MaskTune: Mitigating Spurious Correlations by Forcing to Explore

Autor:
Gianmarco Donnesi
Matr. n. 2152311

Supervisor:
Prof. Scardapane

Abstract

The work analyzed in this report focused on investigating and applying strategies to mitigate spurious correlations in deep learning models, inspired by the recent development of a method called *MaskTune*. This represents an innovative approach to tackle a fundamental challenge in *over-parametrized* models (i.e., models that have a very high number of parameters compared to the number of examples in the training dataset): learning meaningful data representations that produce good performance on a downstream task without overfitting to spurious input features. This method proposes a masking strategy that prevents excessive dependency on a limited number of features, forcing the model to explore new ones by masking those previously discovered. To do this, masking is applied during the fine-tuning of a single epoch. This is a technique for adapting a pre-trained model to a new task or dataset, which allows leveraging the model's existing knowledge, reducing the time and computational resources needed for training, and improving performance on specific tasks compared to training a model from scratch. Finally, an additional selective classification task was implemented, exploiting MaskTune's ability to promote the learning of more robust representations less dependent on potentially misleading or unreliable features. This would allow the model to recognize situations where the main informative features are absent or masked, opting to abstain from classification rather than risking an inaccurate prediction. To measure the effectiveness of MaskTune in the selective classification task, specific metrics were used to evaluate both the accuracy of the predictions, when the model decides to make them, and the ability to abstain from making decisions when the available information is not sufficient for a reliable prediction.

1 Introduction

In the current landscape of deep learning, a significant obstacle to creating highly generalizable models is the presence of spurious correlations in training datasets. These correlations, often resulting from bias in data selection, can lead over-parameterized models to excessively rely on irrelevant input features, compromising their performance on new data. This project aims to overcome this challenge through the implementation and adaptation of MaskTune, using the well-known CIFAR-10 and CelebA datasets and employing two neural network models: VGG and ResNet50. MaskTune represents a cutting-edge approach aimed at improving the models' ability to identify and value new significant features, decreasing their reliance on deceptive ones. Additionally, the project investigated the use of MaskTune in the context of selective classification, a technique that allows models to refrain from making predictions in the absence of sufficient data for reliable judgment. This approach is based on the principle that a decision is made only when there is agreement between the predictions of the pre-trained model and the one refined through fine-tuning. Using this technique, the aim is not only to mitigate the effect of spurious correlations but also to encourage a more aware and critical information processing operation by the models.

2 Methods

In this context, we have focused our efforts on exploring the effectiveness of MaskTune in two specific areas: the use of the CIFAR-10 and CelebA datasets, and the application to two distinct neural architectures, a custom VGG and a pre-trained ResNet50. This choice was motivated by the need to assess the impact of MaskTune in scenarios with different levels of complexity and types of spurious correlations.

2.1 Datasets used

CIFAR-10. A widely recognized dataset consisting of 60,000 color images of size 32x32, divided 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. It offers a variety of challenges for deep learning models, including the relatively low resolution of the images and the variety of angles and backgrounds for objects within the same classes. Despite its apparent simplicity, CIFAR-10 can present difficulties in achieving high classification accuracy due to the small size of the images and the intra-class variety.

CelebA (CelebFaces Attributes Dataset). A large dataset of celebrity face images, containing more than 200,000 images of celebrities, each with 40 *attribute annotations* (such as "young", "glasses", "smile") and five *landmark positions* (specific predefined points on a face that correspond to notable facial features. These landmarks are used to capture the geometry and structure of faces within the images). It is used for tasks ranging from facial attribute classification and face recognition to face generation and style transfer. The richness of attributes and the 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 that are capable of generalizing well across a wide spectrum of facial expressions and configurations.

2.2 Input masking

A key aspect of our approach involves using a masking function, applied once the model has been fully trained. Our aim is to generate a new masked dataset by concealing the most discriminative features identified from the input by the model after its complete training. This goal was pursued through *attention-based masking*, a methodology that employs attention mechanisms to identify the parts of the input considered most relevant or discriminative by the neural network during training.
Attention-based masking operates through the following steps:

1. **Importance Analysis**: initially, the model, through its internal attention mechanisms, assesses which parts of the input contribute most to the classification decision. These attention mechanisms are capable of assigning a relative weight to different areas or features of the input, thereby identifying those that the model perceives as most relevant to its decision.

2. **Feature Selection for Masking**: based on the importance analysis, the features or areas of the input that receive the highest weights (and thus are considered most discriminative by the model) are selected for masking. This process is based on the hypothesis that by concealing the parts of the input deemed most informative, the model will be forced to search for and rely on other, less obvious or previously overlooked features.

3. **Application of Masking**: the selected features are then effectively masked, meaning they are made invisible or altered in such a way as to reduce their importance for the model's decisions. This can be achieved through various methods, such as the application of black masks, the addition of noise, or the use of specific distortion techniques that make these features unrecognizable by the network.

4. **Fine-tuning of the Model**: with the new masked dataset, the model undergoes a fine-tuning phase. This phase allows the network to adapt and learn to recognize and use new features or parts of the input that were not previously considered crucial for classification.

It should be emphasized that the attention-based masking strategy was adopted in place of the technique initially proposed in the original paper 5, which involved the use of xGrad-CAM (Gradient-weighted Class Activation Mapping). This is a technique that highlights the areas of the image that contribute 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 important areas for predicting the class label, effectively allowing for an intuitive understanding of why the model makes its decisions. By overlaying this heatmap onto the original image, it becomes clear which features or parts of the image were deemed most relevant or discriminative by the neural

network during its decision-making process. Although xGrad-CAM provides valuable results, we opted for attention-based masking for its ability to offer a more direct and flexible approach in selectively hiding key features. This choice aims to further strengthen the model's generalization, pushing it to discover and leverage new attributes of the dataset, thereby reducing dependence on potentially misleading or overly specific features of the original training set.

2.3 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.

3 Implementation details and achieved results

3.1 Single-label classification with Spurious Features.

The model used is a reduced version of VGG, named *VGGWithAttention*, characterized by two groups of convolutional layers, with the first group transforming the input from 3 to 32 channels and the second from 32 to 64 channels. After each group of convolutional layers, pooling is applied to reduce spatial dimensions, and three fully connected layers are used to progressively reduce the dimensionality to the number of classes in the CIFAR-10 dataset (10 classes).

```
# Building the Base Model (A reduced version of VGG)
# Defining a New Model with Access to an Intermediate Layer:
# First, we will define a modified version of the model that allows us
# to access the activations of an intermediate layer, which we will use as our "attention map".

class VGGWithAttention(nn.Module):
    def __init__(self):
        super(VGGWithAttention, self).__init__()
        # Convolutional layer 1
        self.conv1 = nn.Conv2d(3, 32, 3, padding=1)  # Input: 3 channels, Output: 32 channels, Kernel: 3x3
        self.conv2 = nn.Conv2d(32, 32, 3, padding=1)
        # Pooling
        self.pool = nn.MaxPool2d(2, 2)  # Kernel: 2x2
        # Convolutional layer 2
        self.conv3 = nn.Conv2d(32, 64, 3, padding=1)
        self.conv4 = nn.Conv2d(64, 64, 3, padding=1)
        # # Convolutional layer 3 (reduced compared to standard VGG)
        self.conv5 = nn.Conv2d(64, 128, 3, padding=1)
        self.conv6 = nn.Conv2d(128, 128, 3, padding=1)
       # Reduced fully connected layers
        self.fc1 = nn.Linear(128 * 4 * 4, 512)  # Reduced da 4096
        self.fc2 = nn.Linear(512, 256)  # Reduced da 4096 a 256
        self.fc3 = nn.Linear(256, 10)  # 10 output classes for CIFAR-10

    def forward(self, x):
        # Applying convolutional layers and pooling
        x = self.pool(F.relu(self.conv2(F.relu(self.conv1(x)))))
        x = self.pool(F.relu(self.conv4(F.relu(self.conv3(x)))))
        # Consider activations here as attention map
        attention_map = F.relu(self.conv6(F.relu(self.conv5(x))))
        x = self.pool(attention_map)
        x = x.view(-1, 128 * 4 * 4)
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        output = self.fc3(x)
        return output, attention_map
```

The training of the model occurs in two phases:

1. **Initial Training**: the model is trained on the CIFAR-10 dataset without applying masking. This phase serves to establish a baseline for the model's performance.

2. **Fine-Tuning**: after the initial training, the model is further optimized by applying masking to the input images, forcing it to explore new features compared to those extracted during the training phase.

The images are preprocessed through normalization. Initially, SGD is used with a learning rate of 0.001, momentum of 0.9, number of epochs equal to 12, and batch size of 4. For fine-tuning, the learning rate is reduced to 0.0001 (a lower rate favors a more stable and gradual convergence towards a local optimum).
The use of the "mps" device in the code is aimed at leveraging *Metal Performance Shaders*, (MPS, a collection of high-performance shaders and image processing functions that are optimized to fully leverage the GPU of Apple devices.) having tested the code on a Macbook with an M3 Pro chip.

3.2 Experimental Results

Table 1: Training results.

Phase	Loss	Accuracy (%)
1, 2000	2.303	10.15
1, 4000	2.303	10.11
1, 6000	2.303	9.57
1, 8000	2.303	10.45
1, 10000	2.286	13.21
1, 12000	2.048	23.36
...	...	...
12, 2000	0.141	95.34
12, 4000	0.155	94.76
12, 6000	0.200	93.20
12, 8000	0.201	92.95
12, 10000	0.217	92.59
12, 12000	0.217	92.47

Table 2: Fine-tuning results.

Phase	Loss	Accuracy (%)
1, 2000	0.076	97.75
1, 4000	0.058	98.30
1, 6000	0.044	98.75
1, 8000	0.040	98.76
1, 10000	0.037	98.78
1, 12000	0.037	98.85

Table 3: Model evaluation on the test set without masking.

Class	Accuracy (%)
plane	78.60
car	84.70
bird	70.80
cat	58.60
deer	77.10
dog	64.80
frog	76.90
horse	83.90
ship	83.40
truck	85.30
Total	**76.41**

Table 4: Model evaluation on the test set with masking

Class	Accuracy (%)
plane	83.10
car	88.90
bird	72.10
cat	62.10
deer	72.70
dog	71.80
frog	85.60
horse	83.90
ship	86.40
truck	89.50
Total	**79.61**

Table 5: Results of selective classification

Metric	Value (%)
Accuracy (excluded rejections)	85.65
Rejection rate	5.28
Agreement rate	88.38

Analyzing this experimental results several significant insights can be drawn about the effectiveness of the VGGWithAttention 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, a significant improvement is observed 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 application shows an accuracy increase for each class. This demonstrates the effectiveness of masking in improving the model's ability to distinguish

between different classes, making the model more robust and reliable in image classification.
Effectiveness of Selective Classification: the results of selective classification show high accuracy (`85.65%`) when rejections are excluded, along with a relatively low rejection rate (`5.28%`) and a high agreement rate (`88.38%`). 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.

3.3 Multi-label classification with Spurious Features

For this task, a model based on the convolutional neural network *ResNet50* (a convolutional neural network (CNN) designed for image recognition, part of the Residual Networks (ResNets) family introduced by Kaiming He et al. in 2015. Architecture developed to overcome the vanishing gradient problem, common in deep networks, through the use of residual connections), was trained without applying any masking mechanism. 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.

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 "force" the model to focus on new features of the images, 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.

Selective classification. As already observed with the VGGWithAttention model, we extended the selective classification approach to ResNet50, applying a similar methodology. This process includes loading the already trained models, comparing their respective predictions on a selected test dataset, and using a confidence threshold to filter out predictions deemed reliable. 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 for lack of reliability, and the frequency of agreement between the two models. This approach highlights the value of attention mechanisms in refining the accuracy and reliability of deep learning models, underlining the importance of advanced classification strategies in significant application contexts.

4 Conclusion

Despite the innovative approach and advanced techniques applied in the project, it is important to underscore a significant challenge encountered during the evaluation phase of the proposed solution for the ResNet50 network. Given the intrinsic complexity of the task and the computational power limitations available, it was decided to test the model's functionalities on a reduced subset of the original CelebA dataset. This choice was dictated by the need to efficiently manage available resources while trying to maintain as accurate and significant a level of analysis as possible.
However, it emerged that, despite efforts to optimize and adapt the learning and evaluation process to the available hardware configuration, it was not possible to obtain metric values that allowed a conclusive and satisfactory evaluation of the goodness of the proposed solution.
In conclusion, although a potentially effective strategy for improving the reliability and precision of deep learning models through the use of attention mechanisms and selective classification techniques has been implemented, the complete evaluation of its effectiveness remains, at this moment, an unfulfilled goal.

5 References

Taghanaki, S. A., Khani, A., Khani, F., Gholami, A., Tran, L., Mahdavi-Amiri, A., and Hamarneh, G. (2022). *Masktune: Mitigating spurious correlations by forcing to explore.* arXiv:2210.00055v2 [cs.LG] 8 Oct 2022.