Learning with Domain Knowledge to Develop Justifiable Convolutional Networks

Rimmon Bhosale RIMMON281996@GMAIL.COM and **Mrinal Das** MRINAL@IITPKD.AC.IN
Indian Institute of Technology, Palakkad, Kerala, India

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

The inherent structure of the Convolutional Neural Networks (CNN) allows them to extract features that are highly correlated with the classes while performing image classification. However, it may happen that the extracted features are merely coincidental and may not be justifiable from a human perspective. For example, from a set of images of cows on grassland, CNN can erroneously extract grass as the feature of the class cow. There are two main limitations to this kind of learning: firstly, in many false-negative cases, correct features will not be used, and secondly, in false-positive cases the system will lack accountability. There is no implicit way to inform CNN to learn the features that are justifiable from a human perspective to resolve these issues. In this paper, we argue that if we provide domain knowledge to guide the learning process of CNN, it is possible to reliably learn the justifiable features. We propose a systematic yet simple mechanism to incorporate domain knowledge to guide the learning process of the CNNs to extract justifiable features. The flip side is that it needs additional input. However, we have shown that even with minimal additional input our method can effectively propagate the knowledge within a class during training. We demonstrate that justifiable features not only enhance accuracy but also demand less amount of data and training time. Moreover, we also show that the proposed method is more robust against perturbational changes in the input images.

Keywords: Justifiable Feature Extraction ; Minimal Extra Supervision ; Image Classification ; Computer Vision ; Deep Learning.

1. Introduction

Convolutional Neural Networks (CNNs) are more popular than any other technique in image classification. The ability to automatically extract required features from the images is one key factor behind the phenomenal success of these models. Image classification being used in risk-critical areas such as medicine, surveillance, etc., CNNs could make a huge impact in these domains. However, when implementing artificial intelligence based systems in such domains, attributing the success of the application to accuracy alone is not sufficient. In such cases, these systems are expected to be justifiable as the decisions made by them may have a huge impact on various factors with high risks.

Recently, it has been observed that CNNs are very much efficient to find the correlation between features and labels and often extract features greedily following this principle (Shwartz-Ziv and Tishby, 2017; Tishby and Zaslavsky, 2015; Chaitin, 2015; Blier and Ollivier, 2018; MacKay, 2003). In this process, often these models may learn coincidental features which may not be justifiable from a human perspective (Shen et al., 2017). Let us

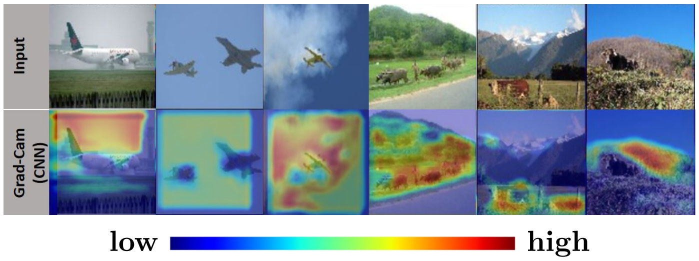


Figure 1: **CNN are correlation learning models:** To demonstrate the fact that CNN cannot distinguish between justifiable features and coincidental features we take example of aeroplane-cow classification using a biased dataset. Along with the aeroplanes, sky is also a common feature across all aeroplane images, while grass is a common feature across all cow images. It is evident from the heatmaps that the model has learnt these coincidental features for classification.

illustrate this using an extreme example. Suppose we have a dataset with images of cows on grasslands and aeroplanes in blue-sky, and that we need to perform binary classification. For a CNN model trained on such a dataset, it can be seen from the Grad-Cam (Selvaraju et al., 2017) outputs that the grass is extracted as a feature of the class 'cow', and the sky is extracted as a feature of the class 'aeroplane' (Fig. 1). Even though we may have high classification accuracy, these heatmaps reveal that the model is using irrelevant features for classification and thus lacks accountability. Such a learning of irrelevant, non-justifiable features during classification is very common in some cases like medical image datasets. One possible solution for eliminating such data bias is to refactor the dataset manually. But in cases where there is data scarcity, it may be difficult to refactor the datasets. Ideally, we expect the classification models to extract only the justifiable features from the images and use only those for classification, despite the dataset size. Conventionally trained CNNs cannot distinguish between the features that are justifiable and those which are not. To this end, our objective is to modify the learning process of the CNN and utilize the available data efficiently to reliably learn the justifiable features from the images.

It is evident from the previous example that the CNNs trained using classification loss alone, cannot guarantee the extraction of justifiable features from the images. Though we have demonstrated the issue using a very simple and trivial example, its implications in risk sensitive areas are severe. With this point of view, we propose to take guidance from humans on what they think is justifiable for a few samples in the dataset. We capture this guidance in the form of activation masks which are binary matrices with 1's representing the justifiable regions in the images. Once we have guidance from the user, we plan to tweak the learning process of the CNN to focus on extracting the required features. We achieve that by modifying the learning objective of the CNN and the backpropagation algorithm then taking care of updating the model parameters accordingly. This simple modification in the training procedure helps us to avoid the learning of spurious correlations between features and the class labels and guides the model to focus more on the justifiable ones. We have experimentally observed that this concept works quite well on a wide range of cases and has proved to be extremely useful in the case of medical datasets.

The main contributions of our work are summarized below:

1. We propose a technique to utilize domain knowledge to guide the CNNs in learning justifiable features.
2. We demonstrate that our method not just improves the learning of justifiable features but also helps in learning efficiently in situations of data imbalance and data scarcity.
3. Additionally, we also show that the features learnt using our method are more robust to various types of image perturbations.

2. Related Work

2.1. Convolutional Neural Network

Convolutional Neural Networks have boosted the progress in the field of computer vision since its inception. Manually designed architectures like LeNet (LeCun et al., 1989), AlexNet (Krizhevsky et al., 2012), VGG16 (Simonyan and Zisserman, 2014) and many more have been proposed in the literature and are shown to achieve state-of-the-art results in various computer vision tasks, beating the traditional machine learning methods by huge margins. In order to simplify the CNN architectures by retaining the spatial structure throughout the network, Springenberg et al. (2014) proposed the all convolutional nets. To interpret the decisions of the CNNs, tools like Grad-Cam (Selvaraju et al., 2017) provide a way to extract and highlight the class discriminatory features learned by the model.

2.2. Learning justifiable features

Learning justifiable features is closely aligned to learning causal features from the data. With this point of view, we also studied the literature working in the frontiers of this area. Research on the topics of correlation and causality has been lately gaining more popularity. Work by Shen et al. (2018) has recently shed some light on the correlational behavior of CNNs in image classification. The very recent work by Xiao et al. (2020), studies the influence of the image background on object recognition. They show that non-trivial accuracy can be achieved by relying just on the background features in the images. A similar study was done in the case of medical images by Maguolo and Nanni (2021), where they showed that CNN models provided a diagnosis for the chest x-ray images even when the lung regions were removed from the input images. The importance of learning justifiable and causal features, specifically in the field of medicine is studied by Castro et al. (2020) and Liu et al. (2019), highlighting the challenges in computer-aided diagnosis. Not much work has been done on improving the learning of justifiable features, especially in the case of small datasets. The closest work that we found to our method is the self-supervised method called Guided Attention Inference Network (GAIN) (Li et al., 2019) which is proposed to improve the priors for the task of weakly supervised image segmentation. The authors present an extended version of this method called the $GAIN^{p}_{ext}$, which uses an additional parameter sharing network with the GAIN architecture for pixel level supervision, that brings up the similarity with our work. We use this model as a baseline in our experiments.

Another stream of literature that deals with the task of improving the visual explanations from CNN for various tasks such as domain generalization, robustness to image

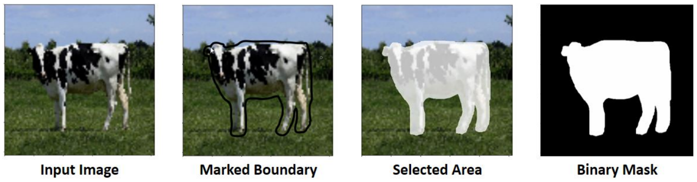


Figure 2: **Activation mask creation:** One can manually annotate and create the activation masks for guidance using an annotation script.

perturbations, etc. have been recently gaining impetus (Zunino et al., 2021; Li et al., 2021; Petsiuk et al., 2021). Work by Zunino et al. (2021) and Li et al. (2021) are very recently published. The method presented by Li et al. (2021) is attention based explainable AI (XAI) method while the one presented by Zunino et al. (2021) proposes to utilize the saliency maps generated using Grad-CAM to directly influence the intermediate activation maps during model training and make them end to end trainable similar to GAIN (Li et al., 2019). Using such operation during model training is computationally expensive. As we will soon see, our proposed approach doesn't need such computation as we include the annotation in the loss function and the model gradually adapts to that during training.

3. Justifiable Convolutional Neural Networks (jCNN)

In this section we describe the proposed method dubbed, *justifiable Convolutional Neural Network* in which we guide the CNN model to learn justifiable features. First, we briefly describe the main concept and then we move to the detailed description of activation masks and the training procedure used in jCNN.

3.1. Main concept behind jCNN

Conventionally, CNNs are trained using the categorical cross-entropy loss that optimizes the correlation between features and the class labels. Thus it is obvious that CNNs may not learn features that are justifiable from a human perspective. Hence, to force the CNNs to learn the features that are justifiable from a human perspective, we propose to guide them through domain knowledge. We provide this domain knowledge in the form of activation masks. Activation masks highlight the justifiable features in the corresponding images. We provide activation masks for a few images in each class during the training process. Then we also modify the learning objective of the vanilla CNN to consider the additional information. During the learning process, the modified model jCNN is able to generalize the information and can extract justifiable features not only from the other images in the training set but also from the unseen test images.

3.2. Activation masks to incorporate domain knowledge

Activation masks are binary images where the justifiable regions are indicated using 1's and those belonging to the context regions are indicated using 0's. We need such masks only for a few samples in each class. For manual activation mask generation, we have used the

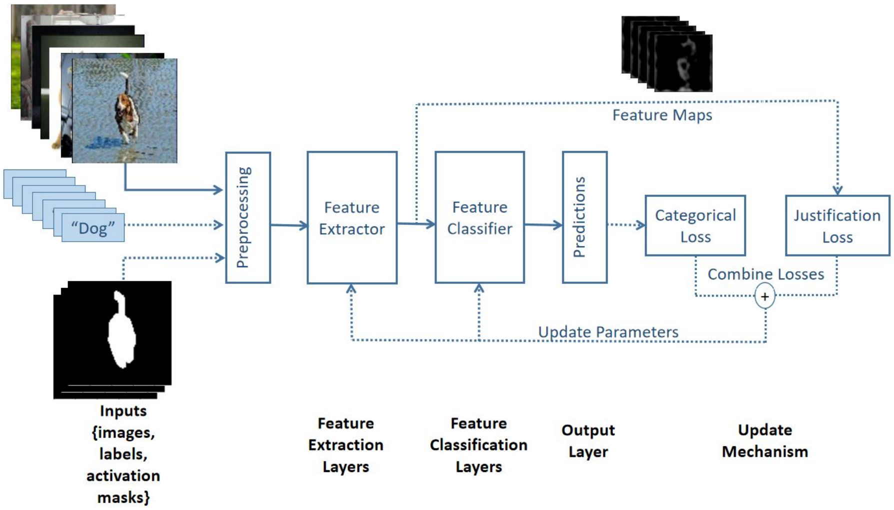


Figure 3: **Illustration of the methodology:** Along with images and labels, activation masks are also input for guidance only during the training phase. Dotted lines denote the flow present only during the training and absent during inference.

'*mpl_interactions*' library available in Python. A user has to select the regions of interest in the given image which we then convert into a binary activation mask and store along with the datasets for later use. A typical step by step procedure for activation mask generation is shown in Fig. 2.

Some datasets like the Brain MRI (Cheng et al., 2015), readily provide binary masks which can be directly used for our purpose. In few other cases, we may have pixel level labels which provide fine grained annotations exactly covering the regions of the objects of interest or bounding boxes annotations, that provide relatively coarse regions of interest. In order to use pixel level annotations as activation masks, the pixel values representing the objects of interest can be set to 1's, and the rest other pixels can be set to 0's. Similarly, when using bounding box annotations, the pixels within the bounding box regions need to be set to 1's while all other pixels outside the box regions need to be set to 0's.

3.3. Learning justifiable feature through activation masks

Convolutional Neural Networks are generally composed of several convolutional layers and pooling layers followed by fully connected layers. The convolutional layers are responsible for extracting different features from the images while the fully connected layers are responsible for classifying the input image into different classes based on these features. The better the quality of features extracted by the convolutional layers, the better will be the performance of the CNN. CNNs are trained end to end by optimizing the categorical cross entropy loss. But as discussed earlier, optimizing just the categorical loss leads to the learning of image features that highly correlate with the class labels, which in turn may not be justifiable from a human point of view.

To mitigate this issue, our approach for learning justifiable features using extra supervision is depicted in Fig. 3. During the model training along with the input images and their labels, we also provide activation masks for a subset of training images as input. The forward pass through the network for a single input image X, with true label y and input activation mask A generates the class probabilities $\hat{y}$ and the feature map outputs $\{\hat{A}_f\}_{f=1}^{F}$ from the last convolutional layer, where F is the number of filters in this layer. Using this notation, we propose to optimize the following loss to train the CNN.

$$\mathcal{L} = \underbrace{-\sum_{i=1}^{C} y_i log(\hat{y}_i)}_{L_{cl}} + \underbrace{\alpha \left(1 - \frac{1}{F} \sum_{f=1}^{F} \left(\frac{\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} (A \circ \hat{A}_f)_{j,k}}{\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} (\hat{A}_f)_{j,k} + \epsilon} \right) \right)}_{L_{jl}}, \tag{1}$$

where C is the total number of classes in the dataset, n_1, n_2 are dimensions of output feature maps and $\epsilon > 0$ is a small quantity to avoid accidental divide by zero error. $\alpha \geq 0$ is the trade-off parameter between the traditional categorical cross entropy loss (L_{cl}) and the proposed *justification loss* (L_{jl}). Also, note that '$\circ$' represents the Hadamard product of activation mask and the feature map output, after either of them is scaled to match the other's size. The greater the value of α, greater is the weightage on the justification loss. Additionally, note that for samples without additional activation masks, we just calculate the categorical cross entropy loss alone.

Apart from the traditional CNNs, we also experimented our justifiable feature learning method with the all convolutional nets proposed by Springenberg et al. (2014). These are a special type of CNN which consist only of convolutional layers skipping the need for fully connected layers in the network. The number of filters in the last convolutional layer of these nets is equal to the number of classes with each feature map output corresponding to each class, thus highlighting only that class specific features. When using jCNN with such a network we calculate the justification loss only with respect to the feature map A_c corresponding to the true class of the image and the input activation mask A as follows:

$$\mathcal{L} = \underbrace{-\sum_{i=1}^{C} y_i log(\hat{y}_i)}_{L_{cl}} + \underbrace{\alpha \left(1 - \left(\frac{\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} (A \circ \hat{A}_c)_{j,k}}{\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} (\hat{A}_c)_{j,k} + \epsilon} \right) \right)}_{L_{jl}}, \tag{2}$$

where c is the index corresponding to the actual class of the image, i.e. $y_c = 1$. This formulation can preserve the spatial structure of the data which is otherwise not maintained by the fully connected layers. Secondly, unlike in Eq. 2 where we calculate justification loss only for the feature maps that represent the true class of the image, in Eq. 1 we are calculating the justification loss on all the feature map outputs which can be comparatively much time-consuming. Additionally, note that both the losses in the case of the all convolutional jCNN are effectively calculated on the outputs of the same trainable layer in the network, leading to better parameter updates and hence better performance.

Table 1: Experimental Setup Details

Dataset	**Filters in Conv. Layers**[a]	**BS**[b]	**LR**[c]	**Epochs**	α
Oxford Pets	128,128,64,64,32,32,16,16	64	1e-4	500	0.8
Aero-Cow	128,128,64,64,32	16	1e-4	200	1.5
Brain MRI	64,64,64,32,32,32	32	1e-5	150	3
IDRiD	32,32,16,16,8,8	5	1e-4	200	0.8

[a] List representing the number of filters in the convolutional layers

[b] Batch Size [c] Learning Rate

4. Experiments

We begin this section by describing the experimental setup including the details of the dataset, baselines and other environmental settings used in our work. In all the later subsections we then present various experiments that we did to evaluate and validate the performance of our method. In that, we first discuss about the effectiveness of jCNN in reliably learning justifiable features in comparison to vanilla CNN. We then demonstrate the effects of varying the trade-off parameter introduced in the objective function of jCNN. Further, we share the results of two interesting experiments where we compare different types of activation masks and the effects of providing guidance only for a subset of classes in the dataset, respectively. And then in the next two subsections we demonstrate the additional benefits of using domain knowledge for image classification. Later, we share the quantitative as well as the qualitative results comparing jCNN with the baseline methods and then conclude the discussion on the results by showing the fact that convergence is still maintained in jCNN.

4.1. Experimental Setup

We have particularly selected the following four datasets - *Oxford IIIT Pets* (Parkhi et al., 2012), *Aeroplane-Cow*, *Brain MRI* (Cheng et al., 2015) and *IDRiD* (Porwal et al., 2018) - for performance comparison with the baselines. These datasets help us in demonstrating the effectiveness of our method across different challenges such as biased data, small dataset size and feature extraction in medical images.

In our experiments, we use the CNN traditionally trained using just the classification loss as our first baseline. As another related method that uses additional pixel level guidance, we use the extended version of GAIN (Li et al., 2019) as our second baseline. In comparison to these, we compare two of our models: one that uses fully connected layers for classification (jCNN-F) and another one that uses all convolutional nets (jCNN-C). For the purpose of comparison on a given dataset, we use same neural network architecture for all four methods. We evaluate the performance of all the models using quantitative metrics like: accuracy, macro f1-score and AU-ROC. A brief summary of the experimental setup that we used is given in Table 1. More details about the dataset and the experimental setup and a few additional results are shared in the supplementary material of the paper.

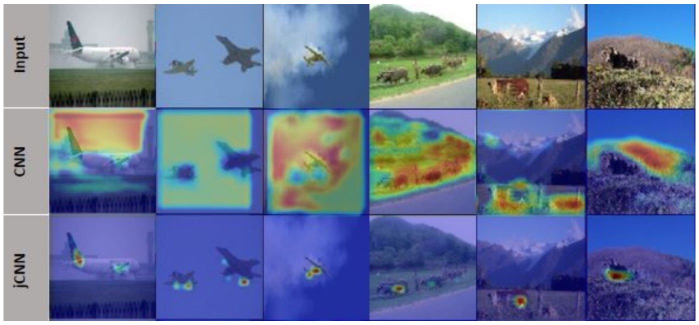


Figure 4: **Demonstrating the effectiveness of the proposed jCNN model:** Here we present the counterpart heatmap images obtained from the jCNN model. We observe that jCNN is successful in detecting the justifiable features from the images and ignores the spurious correlations present in the data.

4.2. Results: Beyond CNNs to detect justifiable features

We begin the discussion on the results by demonstrating the effectiveness of jCNN in learning justifiable features in comparison to the vanilla CNN. In Fig. 4, we present the counterpart heatmaps obtained from jCNN model in comparison to those shared in Fig. 1. The results demonstrate the effectiveness of jCNN in learning justifiable features in comparison to the traditionally trained CNNs. One peculiar property of jCNN to note here is that they learn only the *most class discriminatory justifiable features* present in the images leading to heatmaps that are concentrated in small regions. Additionally, it can be noted that both these models have comparative classification accuracy (Table 2), yet jCNN can be considered to be accurate as well as justifiable unlike the traditionally trained CNN.

4.3. Results: Effect of varying the trade-off parameter in jCNN loss

We performed an experiment to investigate the sensitivity of the setup of human guidance. The trade-off parameter (α) is an important quantity as it trade-offs between the classification loss and the justification loss (Eq. 1, 2). We expect that as the value of the trade-off parameter increases, the model should focus more on the justifiable features present in the images while reducing their dependence on the context features. We see the expected behaviour in Fig. 5.

4.4. Results: Experimenting with Different Mask Types

In this experiment, we compare the performance of jCNN models trained using different type of masks. The objective is to identify how much detailed the masks used in jCNN should be for the models to work as expected. We created a binary class classification dataset for two penguin classes from the Birds 400 dataset available on Kaggle. As shown in Fig. 6 (a), we created four types of masks namely - *coarse, fine-grain, important region* and *non important region* masks. Along with the objects of interest, *coarse* masks can have few background features in them but they are also easy to create. *Fine-grain* masks

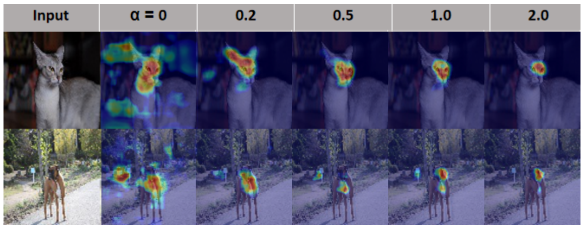
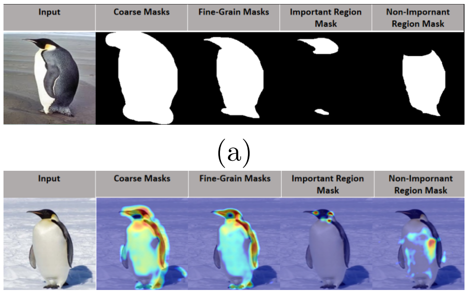


Figure 5: **Illustration of the effects of the trade-off parameter (α) in the jCNN loss :** With an increase in the value of this parameter, the model can focus more on the *class-discriminatory justifiable features* in the input images, i.e. the dependence on the context features present in the images is reducing as expected.



(a)

(b)

Figure 6: **Experimenting with different types of masks:** Here we use the example of classifying images into two penguin species. We created four different types of masks covering different amount of information: *Coarse*, *Fine-grain*, *Important region* and *Non-Important region* masks as shown in (a) and trained four models with these masks. In (b) we present the heatmaps of a test input image.

cover the exact regions of the objects of interest but are relatively hard to create. *Important region* masks are similar to fine grain masks just that they highlight only the important (i.e. most class discriminatory) regions from the objects of interest. Similarly, the *non important region* mask highlight the less class discriminatory regions from the objects of interest.

In Fig. 6 (b), we present the results obtained from the jCNN models trained using these different mask types. Our first observation is with respect to the first three mask types, where we observe that the heatmaps outputs in all the three cases reveal that the models are more or less focusing on similar features from the objects of interest. Additionally, in the case of first two mask types, jCNN is also using few other features from the objects of interest as they are allowed to do so based on the mask types provided as input. It must also be noted that the classification accuracy for all these three models is similar. This brings us to the conclusion that, even though we use coarse masks, jCNN is still able to learn only the justifiable features from the objects of interest. This is important because, such masks are comparatively easy to create. To validate this further, results obtained in helmet detection dataset (MakeML, 2020) where we used coarse masks obtained

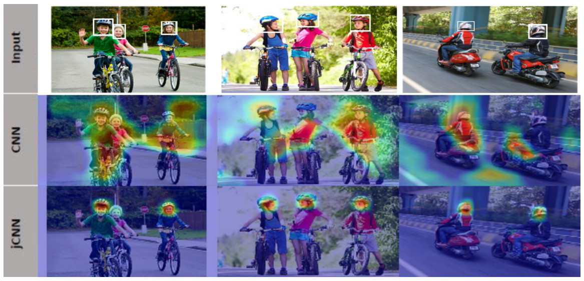


Figure 7: **Results on Helmets(MakeML, 2020) dataset:** We observe that we achieve relatively good visual results when compared with the vanilla CNN model trained using the same dataset containing very few samples.

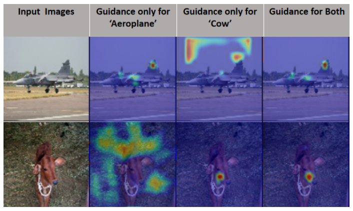


Figure 8: **Effect of partial domain knowledge during learning:** For classes without any domain knowledge like CNN, jCNN also fails to extract justifiable features.

from bounding box annotations are also shared in Fig. 7. In addition to this, our second observation is with respect to the important region masks. Good quantitative as well as qualitative performance from model trained with such masks suggests that, only highlighting the important regions from the objects of interest is also enough for jCNN to work as expected. Lastly, in the case of non-important region masks, as we do not include the most class discriminatory features from the objects of interest in the input activation masks, the model performance significantly degrades especially in terms of classification accuracy which is a trivial observation that one can expect.

4.5. Results: Effect of partial domain knowledge

Through this experiment, we wanted to investigate the situation when we provide domain knowledge for a subset of classes. To demonstrate this, we again take the example of the aeroplane-cow classification. We trained three jCNN models: one with guidance only for the aeroplane class, another with guidance only for the cow class and the last one with guidance for both the classes. The resultant heatmaps from the three models are presented in Fig. 8. It can be observed that, when we provide guidance for only one of the classes, the other class can still suffers from the issue of learning non-justifiable features. Additionally, we can also see that when we provide guidance for the both classes, the model performance improves for both. Thus we can conclude that the learning of features for one class is independent of the masks provided for another classes.

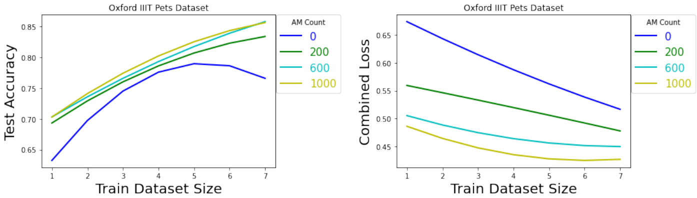


Figure 9: **Ability of jCNN models to learn more with less amount of data:** The above plot shows test results on the Oxford IIIT Pets dataset. With additional guidance, jCNN is able to learn efficiently with same or even less amount of data.

4.6. Results: Learn more with less data

Through this experiment, we show that our method helps not just in learning justifiable features but also helps in learning faster with fewer data. The basic idea is, with more guidance, we reduce the uncertainties in the model for learning justifiable features as the model is explicitly guided to focus on the features that are effective to discriminate. To demonstrate this, we trained jCNN models with same architectures by varying the training dataset size in steps of 1000, from 1000 to 7000 training samples. In each case, we train models with 0, 200, 600 and 1000 activation masks as input. So basically we train 28 (7×4) models in total. For a given dataset size, we expect the models with more activation masks as input to have better performance than those with less number of activation masks as input. We share the results on the unseen test data in Fig. 9 where the classification accuracy and combined loss (arithmetic mean of the categorical cross entropy loss and the justification loss) are used as a metrics for model comparison. Combined loss is a better metric for comparison as it captures the correctness of the models both in terms of classification as well as in terms of learning justifiable features.

We observe that, for a given dataset size, the models with more activation masks as input have lower loss (on test dataset) than the models trained with less number of activation masks as input. This suggests that we can learn more with the same amount of data when provided with more guidance in the form of activation masks. Additionally, we can also see that the lowest loss (on test dataset) attained using 7000 training samples by the models with no activation masks as input is also attained by the other models with just 200 activation masks as input using around 3000 to 4000 training samples. This suggests that with more guidance the models are able to learn more from less amount of data. These two are very crucial observations demonstrating the effectiveness of jCNN in learning efficiently from less amount of data.

4.7. Results: Robustness against simple image perturbations

In the first experiment, we vary the background of the objects of interest and check how our models perform in terms of visual results when compared with both baselines. As the traditionally trained CNNs may rely on the background features for class prediction, they are the ones which are the most affected by the changes in the background of the images. This is evident from the results shared in Fig. 10. On the other hand we expect that, in

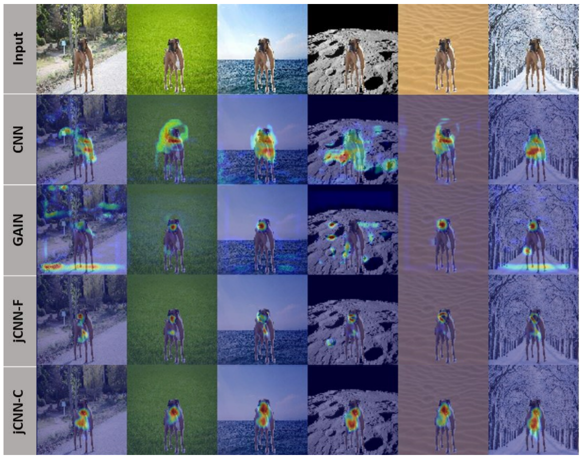


Figure 10: **Learning robust features for changes in the background of the input images:** The above figures show the robustness of feature learning in jCNN against changes in the background of the input images.

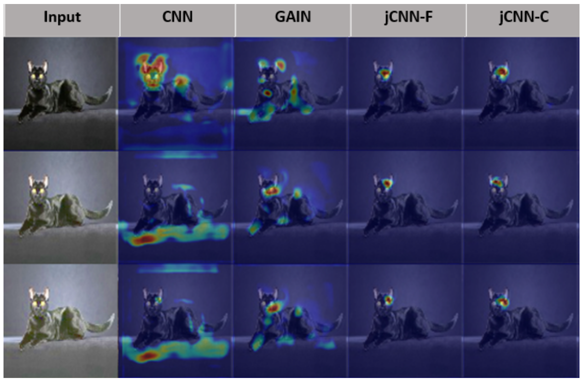


Figure 11: **Learning robust features for changes in the brightness of the input images:** The above figures show the robustness of feature learning in jCNN against changes in the brightness of the input images.

comparison to the CNNs, the features learnt by jCNN would be more stable across such image perturbations as they are capable of extracting only the justifiable features from the images. Again, we observe the expected behavior in the heatmap outputs for the jCNN. This further confirms the effectiveness of our method in learning justifiable features.

Similarly, we also conducted another experiment in which we vary the brightness of the input images as shown in the first row of Fig. 11. Even though, we do not observe significant changes in the input images, we observe that the heatmaps obtained from the CNN vary significantly with such small changes in the input. Such performance raises concerns for the practical use of these models, because in real life practices we expect the data to have much more variations than in a fixed environment used for training these models. Contrarily, we

Table 2: Quantitative results comparing jCNN with the baselines.

Models		**Metrics**	
	Accuracy	**F1-score**	**AUROC**
		Oxford IIIT Pets	
CNN	**0.93±0.014**	**0.92±0.016**	0.97±0.004
GAIN	0.86±0.015	0.84±0.015	0.92±0.004
jCNN-F	0.87±0.015	0.87±0.015	0.97±0.005
jCNN-C	0.91±0.017	0.91±0.016	**0.97±0.003**
		Aeroplane-Cow (Small Dataset)	
CNN	0.74±0.017	0.74±0.015	0.84±0.004
GAIN	0.70±0.022	0.69±0.021	0.72±0.015
jCNN-F	0.76±0.029	0.75±0.018	0.83±0.015
jCNN-C	**0.81±0.012**	**0.80±0.010**	**0.86±0.003**
		Brain MRI (small dataset)	
CNN	0.83±0.033	0.78±0.069	0.96±0.006
GAIN	0.68±0.020	0.61±0.040	0.85±0.20
jCNN-F	**0.85±0.004**	**0.83±0.004**	0.94±0.004
jCNN-C	0.85±0.005	0.82±0.002	**0.96±0.002**
		IDRiD (small dataset)	
CNN	0.80±0.029	0.81±0.016	0.67±0.111
GAIN	0.82±0.009	0.86±0.010	0.71±0.129
jCNN-F	**0.84±0.065**	**0.87±0.053**	**0.73±0.118**
jCNN-C	0.82±0.036	0.83±0.051	0.59±0.151

see that the features learnt by jCNN are more robust in comparison to that of CNNs, for different brightness values of the input images.

4.8. Results: Experiments on benchmark datasets

In this section, we compare the quantitative and the qualitative results obtained on various benchmark datasets. Table 2 compares the performance of various models on several metrics. It can be seen that both jCNN variants perform quiet well especially in the case of very small datasets (*Aeroplane-Cow, Brain MRI and IDRiD*). Further, to validate the correctness of our models in terms of feature learning, we present the comparison of the Grad-CAM heatmaps generated by these models in Fig. 12. We observe that jCNN are not just accurate but also rely on the class-discriminatory justifiable features in the images for classification rather than using any context information.

4.9. Results: Convergence in jCNN

Last but not the least, we would also like to discuss about the convergence of jCNN. As we alter the learning objective in jCNN, we would like to check if the convergence is still maintained or not. We show this through the training plots obtained in our experiments. Fig. 13 shows the training plots of a jCNN model trained on the *Oxford IIIT Pets* dataset.

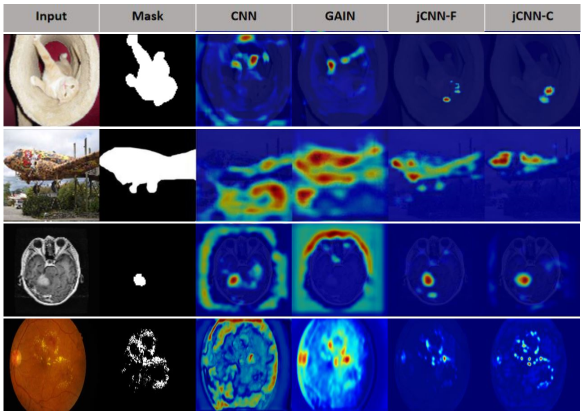


Figure 12: **Qualitative results:** This figure shows the effectiveness of jCNN in learning of justifiable features in the case of different benchmark datasets.

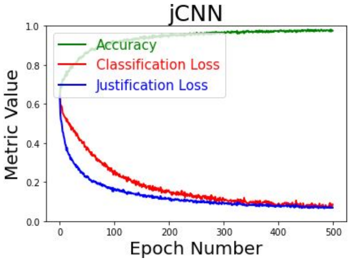


Figure 13: **Convergence of jCNN :** The above training plot on the Oxford Pets dataset shows a steady convergence of the jCNN model.

It can be observed that after few epochs, both the losses in jCNN begin to converge. We have observed similar plots in the case of other datasets as well. This ensures that adding the justification loss does not adversely affect the training of the underlying CNN and shows steady convergence based on the new loss function.

5. Conclusion and Future Work

In this paper we demonstrate that, through minimal domain knowledge it is possible to guide the traditional CNN models to learn the justifiable features from the images. This ability in turn helps jCNN not only to achieve higher accuracy with less data but also to be robust against simple perturbational changes in the images where vanilla CNN usually fails. One limitation that we came across during our experiments was that in the case of datasets where the justifiable features are very small or subtle, jCNN struggles to learn such feature. Eventhough it is inherent from CNN, in future we would like to work on the improvements in this direction. We would also like to investigate methods other than CNN where we can adopt our approach to make them end-to-end trainable for justifiability. We have made our code available at: https://github.com/BRim28/jCNN

References

Léonard Blier and Yann Ollivier. The description length of deep learning models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in NIPS*, volume 31. Curran Associates, Inc., 2018.

Daniel C Castro, Ian Walker, and Ben Glocker. Causality matters in medical imaging. *Nature Communications*, 11(1):1–10, 2020.

Gregory Chaitin. *On the intelligibility of the universe and the notions of simplicity, complexity, and irreducibility.* Akademie Verlag, 2015.

Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, 10(10):e0140381, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in NIPS*, 25:1097–1105, 2012.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Guided attention inference network. *IEEE transactions on pattern analysis and machine intelligence*, 42 (12):2996–3010, 2019.

Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. Scouter: Slot attention-based classifier for explainable image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1046–1055, 2021.

Yuanyuan Liu, Zhouxuan Li, Qiyang Ge, Nan Lin, and Momiao Xiong. Deep feature selection and causal analysis of alzheimer's disease. *Neuroscience*, 13:1198, 2019.

David JC MacKay. *Information theory, inference and learning algorithms.* Cambridge university press, 2003.

Gianluca Maguolo and Loris Nanni. A critic evaluation of methods for covid-19 automatic detection from x-ray images. *Information Fusion*, 76:1–7, 2021.

MakeML. Helmets Dataset | MakeML create neural network with ease, jan 13 2020. [Online; accessed 2022-08-19].

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11443–11452, 2021.

Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, and Fabrice Meriaudeau. Indian diabetic retinopathy image dataset (idrid), 2018. URL https://dx.doi.org/10.21227/H25W98.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Zheyan Shen, Peng Cui, Kun Kuang, and Bo Li. On image classification: Correlation vs causality. *arXiv preprint arXiv:1708.06656*, 2017.

Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. Causally regularized learning with agnostic data selection bias. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 411–419, 2018.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 IEEE Information Theory Workshop (ITW)*, pages 1–5. IEEE, 2015.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Andrea Zunino, Sarah Adel Bargal, Riccardo Volpi, Mehrnoosh Sameki, Jianming Zhang, Stan Sclaroff, Vittorio Murino, and Kate Saenko. Explainable deep classification models for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3233–3242, 2021.