IS MY DEEP LEARNING MODEL LEARNING MORE THAN I WANT IT TO?

Anonymous authors
Paper under double-blind review

ABSTRACT

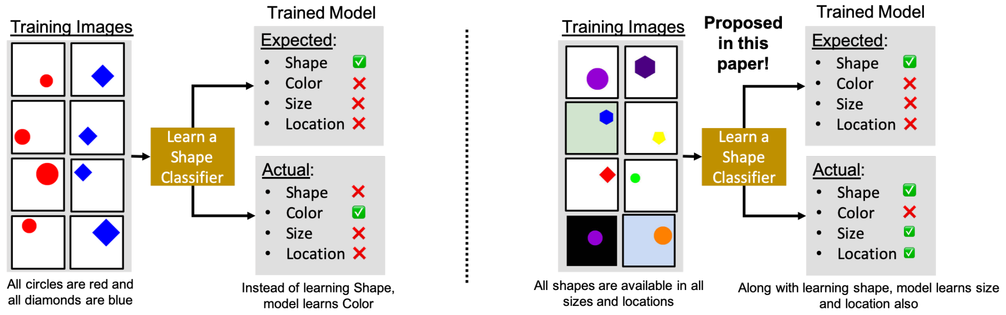
Existing deep learning approaches for learning visual features tend to extract more information than what is required for the task at hand. From a privacy preservation perspective, the input visual information is not protected from the model; enabling the model to become more intelligent than it is trained to be. Existing approaches for suppressing additional task learning assume the presence of ground truth labels for the tasks to be suppressed during training time. In this research, we propose a three-fold novel contribution: (i) a novel metric to measure the trust score of a trained deep learning model, (ii) a model-agnostic solution framework for trust score improvement by suppressing all the unwanted tasks, and (iii) a simulated benchmark dataset, PreserveTask, having five different fundamental image classification tasks to study the generalization nature of models. In the first set of experiments, we measure and improve the trust scores of five popular deep learning models: VGG16, VGG19, Inception-v1, MobileNet, and DenseNet and demonstrate that Inception-v1 is having the lowest trust score. Additionally, we show results of our framework on color-MNIST dataset and practical applications of face attribute preservation in Diversity in Faces (DiF) and IMDB-Wiki dataset.

1 INTRODUCTION

The primary objective of artificial intelligence is to imitate human intelligence *tabular rasa*. Especially, with the advent of deep learning (DL), the models are striving to perform composite tasks by learning complex relationships and patterns available in noisy, unstructured data (Ruder, 2017). With this sudden growth in the consumption of data by models, there has been a lot of study on the privacy and security of the learnt model (Shokri & Shmatikov, 2015). Data governance and model governance frameworks, control and protect sharing of data and model meta information between two entities and also their social implications (Helbing, 2019).

The premise of model privacy has majorly revolved around preserving the model content from human (man-in-the-middle) adversarial attacks (Abadi et al., 2016). However, the model itself could learn all the private information from the data and become much more intelligent than the original intent it was trained for. With the strive for model generalization, including techniques for transfer learning and multi-task learning, the model is encouraged to learn more and more generic features from the data that could be used for more than one task (Søgaard & Goldberg, 2016). Consider the example described in Figure 1, where a classifier is trained to detect the shape of an object from images. However, using the features extracted by the above classifier, the size and location of the object in the image can also be predicted. Thus, a shape classifier is more intelligent than its objective of only predicting the shape of the object. While in certain applications, this is a required property of classification models (such as in, transfer learning and domain adaptation), in most of the privacy preserving applications, the data and its other visual attributes have to be kept private from the model itself. As an additional real-world example, we train a DL model to predict the gender from a face image. However, the DL model learns most generic features from the face image, enabling it to predict the age and the identity of the person. The input face image could be saved securely from a human attacker, however, there is not much focus on securing from the model itself.

Additionally as shown in Figure 1 (a), the task of debiasing is to remove the the bias (color) in learning a specific task (shape). This happens due to the high correlation between the color and shapes in the input images. However, as shown in Figure 1 (b), our task in model trust is to forcefully



(a) **Model Debiasing**: Can we remove the bias in learning a specific task? (b) **Model Trust**: Can we enable the model to learn only a specific task ?

Figure 1: Visually distinguishing the concepts model debiasing and model trust (what we aim to do). The fundamental research motivation in this work is to study if a learning model could be restricted to perform only one or a specific group of tasks.

ensure that the model learns to perform only one or few selected tasks (shape) from the input images and unlearn all other tasks (color, size, location). If multi-class classification tasks could be done from the same image, the research question is, *"How can we ensure that the model is learnt only for one or a few tasks (called as, preserved tasks), and is strictly not learnt for the other tasks (called as, suppressed tasks)?"*. To pursue research on this problem, there are few evident challenges: (i) there is a lack of a balanced and properly curated image dataset where multiple classification tasks could be performed on the same image, (ii) the complete knowledge of both the preserved tasks and the suppressed tasks should be known *apriori*, that is, we cannot suppress those tasks that we don't have information about, and (iii) presence of very few model agnostic studies to preserve and suppress different task groups. In this research, we propose a novel framework to measure the trust score of a trained DL model and a solution approach to improve the trust score during training. The major research contributions are summarized as follows:

1. A simulated, class-balanced, multi-task dataset, *PreserveTask* with five tasks that could be performed on each image: shape, size, color, location, and background color classification.
2. A novel metric to measure the trustworthiness score of a trained DL model. The trust scores of five popular DL models are measured and compared: VGG16, VGG19, Inception-v1, MobileNet, and DenseNet. A generic model-agnostic solution framework to improve the trust scores of DL models during training by preserving a few tasks and suppressing other tasks on the same image.
3. Experimental analysis are performed for the proposed framework in comparison with other existing approaches under different settings. Experimentally, we considered the model with the least trust score, Inception-v1, and showed that the proposed framework aids in improving the overall trust score [1].
4. To demonstrate the practical applications and generalizability of the metric and the solution framework, we show additionally results in colored MNIST dataset and face attribute preservation using two datasets: (i) Diversity in Faces (DiF) (Merler et al.) (ii) IMDB-Wiki (Rothe et al., 2018).

2 Literature Review

There are broadly two different groups of work related to the research problem at hand: (i) k-anonymity preservation and (ii) attribute suppression.

k-anonymity Preservation: The objective here is to preserve the anonymity of certain attributes from being predicted by the model. To quote some earlier works, Boyle et al. (2000), studied to mask out potentially sensitive information from video feeds. In the last decade, face recognition

[1] The benchmark dataset along with the splits, baselines features, results, and the code are made available here: `https://github.com/dl-model-recommend/model-trust`

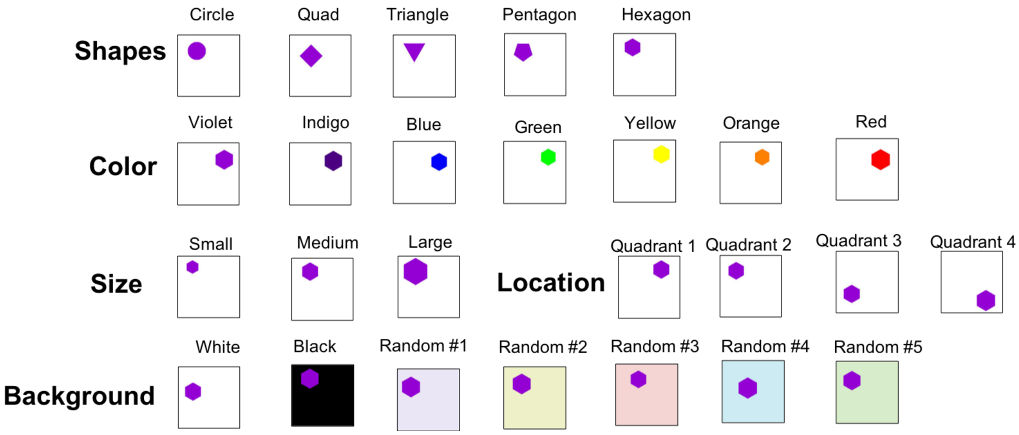


Figure 2: Landscape of the *PreserveTask* dataset describing the set of different possible tasks. Five tasks could be performed on each image and each task has varying number of classes.

has become an important commercial applications and also an application that demanded discussion regarding privacy preservation. Studies focused on extracting only the required meta information from face images while not extracting the identity. This was a required step to make face as a usable biometric. Studies such as Gross et al. (2006), Newton et al. (2005), and Mirjalili et al. (2018) focused on preserving the identity of the face image from the model by performing face de-identification. Studies such as Mirjalili & Ross (2017) and Othman & Ross (2014) focused on anonymizing the face gender information while models could extract the identity.

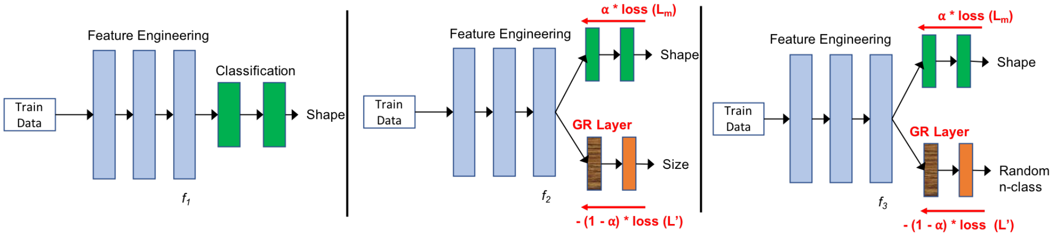
Attribute Suppression: The aim of this group of techniques is to explicitly suppress a few attributes by perturbing the input data to the model. Studies such as Rozsa et al. (2016) and Rozsa et al. (2017) test if the learnt models are robust and protected against adversarial attacks. Chhabra et al. (2018) suggested using a constrained generative adversarial network (GAN) to perturb the input face image and suppress the required attribute. The GANs will generate the attribute free face image of the original face image. The closest related work to our approach, is the study by Jayaraman et al. (2014) where the visual attributes are decorrelated using a negative gradient in the model. The results demonstrate that the classification task could be performed by preserving specific attributes in the image while suppressing the influence of the remaining.

Additionally, there is a good amount of research in bias mitigation while learning models (Zhao et al., 2017) (Kim et al., 2018) (Attenberg et al., 2015) (Li & Vasconcelos, 2019). The primary aim is to debias the model learning from any kind of correlated attributes (Alvi et al., 2018) (Raff & Sylvester, 2018) (Kim et al., 2019) (Wang et al., 2019), which is different from our aim of improving the model's trust. The major gaps in the existing research works are: (i) most of the techniques focus on data perturbation, that is, changing the input data from x to x' such that the suppressed task information is not available in the data. There is not much focus on model perturbation without altering the input data, (ii) most of the existing datasets have only binary attributes and hence suppressing and preserving a few tasks does not actually translate to the classification complexity of multi-class tasks, and (iii) there is a lack of a well curated benchmark dataset to evaluate the privacy preserving capacity of DL models.

3 PRESERVETASK DATASET

Shared tasks performed on the same image carry some common attributes which are often extracted by complex deep learning models. The objective of this is to untangle the shared tasks and enable deep learning models to perform only one (or few) of those tasks. In order to evaluate the performance of such a framework, the dataset should have the following properties:

- Should perform multiple tasks on the same image and each task should have varying number of classes, in order to study the relationship of complexity of classification tasks.



(a) Learning shared tasks

(b) Suppressing known tasks

(c) Suppressing unknown tasks

Figure 3: (a) A deep learning model learning features suited for multiple tasks, more than the intended shape classification task, (b) Existing approaches suppress other known tasks, such as size classification by backpropagation of negative loss or gradient, (c) Proposed approach of suppressing all possible n-class classification task by using random class labels.

- As this research area is nascent, the dataset should be noise-free and class balanced, to avoid other complexities that could influence classification performance.
- Tasks should be designed in such a way that certain tasks, share common attributes and features, while certain tasks should be independent of each other.

There are some similar publicly available datasets in the literature. LFW (Learned-Miller et al., 2016), CelebA (Liu et al., 2015), IMDB-Wiki (Rothe et al., 2018), AwA 2 (Lampert et al., 2009), and CUB (Wah et al., 2011) datasets have multiple binary classification tasks, while only one non-binary classification task. It is challenging to study the influence of complexity of classification tasks using these datasets and hence is not extendable to practical applications. CLEVR (Johnson et al., 2017) dataset provides with four different tasks with variable number of classes. However, each image contains multiple objects with different shape, color, and textures, allowing multiple labels for each task. Task suppression in multi-label, multi-task classification setting provides a very challenging experimental setting.

Inspired from the CLEVR dataset, we create a new *PreserveTask* dataset, which is a multi-task dataset exclusively designed for the purpose of bench-marking models against preserving task privacy. The primary objective is to create easy-to-perform multi-task dataset, where the performance of the individual tasks is high. As shown in Figure 2, *PreserveTask* dataset has five different classification tasks, as follows: (i) *Shape Classification (5):* circle, triangle, diamond, pentagon, hexagon, (ii) *Color Classification (7):* violent, indigo, blue, green, yellow, orange, red, (iii) *Size Classification (3):* small, medium, large, (iv) *Location Classification (4):* quadrant 1, quadrant 2, quadrant 3, quadrant 4, (v) *Background Color Classification (3):* white, black, or colored.

These five tasks are chosen such that few tasks are highly correlated (size, shape), while few tasks are ideally independent of each other (size, color). All the images are generated as 256×256 colored images. There are 5 (shapes) * 7 (color) * 3 (size) * 4 (location) * 3 (background color) = 1260 variations, with 50 images for training and 10 images for testing for each variation, generating a total of $63,000$ training and $12,600$ images. This ensures that there is a perfect class balance across all tasks. It is to be noted that the task of suppression of unknown shared task is a fairly open research problem. Hence, in order to set the benchmark of different frameworks, an easy, straightforward *PreserveTask* dataset is created as a conscious decision without having much noise, such as in DeepFashion (Liu et al., 2016) dataset. As the problem area matures, further extensions of this dataset could be generated and more real world natural objects could be added.

4 Proposed Approach

To understand the current scenario of shared task learning, consider any deep learning model as shown in Figure 3 (a). Assume a deep learning model, say VGG19, is trained for predicting the shape of objects in images. Ideally, the features f_1 obtained from the model should be good for object shape prediction. However, it is observed that different size, color, location prediction classifiers could be trained on top of f_1 demonstrating that f_1 contains more information about the object than just its shape. While this is a required property in multi-task learning and in applications of domain adaptation, from a task privacy preservation perspective this should be controlled. In literature, few technique variants exist to suppress the model from learning a few attributes or tasks (Narayanaswamy et al., 2017). As shown in Figure 3 (b), if the model has to be suppressed from learning the size of the objects, a negative loss or negative gradient is applied to enable features f_2 to not carry any information about the size of the object while retaining all the information about the shape of the object. This comes with an assumption that the information about the tasks to be suppressed are available during training time along with its ground truth class labels for the entire training data.

In our proposed framework, we overcome this assumption and do not expect the suppression task information to be available during model training time. Additionally, we provide a model agnostic approach of suppressing task learning so that the framework could be directly applied to any deep learning model. Let $x \in X$ be the input data and $y_x^{(1)} \in Y^{(1)}$ to $y_x^{(n)} \in Y^{(n)}$ be the n different tasks that could be performed on the image. We learn a model, $g(f(x)) : X \rightarrow Y^{(1)}$, where $f(.) : X \rightarrow Z^{(1)}$, be the feature representation for the given task. Ideally, while only $g(.) : Z^{(1)} \rightarrow Y^{(1)}$ should be possible, we observe that $g(.) : Z^{(1)} \rightarrow Y^{(i)}$ for $i \in (2, n)$ provides high classification accuracy in most cases. To overcome this challenge, we generate random n-class labels in the gradient reversal (GR) branch (Ganin & Lempitsky, 2014) in order to suppress any other n-class classification, as shown in Figure 3 (c),. Multiple gradient reversal branches could be built for varying values of n to suppress all possible other classification tasks. The DL model is trained by a custom loss function as follows,

$$minE_{x\sim P_x(.)}[\lambda L_m(y_x^{(1)}, f(g(x))) - (1-\lambda)L'(rand(R^{Y^{(1)}}), f(g(x)))] \tag{1}$$

where L_m is the loss of the model branch trained for the task to be preserved. L' is the sum of individual losses (L_i) which are to be maximized (task suppressed). λ and $(1-\lambda)$ are the weights given for the minimization and maximization losses which can be chosen based on the amount of sharing between the tasks. Additionally, each of the individual L_i could be distinct loss functions in the model, depending on the task performed. Thus, it can be observed that the proposed framework is both DL model agnostic and loss function agnostic.

4.1 Trust Score

PreserveTask will be used as the benchmark dataset against which the trust score of any trained DL model could be extracted. The trained DL model is evaluated against different tasks in the *PreserveTask* and the entire confusion matrix of performance accuracy is obtained (5×5 corresponding to the five tasks). The behavior of an ideal DL model, would provide 100% accuracy on the leading diagonal i.e., the tasks it was trained for, while providing, random classification accuracies for other tasks. The confusion matrix for such an ideal DL model is shown in Figure 4. For example in the first row, the DL model was trained to learn and predict the color of the object. Hence, color prediction performance should be 1 (denoting, 100% accuracy), while other tasks should provide random $1/n$ accuracy, where n is the number of classes.

Let the ideal performance matrix be denoted as M and the obtained performance matrix for a given trained DL model be T. By intuition, the matrix T that does not deviate much from the ideal matrix M should have a higher trust score. The trust score is mathematically computed as follows,

$$\text{Trust Score} = \frac{\Sigma(|M - T| \cdot W)}{\Sigma W} \tag{2}$$

where, $W = 4 \times \mathcal{I}_5 \cdot \mathbb{1}_5$ provides the weight corresponding to each task pair, I is an identity matrix and $\mathbb{1}_5$ is a ones matrix, each of dimensionality 5×5. Since for each preserving task, there are four suppressing tasks, the deviation of the preserving task from the ideal matrix is scaled by a factor of four to normalize the computation.

Note that if the diagonal elements perform poorly, the concern is on the performance of the model. On the contrary, if the non-diagonal elements has a higher performance, the concern is on the trust of a model from a privacy preservation perspective. The proposed metric implements this notion to compute the trustworthiness of a trained DL model. The trust score is bounded between [0,1]. By

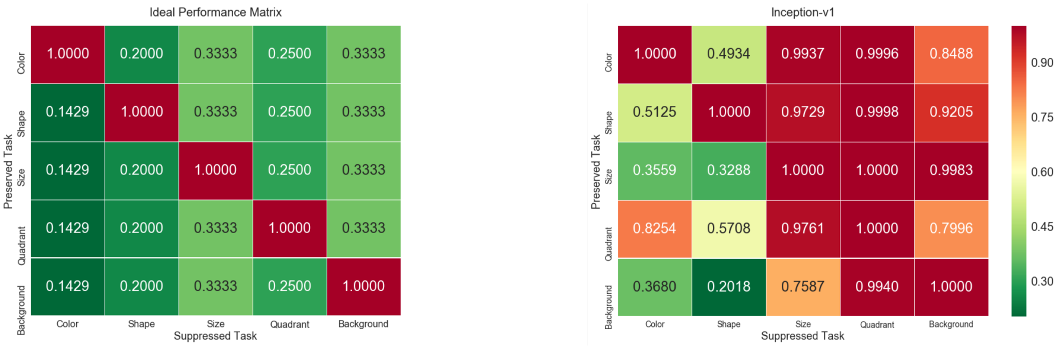


Figure 4: (Left) The accuracy matrix demonstrating the behavior of an ideal trusted DL model. The leading diagonal shows perfect classification while the rest of the values are random classification. (Right) The accuracy matrix detailing the shared task performance of Inception-v1 on the *PreserveTask* dataset.

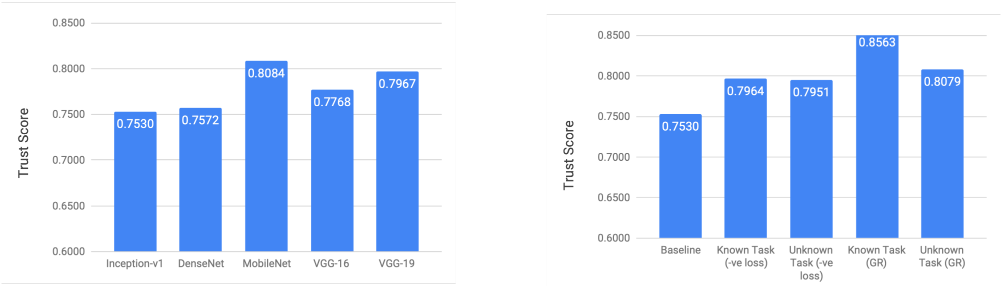


Figure 5: (Left) Trust scores obtained for various DL models. It can be observed that, of the five models, the Inception-v1 and MobileNet has the least and highest trust score, respectively. (Right) Trust scores obtained after various suppression techniques for Inception-v1. It can be observed that using random labels for unknown tasks, we could improve the trustworthiness.

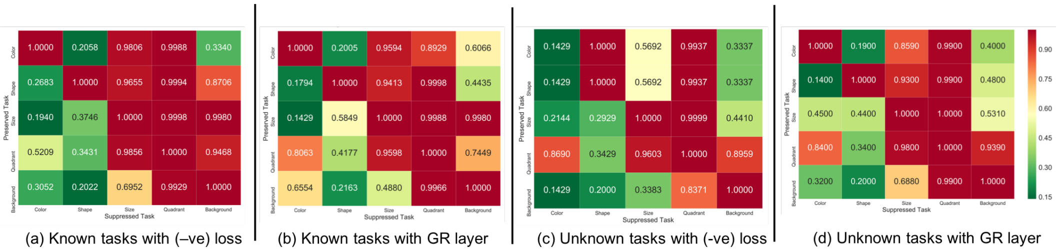
empirical analysis, we observe that a trust score above 0.9 is highly desirable, a trust score between 0.8 and 0.9 is practically acceptable, and any score below 0.8 is considered poor. The trust score of the ideal matrix is 1, while the trust score of a $\mathbb{1}_5$ (all task classification performance is 100%) is 0.6259. To understand the sensitivity of the proposed metric, let us assume that in the ideal matrix, any one non-diagonal element is changed to 1 which results in a trust score of 0.98125. Thus, any reduction of (1 - 0.98125) = 0.0175 in the trust score corresponds to one additional unwanted task being learnt by the classifier.

5 Experimental Results

In this section, we show the experimental results and perform analysis of the proposed framework. Initially, we measure the trustworthiness of the existing models. We then experimentally demonstrate suppression of different tasks in various experimental settings. All the experiments are performed using the *PreserveTask* dataset. For additional results and detailed comparison with other techniques, please refer to the appendix.

5.1 How Trustworthy are Existing Models?

Consider a popular deep learning model, Inception-v1 (Szegedy et al., 2016) consisting of 22 computational layers. The model was trained from scratch using the *PreserveTask* for the task of shape classification, providing 99.98%. In order to study, if this deep learning model learnt additional visual attributes, as well, the last flatten layer's output (4096×1) were extracted. Four different



(a) Known tasks with (–ve) loss (b) Known tasks with GR layer (c) Unknown tasks with (-ve) loss (d) Unknown tasks with GR layer

Figure 6: The performance matrix obtained after suppressing the known tasks in (a), (b) and the unknown tasks in (c), (d). Comparative results between a baseline negative loss function and the proposed GR layer based suppression is also shown. All results are computed for the Inception-v1 model.

two-hidden layer neural network classifiers (512, 100) were trained [2] using the extracted features to predict size, color, location, and background color of the objects. The prediction accuracies were 97.29%, 51.25%, 99.98%, 92.05%, respectively for the four tasks. It can be observed that the performance of size, location, and background prediction are really high proving that the features obtained from Inception v1 model has features corresponding to these tasks as well. Also, it can be observed that the color prediction performance is very low, as shape and color prediction are inherently independent tasks. The similar experiment is repeated for training the Inception v1 model on one task and using the learnt feature to predict the performance of other tasks, and the results are shown in Figure 4. Ideally, only the diagonal elements of this confusion matrix should have higher accuracies (red in color) while the rest of the prediction should have lower accuracies (green in color). Accordingly, the trust score of the trained Inception-v1 model (proposed in section 4.1) was found to be 0.7530, which is very poor.

In order to further demonstrate that this additional intelligence is not a property of just Inception-v1 model, similar experiments are performed using four other popular deep learning models: VGG16, VGG19, MobileNet, and DenseNet. The trust scores of all the DL models are shown in Figure 5. It can be observed that out of these five models, Inception-v1 and DenseNet has the lowest trust score while MobileNet has the highest trust score. While one could argue that the Inception-v1 model learns highly generic features supporting multi-task and transfer learning, from a privacy preservation perspective, the model is found to have a poor trust score. This leads to the open question, "*Do models always needs to be additionally intelligent, and if not, how to suppress them?*"

5.2 How to Suppress Known Tasks?

In this section, we perform experiments to suppress the tasks that are known apriori during training, that is, the ground truth labels of the suppression task is available. For simplicity, in demonstrating the experimental results, we assume that one task is to be preserved and one task is to be suppressed, using the Inception-v1 model. This experimental setting is similar to the approach explained in Figure 3 (b). The gradient reversal (GR) layer unlearns the suppressed task, while learning the preserved task. In order to compare the performance of GR, we also use a customized negative loss function which minimizes the loss obtained for the preserved task while maximizing the loss obtained for the suppressed task, weighted by a constant factor. The features eventually extracted from the flatten layer has to show similar performance on the preserved task while reduced performance on the suppressed task.

Figure 6 (a) and (b) demonstrates the results obtained for Inception-v1 using negative loss function and the proposed GR layer. While the leading diagonal elements showed the same performance, in comparison with Figure 4, it can be observed that prediction results of the suppressed tasks reduced in most of the cases. For example, while preserving the object shape prediction, suppressing the background color prediction performance dropped from 92.05% to 44.35%. This indicates that the extracted features no longer contain information about the background color of the image. The

[2] with default scikit-learn parameters

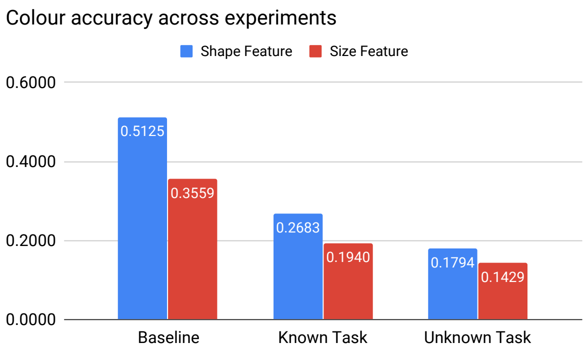


Figure 7: Comparison of color prediction performance with and without using the different task suppression mechanisms. It can be observed that using random labels reduces the performance of color prediction irrespective of whether the preserved task was shape or size prediction.

corresponding trust scores are shown in Figure 5. It can be observed that suppressing known tasks using GR layer improves the trust of the baseline model from 0.7530 to 0.8563.

5.3 How to Suppress Unknown Tasks?

The results obtained in the previous section made the assumption that the ground truth labels of the suppression task have to be available while training the Inception-v1 model. In an attempt to break that assumption, the experimental setting discussed in Figure 3 (c) is performed. Instead of the actual ground truth labels of a particular task, randomly generated n-class labels are used during every mini-batch. Thus, for the same mini-batch training in the next epoch, a different set of random class labels are generated to be maximized. This ensures that the model does not memorize a single suppression task, but, learns to suppress all possible n-class classification tasks.

Figure 6 (c) and (d) demonstrates the results obtained by using random class labels. In comparison with Figure 4, it can be observed that using random class performs well in certain settings. For example, while trying to preserve the shape features and suppressing the prediction capacity of background color, the original model's prediction performance of 92.05% reduced to 87.06% by using the actual labels of background color, while further reduced to 33.37% while using random 3-class labels. It is further highlighted in Figure 7 where color prediction is chosen as the task to be suppressed, while shape and size are independently being preserved. It can be observed that the proposed framework of using random labels, reduces the performance of color prediction from 51.25% to 26.83% when using actual labels and 17.94% when using random labels, when shape prediction was the preserved task. A similar performance reduction from 35.59% to 14.29% is observed when size prediction was the preserved task.

We conclude that using random labels for task suppression produces a comparable trust score to using known labels while producing better results than the baseline trust score of a DL model.

6 Case Study on Challenging Practical Datasets

Colored MNIST Dataset: We introduced two additional tasks of foreground and background color prediction tasks into the MNIST dataset. As shown in Figure 8, colored MNIST images are created by randomly assigning one of the 10 possible foreground colors and one of the different 10 possible background colors. Similar assignment is performed in both training and test dataset, to maintain the standard experimental protocol. MobileNet model was trained from scratch to obtained a baseline trust score of 0.756. After using our framework for task suppression with random labels and gradient reversal based training on the suppression branch, we observed that the MobileNet models trust scores increased to 0.824. In Figure 8 (middle), the TSNE plot shows that when the model is learnt only for shapes, the features for 'red' and 'cyan' colored images are still separable. However, after

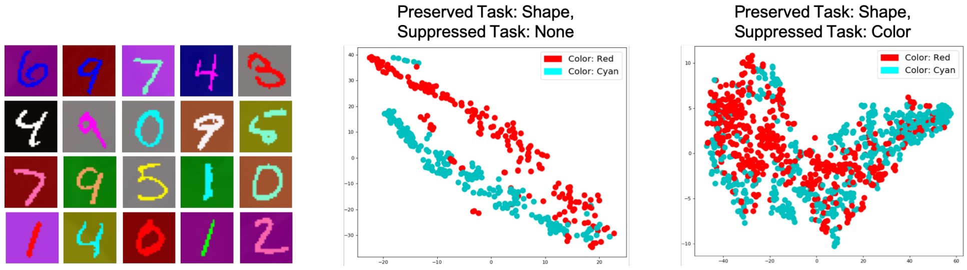


Figure 8: (Left) Sample images from the colored MNIST dataset. (Right) TSNE plot of the feature distribution of 392 images (class 0, foreground color: red and cyan) before and after suppressing the color prediction task.

suppressing the color prediction task using the proposed framework, the features 'red' and 'cyan' colored images are scattered and no longer separable, as shown in Figure 8 (right).

Diversity in Faces (DiF) Dataset: In DiF dataset (Merler et al.), we considered the tasks of gender (two class) and pose (three class) classification. The aim is learn (preserve) only one of these while suppressing the other. Since, the dataset was highly skewed for different classes, we considered a subset of 39296 images with equal class balance[3]. We trained Inception-v1 model on this dataset from scratch and obtained a trust score of 0.7497. Using our framework for task suppression with GR layer and known class labels, the trust score of the model increased to 0.8606. Additionally, with random unknown class labels, we observed that the model's trust scores increased to 0.9069.

IMDB-Wiki Dataset: In IMDB-Wiki dataset (Rothe et al., 2018), we considered the tasks of gender (two class) and age (ten class) classification. The cropped face images of the Wiki dataset are used to train the DenseNet model (the second least trusted model according to our trust scores). The trained model provided a baseline trust score of 0.7846. After using our framework for task suppression and known class labels, the trust score of DenseNet model increased to 0.7883. Also, with random unknown class labels, we observed that the model's trust scores increased to 0.7860.

Thus, our framework for measuring and improving a DL model's trust has lots of practical applications. A face recognition system or a face image based gender recognition system can now be deployed with an additional trust on the model's intelligence level.

7 Conclusion and Future Research

In this research, we showcased a model-agnostic framework for measuring and improving the trustworthiness of a model from a privacy preservation perspective. The proposed framework did not assume the need for the suppression task labels during train time, while, similar performance could be obtained by training using random classification boundaries. A novel simulated benchmark dataset called *PreserveTask* was created to methodically evaluate and analyze a DL model's capability in suppressing shared task learning. This dataset opens up further research opportunities in this important and practically necessary research domain. Experimentally, it was shown that popular DL models such as VGG16, VGG19, Inception-v1, DenseNet, and MobileNet show poor trust scores and tend to be more intelligent than they were trained for. Also, we show a practical case study of our proposed approach in face attribute classification using: (i) Diversity in Faces (DiF) and (ii) IMDB-Wiki datasets. We would like to extend this work by studying the effect of multi-label classification tasks during suppression.

References

Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

[3] Please refer to the appendix for the exact data distribution and the detailed performance matrix obtained

Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *European Conference on Computer Vision*, pp. 556–572. Springer, 2018.

Joshua Attenberg, Panos Ipeirotis, and Foster Provost. Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality (JDIQ)*, 6(1):1, 2015.

Michael Boyle, Christopher Edwards, and Saul Greenberg. The effects of filtered video on awareness and privacy. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pp. 1–10. ACM, 2000.

Saheb Chhabra, Richa Singh, Mayank Vatsa, and Gaurav Gupta. Anonymizing k-facial attributes via adversarial perturbations. *arXiv preprint arXiv:1805.09380*, 2018.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. *arXiv preprint arXiv:1409.7495*, 2014.

Ralph Gross, Latanya Sweeney, Fernando De la Torre, and Simon Baker. Model-based face de-identification. In *null*, pp. 161. IEEE, 2006.

Dirk Helbing. Societal, economic, ethical and legal challenges of the digital revolution: from big data to deep learning, artificial intelligence, and manipulative technologies. In *Towards Digital Enlightenment*, pp. 47–72. Springer, 2019.

Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1629–1636, 2014.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 1988–1997. IEEE, 2017.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. *arXiv preprint arXiv:1812.10352*, 2018.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9012–9020, 2019.

Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 951–958. IEEE, 2009.

Erik Learned-Miller, Gary B Huang, Aruni RoyChowdhury, Haoxiang Li, and Gang Hua. Labeled faces in the wild: A survey. In *Advances in face detection and facial image analysis*, pp. 189–248. Springer, 2016.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9572–9581, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1096–1104, 2016.

Michele Merler, Nalini Ratha, and Rogério S Feris.

Vahid Mirjalili and Arun Ross. Soft biometric privacy: Retaining biometric utility of face images while perturbing gender. In *Biometrics (IJCB), 2017 IEEE International Joint Conference on*, pp. 564–573. IEEE, 2017.

Vahid Mirjalili, Sebastian Raschka, Anoop Namboodiri, and Arun Ross. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In *2018 International Conference on Biometrics (ICB)*, pp. 82–89. IEEE, 2018.

Siddharth Narayanaswamy, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. Learning disentangled representations with semi-supervised deep generative models. In *Advances in Neural Information Processing Systems*, pp. 5925–5935, 2017.

Elaine M Newton, Latanya Sweeney, and Bradley Malin. Preserving privacy by de-identifying face images. *IEEE transactions on Knowledge and Data Engineering*, 17(2):232–243, 2005.

Asem Othman and Arun Ross. Privacy of facial soft biometrics: Suppressing gender but retaining identity. In *European Conference on Computer Vision*, pp. 682–696. Springer, 2014.

Edward Raff and Jared Sylvester. Gradient reversal against discrimination: A fair neural network learning approach. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 189–198. IEEE, 2018.

Rasmus Rothe, Radu Timofte, and Luc Van Gool. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision*, 126(2-4): 144–157, 2018.

Andras Rozsa, Manuel Günther, Ethan M Rudd, and Terrance E Boult. Are facial attributes adversarially robust? In *Pattern Recognition (ICPR), 2016 23rd International Conference on*, pp. 3121–3127. IEEE, 2016.

Andras Rozsa, Manuel Günther, Ethan M Rudd, and Terrance E Boult. Facial attributes: Accuracy and adversarial robustness. *Pattern Recognition Letters*, 2017.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Reza Shokri and Vitaly Shmatikov. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321. ACM, 2015.

Anders Søgaard and Yoav Goldberg. Deep multi-task learning with low level tasks supervised at lower layers. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pp. 231–235, 2016.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Haotao Wang, Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Privacy-preserving deep visual recognition: An adversarial learning framework and a new dataset. *arXiv preprint arXiv:1906.05675*, 2019.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*, 2017.

A Appendix

This supplementary material contains all the detailed hyper-parameters used by different models that we trained, to aid in reproducing the results that we showed in the research paper. Additionally, we provide more detailed analysis and visualizations of the results, that could not be included in the paper due to space constraints.

A.1 Baseline Deep Learning Models

Five different baseline deep learning models were used in the experiments: Inception-v1, VGG16, VGG19, DenseNet, and MobileNet. The different parameters and the training process used in these experiments are shown below:

- The data is z-normalized to have a zero mean and unit standard deviation, before being provided to the models for training.
- The standard architectures of Inception-v1, VGG16, VGG19, DenseNet, and MobileNet are borrowed from the default implementations in the Keras library.
- The deep learning models were trained with *categorical cross-entropy* and *Adam* optimizer with parameters as learning rate = 0.0001 and amsgrad set as $False$.

A.2 Classifier Models

For all the experiments, a two hidden layer neural network is used as a classifier. This is to maintain consistency of the same classifier across all the experiments.

- The architecture is Dense (512) $\rightarrow$ Dropout (0.5) $\rightarrow$ Dense (100) $\rightarrow$ Dropout (0.3) $\rightarrow$ Dense (num_of_classes)
- Each of the Dense layer has a $ReLU$ activation function.
- *categorical cross-entropy* is used as the loss function with *Adam* as the optimizer, having parameter values as learning rate = 0.0001 and amsgrad set as $False$.
- 20% of the data is used as validation data and the model is trained for 100 epochs with early stopping.
- Batch size of 32 was used to make the computation faster and the experiments were run using $1 \times K80$ GPU.

B Experimental Results and Analysis

In this section, we are including additional analysis, visualizations, and charts of the results presented in the main paper. In order to aid better comparison, we include the charts and results presented in the main paper also here, so that the supplementary could be read in an independent manner.

B.1 How Trustworthy are Existing Models?

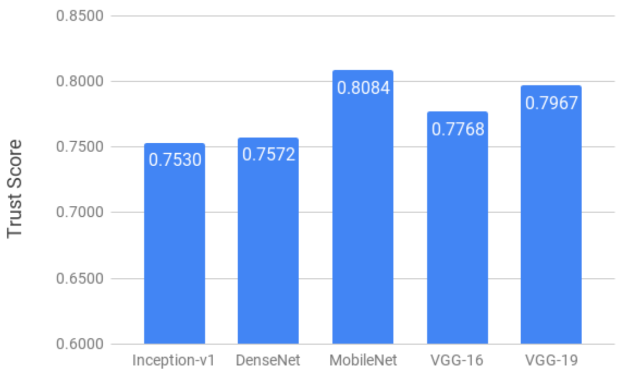


Figure 9: Trust scores obtained for various DL models. It can be observed that, of the five models, the Inception-v1 and DenseNet has the least trust score while MobileNet has the highest.

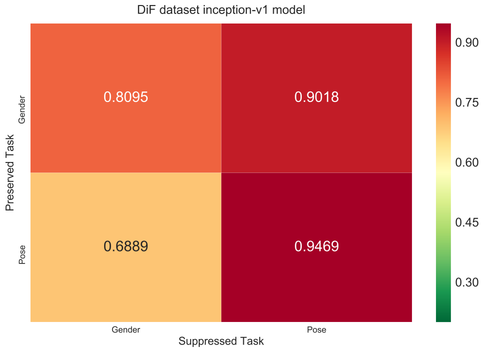


Figure 10: The performance matrix heat-map detailing the shared task performance of Inception-v1 model on the *PreserveTask* dataset.

B.2 How to Suppress Tasks?

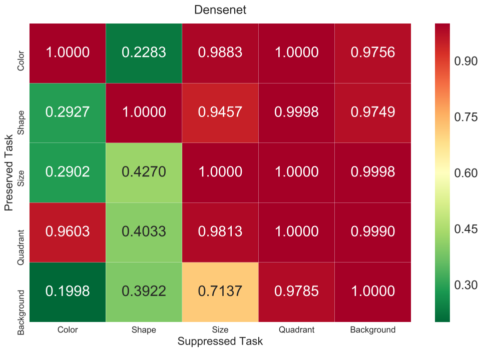


Figure 11: The performance matrix heat-map detailing the shared task performance of DenseNet model on the *PreserveTask* dataset.

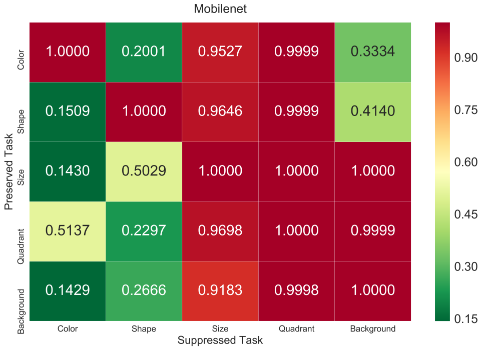


Figure 12: The performance matrix heat-map detailing the shared task performance of MobileNet model on the *PreserveTask* dataset.

C Case Study: Face Attribute Preservation

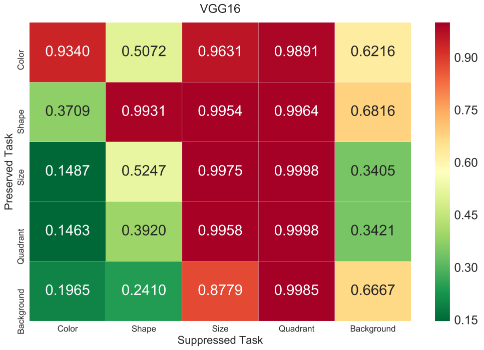


Figure 13: The performance matrix heat-map detailing the shared task performance of VGG-16 model on the *PreserveTask* dataset.

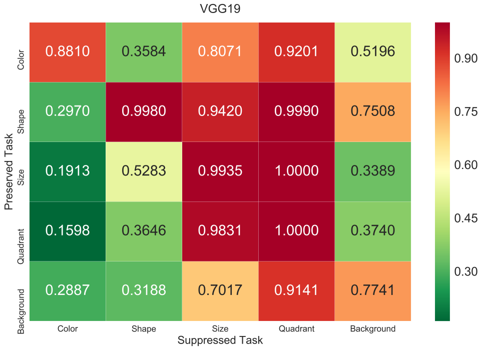


Figure 14: The performance matrix heat-map detailing the shared task performance of VGG-19 model on the *PreserveTask* dataset.

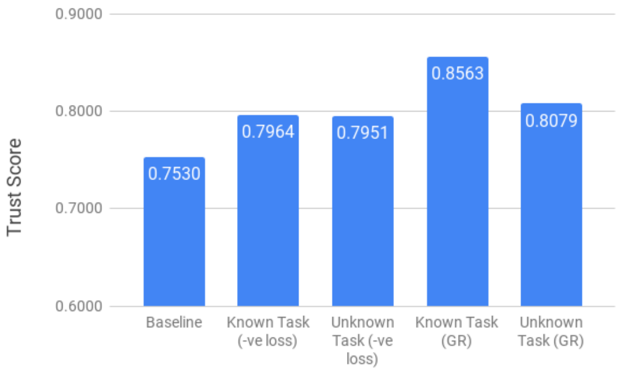


Figure 15: Trust scores obtained after various suppression techniques. It can be observed that even using random labels for unknown tasks, we could improve the trustworthiness of the Inception-v1 model on the *PreserveTask* dataset.

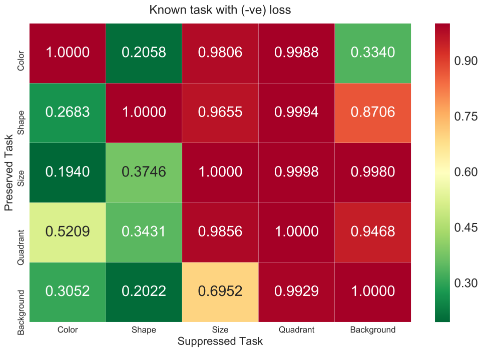


Figure 16: The performance matrix heat-map, after suppressing a known task using negative loss, detailing the shared task performance of Inception-v1 model on the *PreserveTask* dataset.

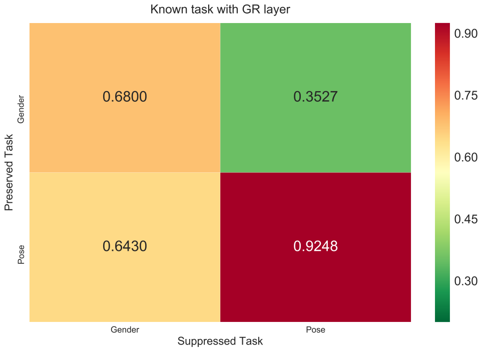


Figure 17: The performance matrix heat-map, after suppressing a known task using GR layer, detailing the shared task performance of Inception-v1 model on the *PreserveTask* dataset.

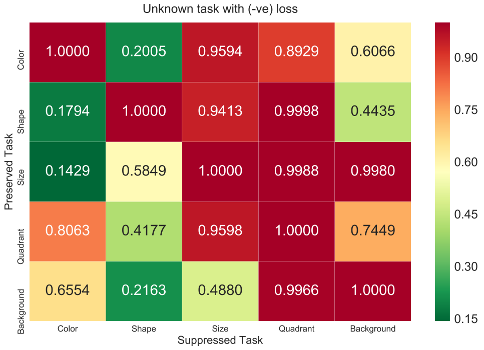


Figure 18: The performance matrix heat-map, after suppressing a unknown task using negative loss, detailing the shared task performance of Inception-v1 model on the *PreserveTask* dataset.

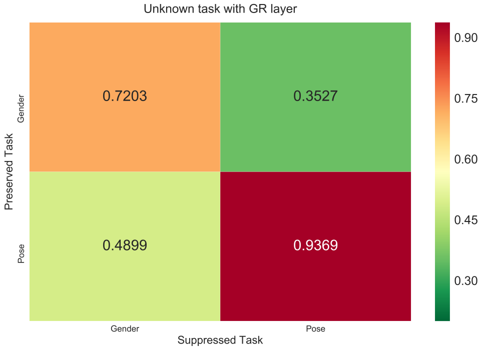


Figure 19: The performance matrix heat-map, after suppressing a unknown task using GR layer, detailing the shared task performance of Inception-v1 model on the *PreserveTask* dataset.

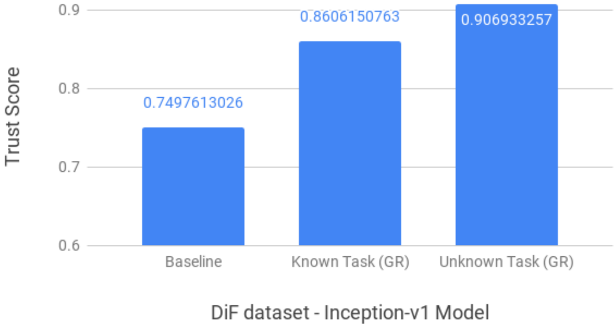


Figure 20: Trust scores obtained in the Diversity in Faces (DiF) dataset after various suppression techniques. It can be observed that even using random labels for unknown tasks, we could improve the trustworthiness of the Inception-v1 model.

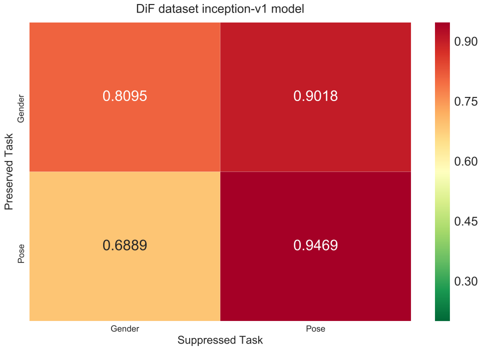


Figure 21: The performance matrix heat-map detailing the shared task performance of Inception-v1 model on the Diversity in Faces (DiF) dataset.

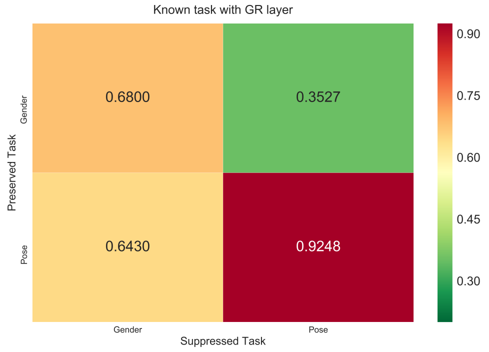


Figure 22: The performance matrix heat-map obtained after suppressing the known tasks, detailing the shared task performance of Inception-v1 model on the Diversity in Faces (DiF) dataset.

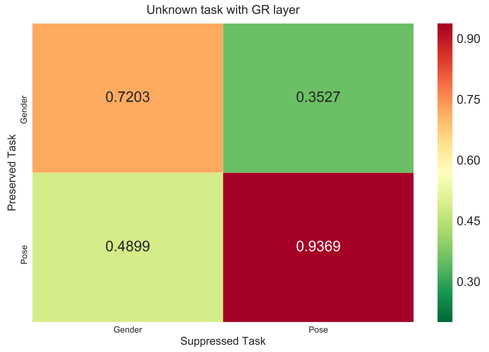


Figure 23: The performance matrix heat-map obtained after suppressing the unknown tasks, detailing the shared task performance of Inception-v1 model on the Diversity in Faces (DiF) dataset.

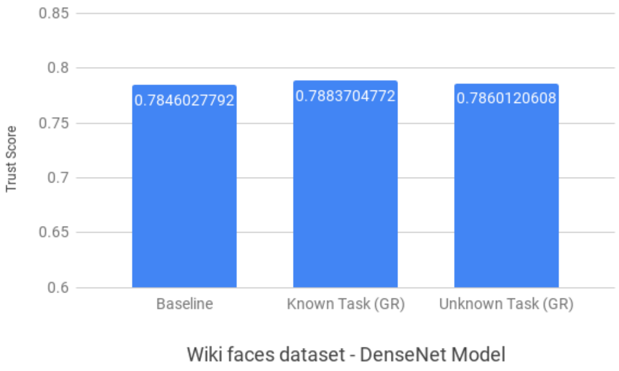


Figure 24: Trust scores obtained in the WIKI face dataset after various suppression techniques. It can be observed that even using random labels for unknown tasks, we could improve the trustworthiness of the Inception-v1 model.

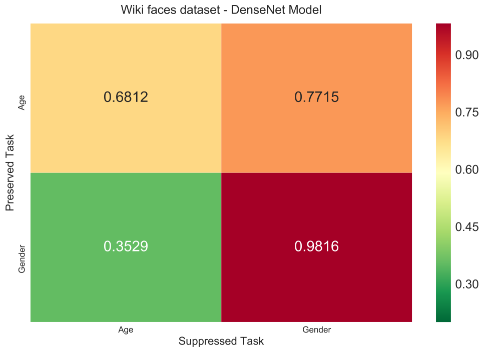


Figure 25: The performance matrix heat-map detailing the shared task performance of DenseNet model on the Wiki face dataset.

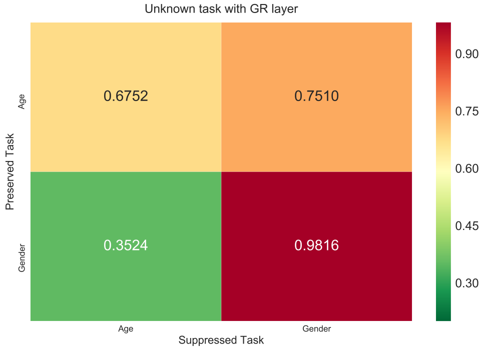


Figure 26: The performance matrix heat-map obtained after suppressing the known tasks, detailing the shared task performance of DenseNet model on the Wiki face dataset.

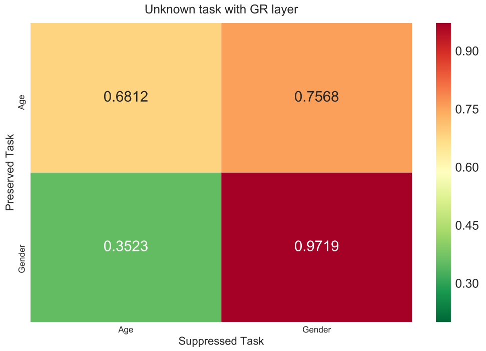


Figure 27: The performance matrix heat-map obtained after suppressing the unknown tasks, detailing the shared task performance of DenseNet model on the Wiki face dataset.