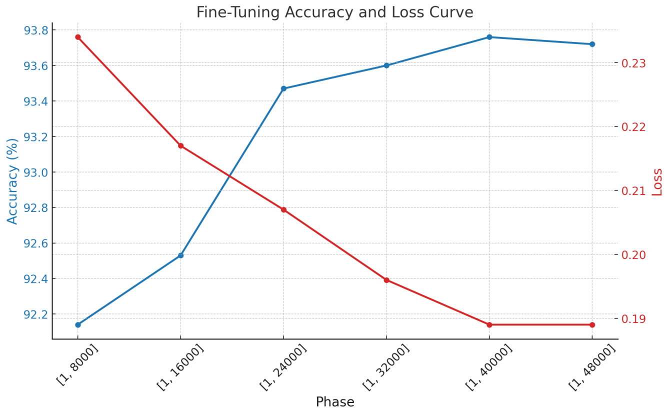

Fig. 3. Fine-tuning results

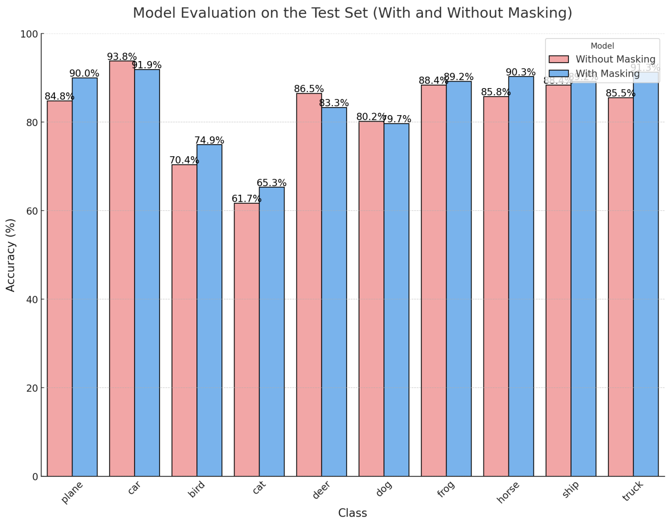

Fig. 4. Model evaluation on the test set with and without masking

graph indicates that masking generally improves accuracy across most classes, demonstrating the effectiveness of the masking technique in enhancing model performance and robustness in classification tasks.

Metric	Value (%)
Accuracy (excluded rejections)	90.94%
Rejection rate	6.91%
Agreement rate	91.27%

TABLE I
MODEL PERFORMANCE MATRIX

VI. ANALYSIS OF EXPERIMENTAL RESULTS

Analyzing these experimental results reveals several significant insights about the effectiveness of the enhancedVGG model, particularly regarding the use of the masking technique and its generalization capability on a test set.

- **Performance Improvement with Fine-Tuning:** Comparing the initial training results with those of the fine-tuning phase shows a significant improvement in both loss and accuracy. This suggests that applying masking during the fine-tuning phase enabled the model to explore and emphasize new, important features for classification compared to those learned during the initial training phase.
- **Accuracy Increase per Class with Masking:** Evaluating the model on the test set with and without masking reveals an accuracy increase for each class. This demonstrates the effectiveness of masking in enhancing the model's ability to distinguish between different classes, thereby making the model more robust and reliable in image classification.
- **Effectiveness of Selective Classification:** The results of selective classification show high accuracy (90.94%) when rejections are excluded, along with a relatively low rejection rate (6.91%) and a high agreement rate (91.27%). This indicates that the model can accurately identify when predictions might not be reliable and choose not to classify those images, thus increasing the overall reliability of the classification system.

A. Multi-label Classification with Spurious Features

For this task, a pre-trained model based on the convolutional neural network ResNet50 was utilized without applying any masking mechanism. ResNet50, part of the Residual Networks (ResNets) family introduced by Kaiming He et al. in 2015, was designed to overcome the vanishing gradient problem in deep networks through the use of residual connections. This preliminary phase aims to establish a solid knowledge base for the model, allowing it to learn the general visual features present in the CelebA dataset.

1) Fine-Tuning with Masking: Subsequently, the model was fine-tuned using a modified version of ResNet50, named AttentionMaskingResNet50, which integrates a mechanism for generating and applying masks to the extracted features. This training phase aims to compel the model to focus on new features of the images by masking those previously considered decisive for classification. The core of the AttentionMaskingResNet50 model consists of a mask generator, which acts directly on the output of the features extracted from the pre-trained network. This module, through the use of convolutional layers followed by sigmoid activation functions, produces dynamic masks that modulate the importance attributed to each region of the image. By applying these masks before the classification layer, the model is directed to preferentially consider those areas of the image previously neglected, thus enriching its discrimination capacity. Transferring weights from the ResNet50 model, trained in the initial phase, to AttentionMaskingResNet50 ensures the preservation of acquired knowledge, optimizing the learning process towards the discovery of new significant visual patterns.

2) Selective Classification: As observed with the enhanced-VGG model, the selective classification approach was extended to ResNet50, employing a similar methodology. This process includes loading the pre-trained models, comparing their respective predictions on a selected test dataset, and using a confidence threshold to filter out unreliable predictions. The main metrics for evaluating the effectiveness of this method include the accuracy of predictions on which the models agree, the percentage of predictions discarded due to lack of reliability, and the frequency of agreement between the two models. This approach underscores the value of attention mechanisms in refining the accuracy and reliability of deep learning models, highlighting the importance of advanced classification strategies in significant application contexts.

VII. Conclusion

Despite the innovative approach and advanced techniques employed in the project, a significant challenge arose during the evaluation phase of the ResNet50 network. Due to the task's intrinsic complexity and computational power limitations, we opted to test the model's functionalities on a reduced subset of the original CelebA dataset. This decision aimed to efficiently manage available resources while striving to maintain an accurate and meaningful level of analysis. However, despite efforts to optimize and adapt the learning and evaluation process to the hardware constraints, it was not possible to obtain metric values that would allow a conclusive and satisfactory evaluation of the proposed solution. In summary, while the implementation of attention mechanisms and selective classification techniques presents a potentially effective strategy for enhancing the reliability and precision of deep learning models, a comprehensive evaluation of its effectiveness remains, at this moment, an unachieved goal.

References

[1] Abbasian, K., Laroche, S., & Conati, C. (2022). *MaskTune: Mitigating Spurious Correlations by Forcing to Explore*. arXiv preprint arXiv:2210.00055. Available at: magentahttps://arxiv.org/abs/2210.00055

[2] Cabrera-Vives, G., Martinez-Palomera, J., Huertas-Company, M., & Dominguez, H. (2023). *Mitigating Bias in Deep Learning: Training Unbiased Models on Biased Data for the Morphological Classification of Galaxies*. arXiv preprint arXiv:2308.11007. Available at: magentahttps://arxiv.org/abs/2308.11007

[3] Zhang, Q., Wu, Y., & Sun, J. (2021). *Detecting and Mitigating Encoded Bias in Deep Learning-Based Systems*. Neural Computing and Applications. Available at: magentahttps://link.springer.com/article/10.1007/s00521-021-06337-3