Advancing Post-Hoc Case-Based Explanation with Feature Highlighting Appendix

Author Name
Affiliation
pcchair@ijcai-23.org

1 Experiment 2: Training Hyperparamters

For the training ablation studies ResNet50 and ResNet34 was used. ImageNet experiments used the pre-trained ResNet50 model available on Pytorch. CUB-200 fine tuned ResNet34 with the following hyperparamters.

CUB-200. Epochs 500. 1 GPU used. Number of workers was 2. Batch size 12. Data transformations during training were done with Pytorch transforms: RandomResizedCrop(224), RandomRotation(45), RandomHorizontalFlip(0.5), ColorJitter(brightness=0.126, saturation=0.5). A multiplicitive learning rate decay of 0.999 was used. The epoch with best test accuracy was used for the model, which was epoch 41.

ImageNet. This was a pre-trained model from PyTorch.

Standard training and testing splits were used for ImageNet, and CUB-200. However, as is standard practice, the validation data was used for ImageNet, since the test data labels are unavailable.

Fine-Tuning. When finetuning the CNNs, the exact same hyperparamters were used (and the same ones from the original ResNet paper for ImageNet), except that the learning rate was decreased by 10^{-1} from each methods initial learning rate when training the initial CNN model.

2 Computational Costs

2.1 Hardware

Expt. 1 had the CNN trained on a single Nvidia K80 GPU, 2vCPU @ 2.2GHz, and 13GB RAM. The experiments were subsequently run on MacBook Pro, processor 2.9 GHz Intel Core i5, memory 16 GB 2133 MHz LPDDR3.

Experiments 2-3 were run on a Dell R740XD with an Nvidia V100 (32GB) GPU: 256GB RAM. Storage of the dataset was on scratch storage 220TiB.

Computational time for experiments. Expt. 1 took approximately 48hrs and 6hrs to run for ImageNet and CUB-200, respectively, presuming the use of a single Nvidia V100 (32GB) GPU. Expt. 2 took 2+ weeks to run the beta experiment on ImageNet, one day for the alpha, and 2+ weeks days for the comparative tests. CUB-200 by comparison took less than 2 days for all tests.

3 Example Explanations

Here more example explanations are showcased with correct classifications, and incorrect classifications. In addition, a misclassification with three salient regions and multiple NNs (rather than just one) to illustrate this explanation option is shown.

Fig. 1 shows six correct classifications on ImageNet by ResNet50, alongside an explanation for them. Firstly, the explanation comprises of a nearest neighbor from a pool of n candidate cases retrieved (50 in our experiments), which alone is already considered a "good" explanation by non-experts [Jeyakumar *et al.*, 2020; Kenny *et al.*, 2021]. However, the explanations go further by pinpointing a salient region in the test image which was "learned" from the NN. This type of explanation not only informs the user of what important feature was used in the explanation, but also from *where* it arose in the first place so it may be further contextualized.

Fig. 2 shows two incorrect classifications in ImageNet from our user study. The first image is a "Kimono" misclassified as a "Violin", were the CNN confused the pipe in the test image with a violin bow. Fig. 2b shows a "Hammer" misclassified as a "Shovel". Here the CNN saw similarity in the test image's wooden handle to a previous training image of a "Shovel", and hence classified the image as a shovel.

Fig. 3 shows another two incorrect classification in ImageNet from our user study. Fig. 3a shows an "Acoustic Guitar" which is misclassified as an "Electric Guitar". The CNN conflates the fretboard as being indicative to an electric guitar, but it is also important to an acoustic guitar, and hence the misclassificaiton arises. Fig. 3b shows a misclassification of a "Flute" as a "Horizontal Bar". The visual similarity of the horizontal bar in the NN image is confused with the wooden flute in the test image.

Fig. 4 shows another way of visualizing salient regions with multiple NNs shown. Specifically, the three most salient salient regions are shown alongside the three closest representations of them in the pool of NNs retrieved. Interestingly, in Fig. 4, the third salient region seems to be confused between a Leopard's tail and leg, possibly contributing to the misclassification.

Fig. 5 shows another example of using three latent-based salient regions for the correct diagnosis of Covid-19 in a patient's x-ray.

Fig. 6 shows another example of using SP-salient regions

to explain misclassifications on ImageNet. (A) A "Knot" and (B) "Rifle" are misclassified by ResNet50 apparently due to bias in the CNN by associating, what could be called, "wooden background", and "snowy background" with the respective classes.

References

[Jeyakumar *et al.*, 2020] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. How can i explain this to you? an empirical study of deep neural network explanation methods. *Advances in Neural Information Processing Systems*, 33, 2020.

[Kenny *et al.*, 2021] Eoin M Kenny, Courtney Ford, Molly Quinn, and Mark T Keane. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in xai user studies. *Artificial Intelligence*, page 103459, 2021.

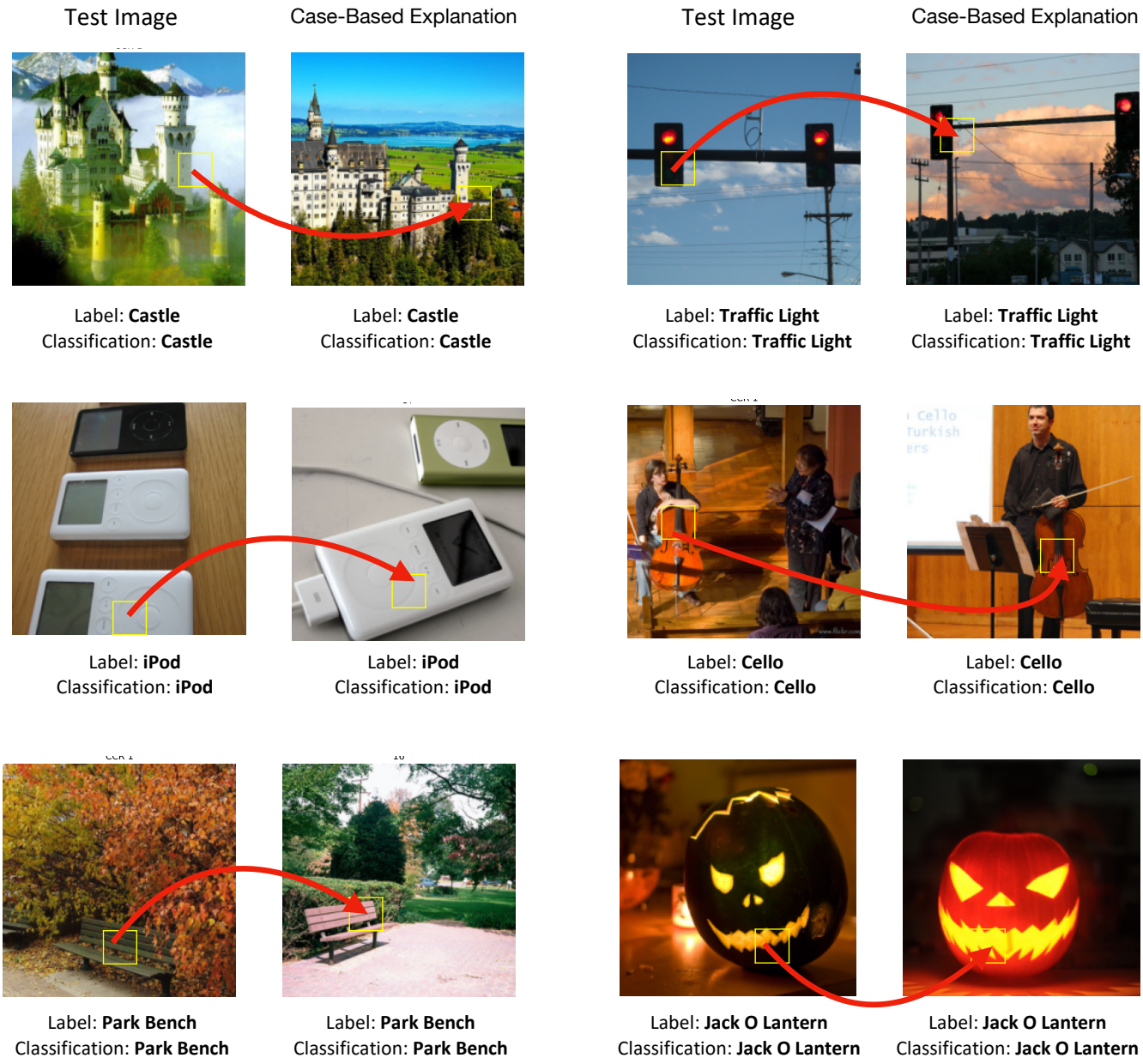

Figure 1: Correct examples: Starting from the top left (and going in reading order), we see correct classifications and a NN explanation showing the salient region used in the classification. Namely, the images show correct classifications of a "Castle", "Traffic Light", "iPod", "Cello", "Park Bench", and "Jack O Lantern"

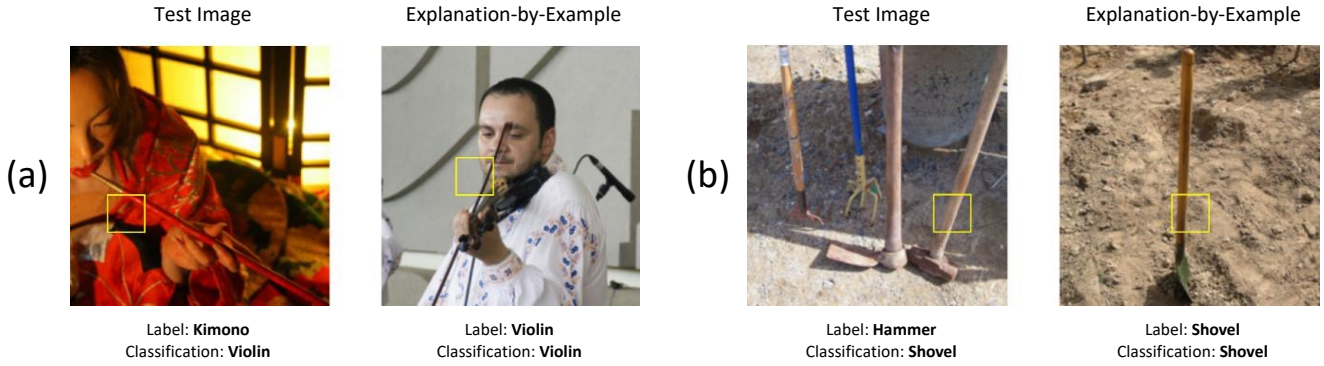

Figure 2: Incorrect examples: (a) A "Kimono" is misclassified as a "Violin", the salient region shows the CNN confused the pipe in the test image as the violinist's bow. (b) A "Hammer" is misclassified as a "Shovel", the salient region shows the CNN learned to focus on the wooden handle of shovels when classifying them, which it partly learned from the training image shown.

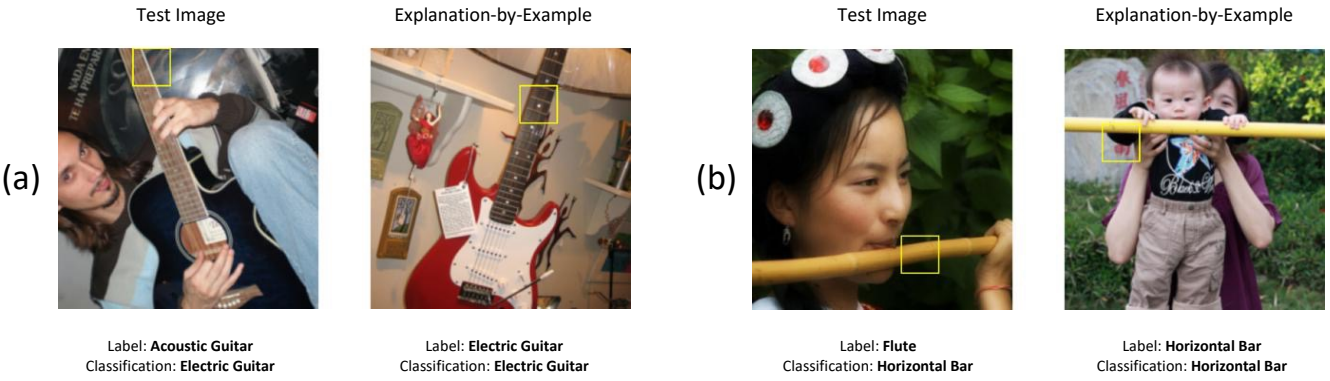

Figure 3: More Incorrect examples: (a) An "Acoustic Guitar" is misclassified as an "Electric Guitar", the salient region shows the CNN has learned to associate the guitar's fretboard with electric guitars, and neglected the rest of the image when classifying the test image. (b) A "Flute" is misclassified as a "Horizontal Bar", the salient region shows the CNN seems to have focused on the qualitative similarity between the horizontal bamboo bar in the training image, and the bamboo flute in the test image.

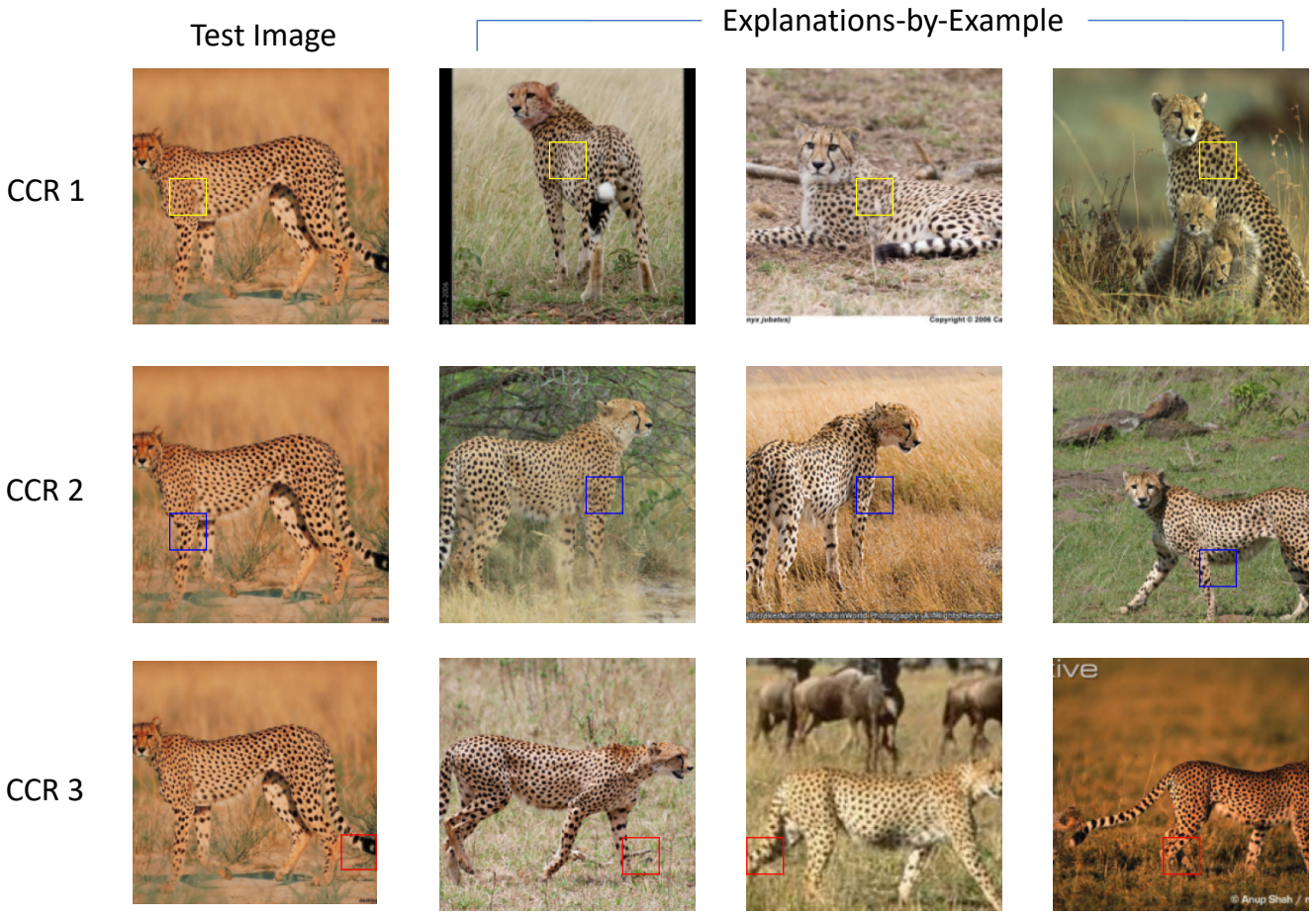

Figure 4: Multiple salient regions: An incorrect classification of a "Leopard" as a "Cheetah". Three of the most salient salient regions are shown for the test image, alongside their three closest representations in the pool of NNs. Showing multiple close matches for the test image salient regions helps contextualize the feature further which may be useful, especially if the L_2 difference between the NN-salient regions is quite small.

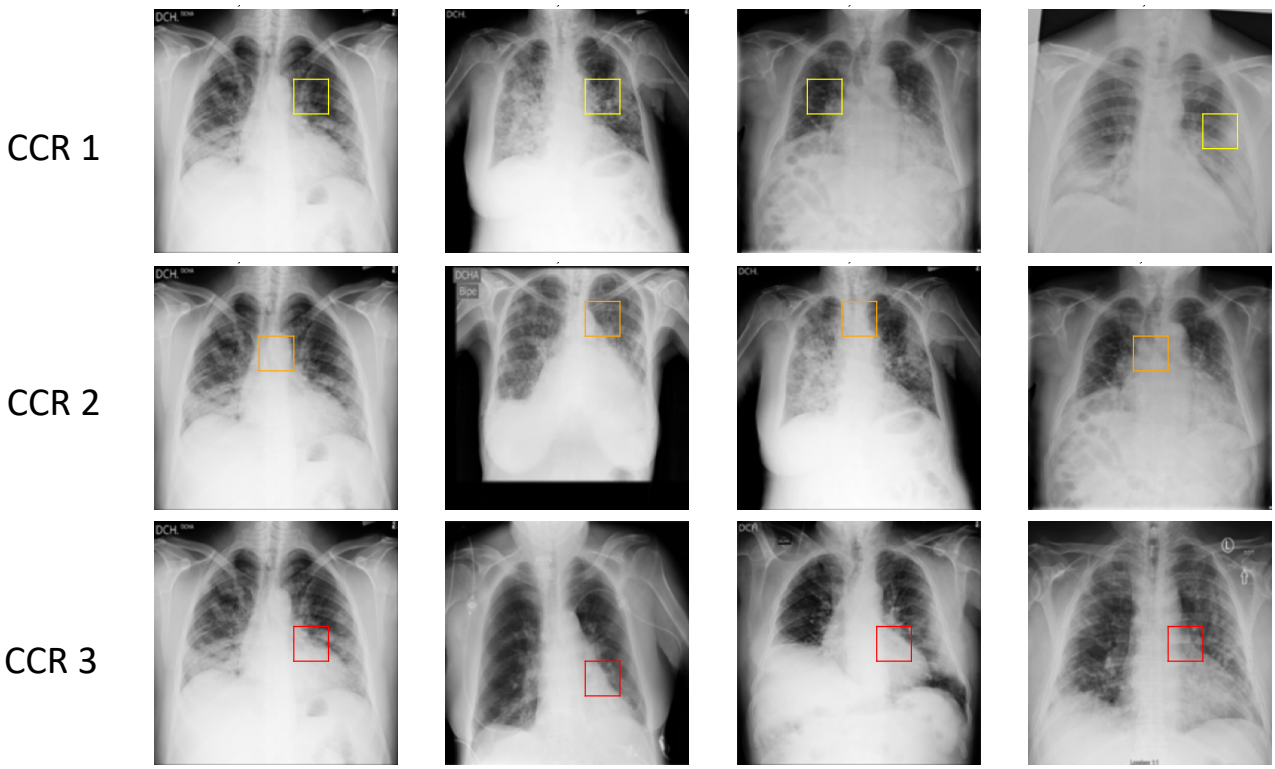

Figure 5: Multiple salient regions: A correct classification of "Covid-19". Three of the most salient salient regions are shown for the test image, alongside their three closest representations in the pool of NNs. Showing multiple close matches for the test image salient regions helps contextualize the feature further which may be useful, especially if the L_2 difference between the NN-salient regions is quite small.

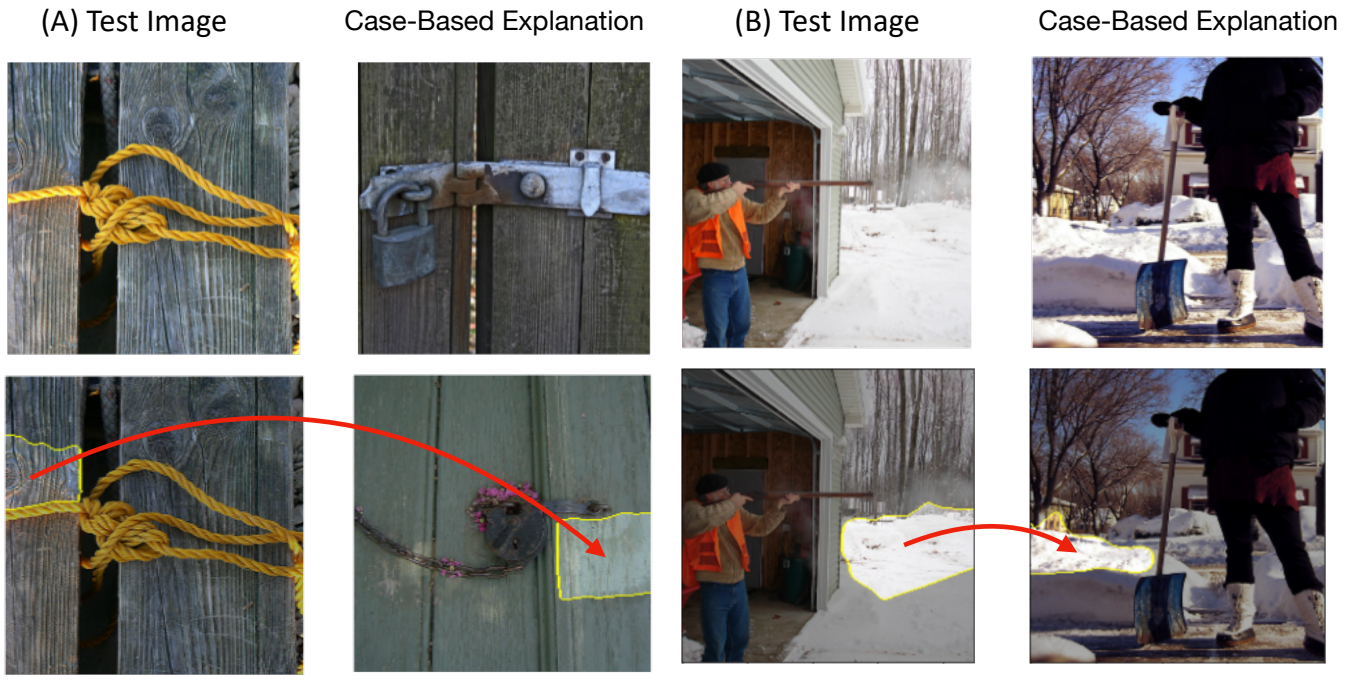

Figure 6: Incorrect ImageNet Classifications: SP-salient regions are used to explain two misclassification (A) A “Knot” is misclassified as a “Padlock”, were the CNN has learned to associate a “wooden background” feature with the predicted class, causing a bias. (B) A “Rifle” is misclassified as a “Shovel”, where a bias in the CNN has learned to associate a “snowy background” feature with the class “Shovel”.