Squad-CAM

Andree Hultgren **Topias Tyystjärvi** **Ravi Bir**

Abstract

The ability to explain deep convolutional neural networks allows for explanations and insights into the model. Extracting this information can help with optimization, bias minimization and much more. Grad-CAM has been widely used in many situations. This paper proposes a method to square the weighted linear combination of Grad-CAM to create Squad-CAM. This proposed method yields a slightly more accurate and intuitive explanations measured with both error measure and human studies.

1 Introduction

Deep learning models produce predictions that are often difficult to interpret. As they get more profound and more complex, they become less intuitive and interpretable to humans. The decrease of interpretability lowers the trust the user has in a model and makes them less confident in the decisions that are made based on the models' predictions. Including information about the reasoning behind a prediction would make models more useful and trustworthy for human users and would provide insight into possible ways the model could be improved.

Applying explanation methods to computer vision tasks would be very beneficial. The use of Convolutional Neural Networks (CNNs) has seen breakthroughs in areas such as image classification, object detection and image captioning. However, the models used for these tasks are highly complex and lack interpretability. Grad-CAM (1) was developed as a solution to this. It produces visual explanations for any CNN based model for a variety of computer vision tasks. Retraining of the model is not required, and its architecture remains unchanged.

In this paper, Grad-CAM is reimplemented as well as Guided Backpropagation (2). These methods have been applied to VGG-16 and Resnet50. We experimented with different combinatory methods and activation functions to produce a localisation explanation. The resulting explanation method is given the acronym "Squad-CAM", which is short for Squared Grad-CAM. This explanation method was evaluated through human studies and localisation error through the correct labels.

2 Related Work

Research has been conducted into alternate methods to generate explanations for CNNs. Simonyan et al. (3) utilise effective pixels (pixels that make the most considerable contribution to the prediction). The partial derivative of a class' score with respect to the pixel intensities is used to visualise a CNN prediction. However, it was shown that this method, and methods similar to it, were not class-discriminative (4).

Another alternative is a weakly-supervised localisation approach such as Class Activation Mapping (CAM) (5). In order to produce class-specific feature maps, unlike Grad-CAM, CAM makes modifications to the CNN architecture. The fully connected layers are replaced with convolutional layers and global average pooling. CAM can only be applied to a specific type of CNN where the prediction is made directly after the global average pooling stage. This restriction severely limits the application of CAM and demonstrates the need for Grad-CAM, which can be applied to any CNN. It can be shown that Grad-CAM is a generalisation of CAM (1).

Project report for DD2412 or FDD3412

3 Methods

3.1 Guided Backprop

Guided backprop (2), as described by Springenberg, entails replacing the backprop of gradients in a certain way. Both Resnet and VGG-16 use ReLU activation layers in their models. ReLU and the guided backpropagation are described in equation 1 and equation 2 respectively.

$$f_i^{l+1} = ReLU(f_i^l) = max(f_i^l, 0) \tag{1}$$

$$R_i^l = (f_i^l > 0) \cdot (R_i^{l+1} > 0) \cdot R_i^{l+1} \tag{2}$$

When the backprop has reached the input layer, it can be used to explain the decision of the model.

3.2 Grad-CAM

Grad-CAM focuses on highlighting the regions of an image that contribute to the classification of the model. Each neuron in the last convolutional layer of the CNN is assigned a weight based on how important it is in making the desired classification decision. In line with other research, the last layer is used because it has the "best compromise between high-level semantics and detailed spatial information."(1)

The neuron importance weights, α_k^c, are calculated using equation 3

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3}$$

where A^k is the feature map activations of the last convolutional layer, c is the class, k is the feature map, Z is the number of pixels and (i, j) is the pixel location. The localisation map of class c is computed by a weighted combination of forward activation maps. A ReLU activation function is then applied to remove all negative values since we are only interested in pixels that have a positive influence on c, as described in equation 4.

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k) \tag{4}$$

The guided Grad-CAM is where $L_{Grad-CAM}^c$ is upscaled with bilinear interpolation to the same size as the guided backprop image. Elementwise multiplication between the localisation and guided backprop results in the Guided Grad-CAM explanation.

3.3 Squad-CAM

One of the issues with equation 3 is that negative neuron importance weights alter the resulting localisation. A negative weight combined with a positive feature value would disable the localisation at that position whilst a negative weight combined with a negative feature would enable the localisation at that position.

In order to alter this behaviour, multiple different combinations and activation functions were tested as a replacement for the linear combination seen in equation 4. Different activation functions and combinations were tested, see Appendix A for these tests. The method that proved to give a better and more accurate localisation was squaring the neuron importance weights as described in equation 5.

$$L_{Squad-CAM}^c = ReLU(\sum_k {\alpha_k^c}^2 A^k) \tag{5}$$

The combination is no longer a linear combination. With the squared weights, the values that are allowed through the ReLU activation functions have changed. This change is visualized in figure 1.

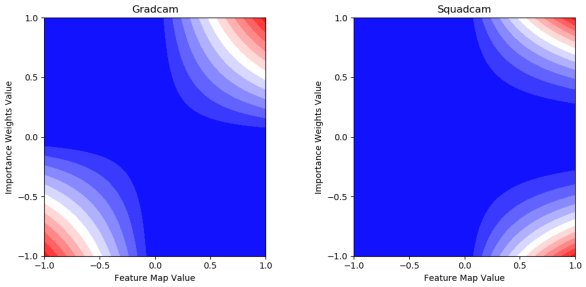


Figure 1: A comparison of the different combination methods used in Grad-CAM and Squad-CAM for feature map k at position i, j.

The logical explanation is that the combination should focus on the value of the importance weight rather than the sign. The squared method proved to be a viable method of doing so.

The guided Squad-CAM is where $L^c_{Grad-CAM}$ is upscaled with bilinear interpolation to the same size as the guided backprop image. Elementwise multiplication between the localisation and guided backprop results in the guided Squad-CAM explanation.

4 Data

As Grad-CAM and Squad-CAM do not require training, no data is needed for the development of the method. Data was needed for human studies that will evaluate the accuracy of the explanation method. The data used in this project are cherry-picked images from the "ImageNet 2012 val" and "Pascal-VOC val" datasets. The selected images were images with at least two possible different classifications. Random samples of the "Imagenet 2012 val" set were used for the localisation and failure mode experiments.

The models that will be under investigation for the explanation methods are pre-trained versions of VGG-16 and Resnet50. Since both models require input images of size (224 x 224), each image is rescaled to fit the desired input size.

5 Experiments and Findings

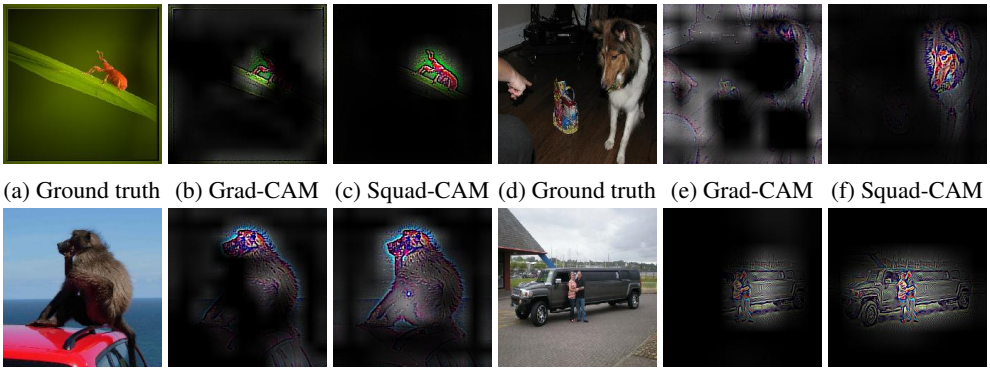
We performed two human studies and a localisation experiment. In order to evaluate the explanatory abilities of each method, humans needed to be involved. The studies performed were designed in the same way as the studies made in Grad-CAM. For human studies, a website was used to present the questions and record the participants' answers. The link to the website is `http://squadcam.hultan.com`. The website will remain active for a while, so please visit the link to see how this was done. Some examples of our results can be seen in figure 2.

5.1 Explanatory accuracy

In the first study, the objective was to observe how well the participants understood the explanations given. The quality of explanations should reflect the accuracy of the models. In total, 90 people were asked 30 questions each from a set of 606 possible permutations of images, models and explanation method providing a total of 2700 data points. The distribution over models, images and explanation methods were uniform.

The target group was presented with an explanation image from either explanation method and model. The user was then provided with choices of two objects present in the ground truth image. The user was asked to pick which option they see on the explanation visualisation. The accuracy of the predictions can be seen in Table 1. This accuracy is a direct reflection of the explanatory accuracy of each method applied to a model.

From Table 1 it is evident that the proposed method "Guided Squad-CAM" is the best at explaining the decision of the VGG-16 model. Humans were able to correctly identify the category 79.33% of



(a) Ground truth (b) Grad-CAM (c) Squad-CAM (d) Ground truth (e) Grad-CAM (f) Squad-CAM

(g) Ground truth (h) Grad-CAM (i) Squad-CAM (j) Ground truth (k) Grad-CAM (l) Squad-CAM

Figure 2: Example results from Guided Grad-CAM and Guided Squad-CAM for different ground truths.

Table 1: Explanatory accuracy for method and model.

	Guided Backprop	Guided Grad-CAM	Guided Squad-CAM
VGG-16	64.81%	76.16%	**79.33%**
Resnet50	61.58%	**73.27%**	72.20%

the time, which is higher than both Guided Grad-CAM (76.16%) and Guided Backprop (64.81%). It can be concluded that Guided Squad-CAM is the most class-discriminative explanation method. The results of the paper have been successfully reproduced, with Guided Grad-CAM giving a higher human accuracy than Guided Backpropagation (1). The higher accuracy values obtained in our experiment can be attributed to the slightly different datasets used. However, we have improved on the results given in the paper with Guided Squad-CAM, which produced a human accuracy 3.17% higher than Guided Grad-CAM. For the Resnet50 model we can see that Guided Backprop has a significantly lower accuracy (61.58%) compared to the other two methods. However, for this model, the accuracies of Guided Grad-CAM and Guided Squad-CAM are so similar (1.07% difference) that we can not distinguish which explanation model is more accurate and class discriminative than the other. These results may be due to Resnet50s last convolutional layer being lower resolution (7x7 as opposed to 14x14 for VGG-16), which indicates an architecture-dependent limitation in applicability.

Combining the results from all experiments and comparing Guided Grad-CAM with Guided Squad-CAM, we get a statistical p-value of 0.2692. The p-value value can be extracted from results on the project website as described in section 5. This p-value is not enough to conclude that Guided Squad-CAM outperforms Guided Grad-CAM.

5.2 Establishing trust

In the second study we evaluated the trustworthiness of the explanation methods. It is known that Resnet50 is a slightly more accurate model than VGG-16. This experiment aims to see whether humans with no knowledge of the models can correctly come to this conclusion. We will see if they can guess the more accurate model using only the visualisations produced by the explanation methods.

For this experiment, 66 different people were asked to compare two explanations. Only images where both models predicted the same label was used. For each image, the human was shown the explanation produced from the two models (the same explanation method is used for both models). Using only these explanations, they will then be asked to rate how reliable each model is compared to the other. A score is given based on the given answer. A score of ± 2 is given if one model is deemed more or less reliable than the other. A score of ± 1 is given if one model is deemed slightly more or less reliable than the other. Finally, a score of 0 is given if both models seem equally reliable. To avoid bias, each model has a 50% chance of being presented as "Model A".

The results in Table 2 are different from those obtained by the paper (1) since we do not compare VGG-16 with Alexnet, but rather with Resnet50. In our experiments, Guided Backprop with VGG-16 received an average score of 0.52, which means the participants deemed VGG-16 slightly more reliable than Resnet50. Guided Grad-CAM and Guided Squad-CAM, both had average scores close to 0 (-0.19 and 0.17). This negligible difference means that one model was not deemed more or less trustworthy than the other.

It can be concluded that the visualisations produced by our implemented explanation methods could not be used to determine that Resnet-50 is a more accurate classifier than VGG-16, and so they fail in helping the user place trust in the correct model. In the paper, the more accurate classifier could be identified using the explanations. The reason for this disparity in results could be because different models were used. There is a more considerable difference in model performance between VGG-16 and Alexnet compared to VGG-16 and Resnet50. VGG-16 and Resnet50 have almost identical accuracies, with Resnet50 only slightly outperforming VGG-16 (6). This similarity would affect the explanations and could be the reason for our Grad-CAM and Squad-CAM, producing explanations that were deemed equally reliable for the two models. Therefore Grad-CAM and Squad-CAM gave the correct result because VGG-16 and Resnet50 are almost equally reliable. If we truly wanted to test whether Squad-CAM could be used to identify a more accurate model to put trust into, two models that have a more considerable difference in performance (such as AlexNet and VGG-16) should be tested.

Table 2: Trust factor comparison between VGG-16 and Resnet50.

	Backprop	Grad-CAM	Squad-CAM
VGG-16	0.52	-0.19	0.17
Resnet50	-0.52	0.19	-0.17

5.3 Localisation

The localisation capability of Grad-CAM was tested similarly to the original paper. A bounding box was generated by thresholding the Grad-CAM heat map at 15% of the max intensity, choosing the largest thresholded segment, and enclosing it in a rectangle.

A random sampling of 50 images from the ILSVRC-12 validation set (7) was used to determine the top-5 localisation accuracy according to the ILSVRC evaluation as follows. For each image, five bounding boxes are generated, one for each class in the top 5 predicted. The image contains n classes with M instances of each class. For each ground truth class, the error is 0 if there is a prediction with the correct label and a larger than 50% overlap with one of class instance bounding boxes, 1 otherwise. The errors for each class within an image are averaged.

Localisation errors are shown in table 3. Squad-CAM outperforms Grad-CAM with VGG, and results are identical with resnet. All error rates are significantly higher than the original paper's 46.41% - this difference could be explained by the different dataset (ILSVRC 2012 instead of 2015), or possibly an erroneous difference in the evaluation metric.

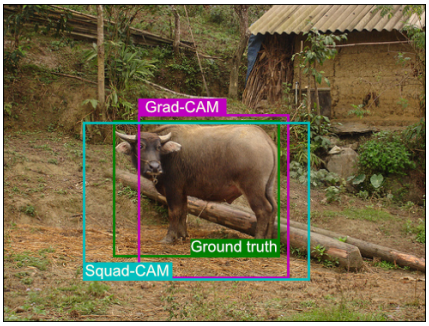
A disadvantage of using Grad-CAM for localisation is that it tends to highlight all instances of a class instead of separating. This artefact can be seen in figure 3b - here, Squad-CAM was more discriminative, but not generally across all images.

Table 3: Top-5 localisation error % according to ILSVRC 2015 for Grad-CAM and Squad-CAM.

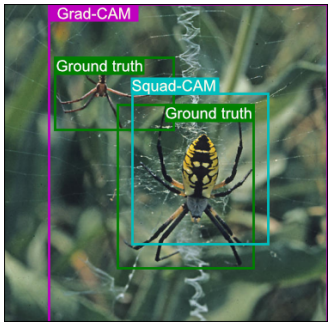
	Grad-CAM	Squad-CAM
VGG-16	70%	**58%**
Resnet50	70%	70%

5.4 Analysing failure modes

We generated Guided Grad-CAM and Guided Squad-CAM for images with incorrect top 5-predictions on ILSVRC 2012 using VGG-16, handpicked examples with unreasonable predictions, and evaluated



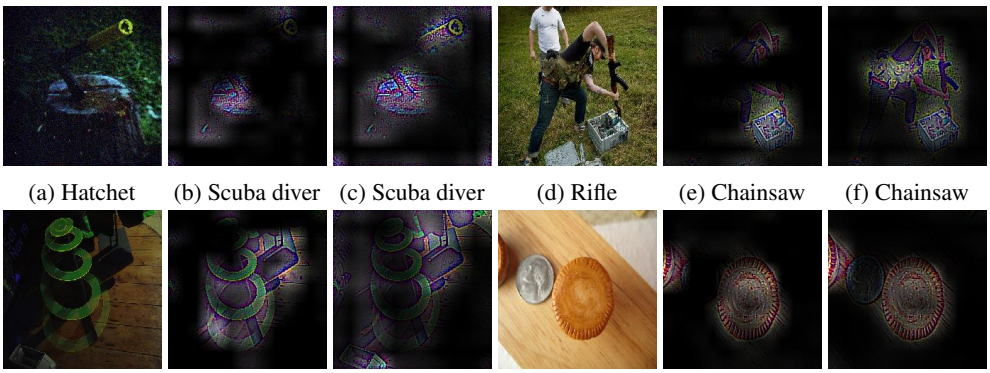
(a) Correct bounding boxes. Squad-CAM gives a larger area.



(b) An incorrect bounding box for Grad-CAM, correct for Squad-CAM. Several instances of the same object make the task difficult.

Figure 3: Grad-CAM and Squad-CAM used for localisation. Ground-truth in green, Grad-CAM in magenta, Squad-CAM in cyan.

if the methods provide reasonable explanations. In figures 4a-4f, the methods highlight informative areas (hatchet handle resembling snorkel or mouthpiece, rifle in contact with a box resembling chainsawing action). The majority of failures, however, gain little information from the methods, as shown in figures 4g-4l.



(a) Hatchet (b) Scuba diver (c) Scuba diver (d) Rifle (e) Chainsaw (f) Chainsaw

(g) Coil (h) Birdhouse (i) Birdhouse (j) Potpie (k) Acorn (l) Acorn

Figure 4: Guided Grad-CAM and Guided Squad-CAM explanations for incorrect predictions. a-f: Informative for failure analysis. (b), (e): Guided Grad-CAM. (c), (f): Guided Squad-CAM. g-l: Uninformative for failure analysis. (h), (k): Guided Grad-CAM. (i), (l): Guided Squad-CAM.

6 Challenges

The main challenge when implementing Grad-CAM was the computations of gradients. The aim was for the implementation to be pure TensorFlow. In order to compute the gradient of an intermediate layer in TensorFlow, tf.GradientTape has to be used. Implementing GradientTape proved to be quite tricky, and the solution was much more straightforward with keras.backend.gradients. This modification yielded the desired results.

When implementing the guided backpropagation, many different methods were tried. The creation of the modified activation functions was not complicated. A quick search on stack overflow gave many examples of guided backpropagation activation functions. There were two methods of modifying the activation function. Either create a customised activation function and replace the existing activation function or modify the used function inside TensorFlow. We opted for the first as it felt more intuitive compared to modifying the core functions. Looping through the model, and replacing wherever a

ReLU activation function proved to work. The implementation was not a consequence of lacking information in the paper. It was a lack of knowledge of how to implement the methods described in the paper in TensorFlow.

Another challenge we faced was implementing Alexnet. We were able to implement Alexnet, but the pre-trained weights were incompatible with our code. We were not able to train Alexnet in time for the project. Due to this issue, we instead decided to implement Resnet50 to compare the VGG-16 model. Using Resnet50 instead of Alexnet as a comparison model may explain why the results and the conclusions we obtained were slightly different from those obtained in the Grad-CAM paper.

7 Conclusion

In this work, we have proposed Squad-CAM. It is an improved version of Grad-CAM, which involves squaring the neuron importance weights, thus focusing on the value of the weights rather than their sign. We implemented Guided Backprop, Guided Grad-CAM and Guided Squad-CAM and recreated some of the experiments performed in the original paper. Our human experiments show that Guided Squad-CAM is the most class discriminative method, and is the best at explaining the decision of a model. However, we are unable to concretely determine whether Squad-CAM can expose the trustworthiness of a model because the models we used in this experiment had almost equally high accuracies. The localisation capability of our implemented Grad-CAM was significantly worse than the papers'; however, Squad-CAM seems to have a better localisation capability than Grad-CAM.

In the future, we would like to compare Squad-CAM's performance with other explanation methods. Concepts from Grad-CAM++ (8) could also be implemented to see if performance would be enhanced. We would also like to confirm Squad-CAM's ability to establish trust in a model by redoing experiment 5.2, but instead using the same models and datasets as the paper. The other experiments can also be repeated using more datasets and models to confirm their results.

8 Ethical Consideration

Using human participants in the experiments meant that ethical considerations had to be made. Informed consent was received from each participant. Participants were informed of the objective of the experiment and then gave their consent to take part. Each participant was over the age of 18; therefore, they could give their consent. The participants' data will remain confidential and will not be disclosed to anyone. Each participant was made aware of their right to withdraw, meaning they can withdraw their data at any point in the future.

9 Self Assessment

We believe that this project has been completed to a Grade A standard. The original Grad-CAM method has been successfully implemented from scratch. An improved version known as Squad-CAM has also been developed and implemented, which involves squaring the neuron importance weights. No other literature has attempted to improve on Grad-CAM, so this idea was developed entirely by us. Experiments that were carried out in the original paper have been replicated, with Squad-CAM incorporated into them. The experiments go above and beyond the requirements by developing a platform to facilitate the questionnaires. All results have been discussed, and potential reasons have been given for the results differing from the paper. Everything that was mentioned in the project proposal has been completed. The report is also well written and organised.

References

[1] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[2] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[3] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: visualising image classification models and saliency maps. corr abs/1312.6034 (2013)," *arXiv preprint arXiv:1312.6034*, 2013.

[4] A. V. A. Mahendran, "Salient deconvolutional networks," *European Conference on Computer Vision*, 2016.

[5] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[6] Keras, "Keras applications," https://keras.io/api/applications/.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[8] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 839–847.

Appendix A - Activation functions and combinations

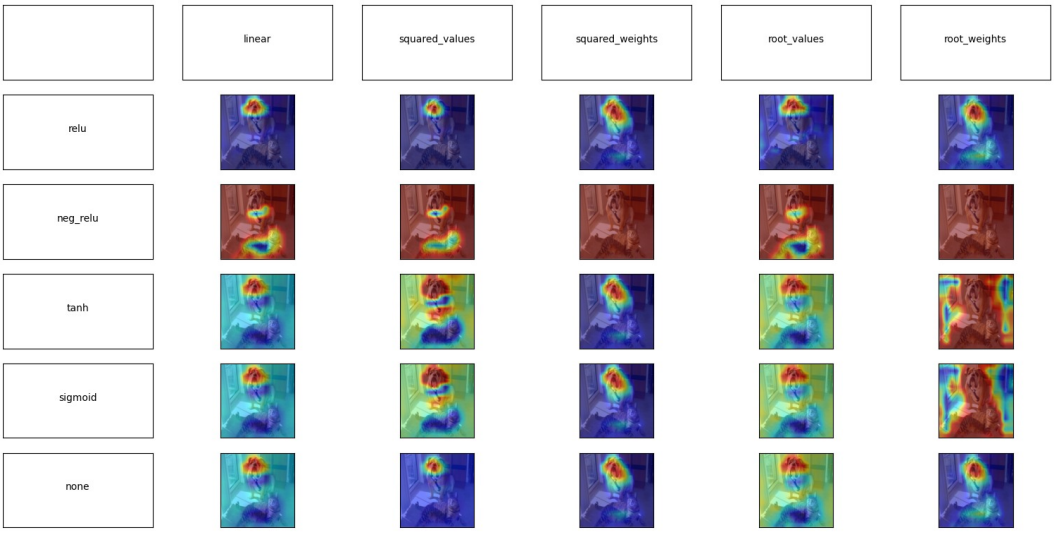


Figure 5: Different combinations and activation functions tested in order to find Squad-CAM.

Appendix B - Source Code

All code for this project can be found at `https://gits-15.sys.kth.se/andreehu/DD2412-project`.

The human studies were conducted through a hosted website specifically built for this project. This website will be available until 2020-12-31 at `squadcam.hultan.com`.