Explainable AI in Pathology - Concept Based Explainability for Mitotic Figure Detection in Whole Slide Images

Adam Tegart

Dublin City University, Dublin, Ireland

Abstract. Advances in Artificial Intelligence (AI) are allowing the performance of models to reach new heights. These advances bring added complexity to models and reduce their explainability. AI is being applied to many interesting use cases, such as in healthcare, but these complex models that are considered "black-box" in nature are difficult to interpret. This brings with them ethical implications when the autonomy and livelihood of individuals is at stake. This project aims to discover if automated approaches can be used to bootstrap concept-based explainability approaches for use in mitotic figure detection. This would allow for the use of human-interpretable concepts without the need for an expert to curate a set of concepts manually. The approach used in this paper builds on previous work that creates segments from images and clusters those that are visually similar. These clusters act as potential visual concepts, which can then be tested using statistical methods. The automated approach outlined in this paper creates a good baseline set of concepts, but requires an iterative approach to refine these concepts for complex use cases that typically require an expert.

1 Introduction

The progress made in the field of Artificial Intelligence over the past decade cannot be understated. From the release of ChatGPT [1] as an intelligent chatbot, to the increasing image fidelity in image generation tasks making use of methods such as StableDiffusion [2], there have been meaningful and exciting advances in the space of Artificial Intelligence. These advancements have come at the cost of reduced explainability as the models are becoming increasingly complex [3], [4] with an emphasis on increasing accuracy rather than creating a simple, interpretable model. Models such as these are considered black-box in nature, as the internals are obscured and the steps taken to reach a prediction are not clear. For these models to gain the trust of the general public they must offer reasoning for their choices so we can comprehend the decision process. Explainable AI (XAI) can be seen as the sine qua non for the uninterrupted advancement of AI, and this is clear from looking at the number of XAI publications in recent years [5]. There is a clear upward trend that will likely persist until the internals of models become easy to comprehend and these black-box models are converted to glass-box.

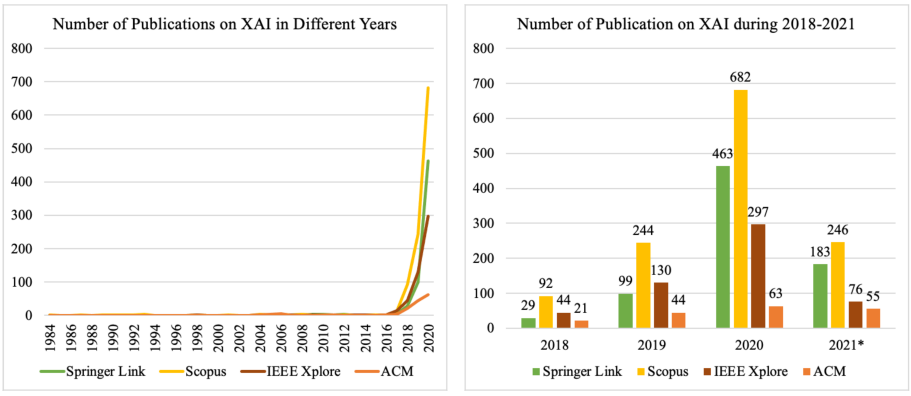

Fig. 1. Graphics showing number of XAI publications from 4 bibliographic databases [5]. The asterisk for 2021 signfies partial data up to June of 2021.

The aforementioned advances in AI models have led to a greater potential for these models to have a positive impact on society and increase our quality of life. There are however some ethical implications to consider with regard to the use of AI in regulated spaces or where the livelihoods or autonomy of individuals is at stake, such as in the health sector. These implications include transparency around decisions and who should be accountable for damages [6]. In the event that a decision from the model negatively affects the livelihood of individuals, it is not clear who is held accountable. Solving the issue of transparency in these models will allow us to better understand if there was negligence involved or if every reasonable precaution was taken. Transparency will also allow for patient satisfaction with the decision process as they are informed of the reasoning behind any diagnosis made by models. Lastly, transparency will allow these models to be integrated in a safe manner as medical professionals can disregard decisions that are made based on criteria which are not typically causal factors.

This project is concerned with applying XAI techniques to the use case of detecting mitotic figures within histopathology slides. Mitotic figures are cells which are undergoing mitosis and in large densities can be indicative of a cancerous region of tissue. A trained pathologist will make use of several visual features to determine if a cell is a mitotic figure [7]. There is a shortage of pathologists globally with only 0.8% of medical doctors being pathologists and a workload that is increasing [8]. Humans are also prone to error and there is the possibility for inter-observer variability between pathologists in relation to the counts of mitotic figures. The increased workload and possibility of varying results between pathologists motivates the use of AI, but adequate XAI techniques must allow the detections to be understood by pathologists. This will allow pathologists to ensure the model is detecting based on the correct features and will allow them to be more efficient. Transparency is a key element to the success of AI as a tool in healthcare as decisions can carry the weight of individuals' lives and every

reasonable precaution should be taken to ensure these models don't result in harm.

2 Related Works

2.1 Object Detection Architectures and Advances

Object detection has been a very well researched computer vision task over the last 20 years [9], [10]. The premise is to take a step beyond classification of objects within an image and predict the location of objects within an image along with the class. Initially, the task was approached in a way that involved a heavy focus on feature engineering with methods such as HOG [11] and DPM [12]] as compute was limited. Convolutional neural networks (CNNs) lead to the next advancements with the R-CNN [13] and SPPNet [14] being introduced. R-CNN was improved with Fast R-CNN [15] and Faster R-CNN [16], which introduced a region proposal network (RPN) to improve the regions selected for potential objects. Feature Pyramid Networks (FPN) [17] were introduced as a means of incorporating semantic information from several levels into the detection. These models mentioned are all two-stage approaches involving region proposal and detection within these regions.

One-stage detectors require no region proposals, YOLO [18] was the pioneer for one-stage detectors. The image is divided into an S x S grid and objects are detected within each cell, though the model struggles to localise smaller objects with accuracy. Improvements were made on the initial YOLO implementation, the most notable being YOLOv3 [19], YOLOv4 [20], YOLOR [21] and YOLOv7 [22]. The main contributions of these advancements are the use of both implicit and explicit knowledge in YOLOR and the use of "bag of freebies" and "bag of specials" methods in YOLO v4 and v7 which allow for additional computation during training and inference respectively for an increase in accuracy. Other popular one-stage detectors include Single Shot MultiBox Detector [23], RetinaNet [24] which introduced "focal loss", CentreNet [25] and EfficientDet [26] which introduced a bi-directional FPN.

Transformers [27] have had a huge impact in the field of natural language processing, this has led to questioning if transformers can be applied to computer vision tasks. There has been progress in this space with ViT [28], DeTR [29] and Swim Transformer [30], though these require considerable data, have more parameters and lack inductive bias.

2.2 Explainability Methods in Medical Image Analysis

Non-invasive medical imaging techniques have become more popular and this has led to an increase in medical imaging data. This increase in readily available data has spurred on the application of AI to this domain where explainability is key. We will look at the various explainability techniques that can be seen in the literature for AI in the context of medical imaging [31]–[33].

Backpropagation approaches make use of partial derivatives to determine the contributions of input to a prediction [34]–[36]. Class activation maps (CAM) [37], and its variant [38], aim to determine which values in the input led to the predictions of a certain class. Layer-wise relevance propagation (LRP) [39] is model-agnostic and the prediction is back propagated through the network until it reaches the input. Deep SHAP [40] makes use of SHapley Additive exPlanation (SHAP) values [41], these were first used in game theory to determine the contribution of each player by measuring performance with and without them included. This can be applied to CNN's by making use of DeepLIFT [42] to approximate the SHAPley values for the network's input, improving the performance.

Perturbation approaches perturb the input slightly to understand the importance of different features in the input. Local interpretable model-agnostic explanations (LIME) [43] makes use of a simple model to approximate the complex model locally. Occlusion sensitivity [35] is a perturbation technique that occludes part of the image and assesses the quality of the prediction so the pixels that have the most impact can be found. Meaningful perturbations [44] alter in a plausible fashion by using a constant value, blur or noise. It has been found that this approach is not suited for medical imaging as these natural phenomena are typically present.

Generative adversarial networks (GANs) [45] have been used as a method to produce counterfactual examples. CycleGAN [46] is an approach that takes an image X in some domain and aims to translate it into Y in the target domain, making it ideal for generating counterfactuals. Gifsplanation [47] generates counterfactuals in a similar way, the latent representation is incrementally moved in the most semantically meaningful direction for a classification and an image is generated at each increment.

Testing with Concept Activation Vectors (TCAV) [48] is an approach that measures the sensitivity of models to certain human-interpretable concepts. A concept activation vector is found by gathering examples with and without the concept and training a linear model. The normal to the hyperplane separating these 2 groups is considered the concept vector. Regression Concept Vectors [49] were introduced, which reframe the problem as a regression problem, allowing the use of continuous features. Automated Concept-based Explanation (ACE) [50] is a method to derive and cluster concepts from a set of images and determine their importance using TCAV. The benefit is that only the cluster would need to be labelled to understand the set of concepts produced, reducing the effort needed to create meaningful explanations.

3 Methodology

The research questions this project aims to answer are how well concept-based explainability methods can be applied to the use case of mitotic figure detection and to what degree an automated approach can be taken to bootstrap the concepts for a visually complex task. Automated concept generation methods

have been used previously [50], though not for visually challenging tasks that typically require an expert. Understanding the extent to which these automatically generated concepts explain the predictions will require input from experts in the field, and so it is important that they can understand the results in a meaningful way. This is the motivation behind the exploration into the visual means of displaying these concepts to understand and encapsulate their meaning in an easy to digest format.

I propose the use of the Testing with Concept Activation Vectors [48] approach, while making use of the Automated Concept Extraction [50] method to produce potential concepts with no prior commitment to develop a bespoke set of concepts. These methods cover the use of CAVs, and as a means of displaying them I propose the use of techniques similar to those used in the Gifsplanation [47] implementation. This will involve training an autoencoder to reconstruct the images and adding information from the CAVs to the activations. This approach would allow for concepts to be validated and would allow for better uptake as there is increased transparency around the predictions made by models.

The dataset that I intend to use for my project is the Mitosis Domain Generalisation challenge dataset [51], consisting of 405 whole slide images covering 6 different types of tissue between human and canine sourced from various scanners, allowing for some variation in the dataset. There are 9,501 mitotic figure annotations among 5 of the tissue types, one tissue type was withheld by the organisers to test submissions. This dataset is ideal for my use case as there are annotations provided and the task is object detection for mitotic figures. The model I intend to make use of is a Faster R-CNN [16], I decided to take this route as the performance of this model is reasonable and the implementation is less complex than more high-end models.

Preprocessing the dataset involved extracting image snippets from these slides that contained the annotations for use. This removed redundant image data from the dataset while retaining all annotations, lowering the storage requirement by over 90%. A tiling approach which allowed for overlap was used for sectioning the tiles, the motivation for this choice was to avoid annotations being cut-off if they occurred on the boundary of 2 tiles. Tiling was a necessary approach to take as the whole slide images are too large to load into memory and sectioning them up is a more efficient approach.

The Faster R-CNN model consists of a ResNet50 [52] backbone that incorporated a feature pyramid network (FPN) [17]. The motivation is that ResNet backbones are commonly used, are simple to implement and produce good results, pairing this with a FPN allows for context from multiple scales to be passed on to the region proposal network (RPN) and Fast R-CNN classifier. The pre-trained weights for the COCO dataset [53] were used and the model was trained using a 90:10 train and validation split and the model is jointly trained on 4 losses, the classification and regression loss for the region proposal network (RPN) and the classification and regression loss for the Fast R-CNN predictor. I made use of the Adam optimizer [54] with a learning rate of 0.005 and weight decay of 0.005 and a warmup for the first 1,000 iterations. A batch

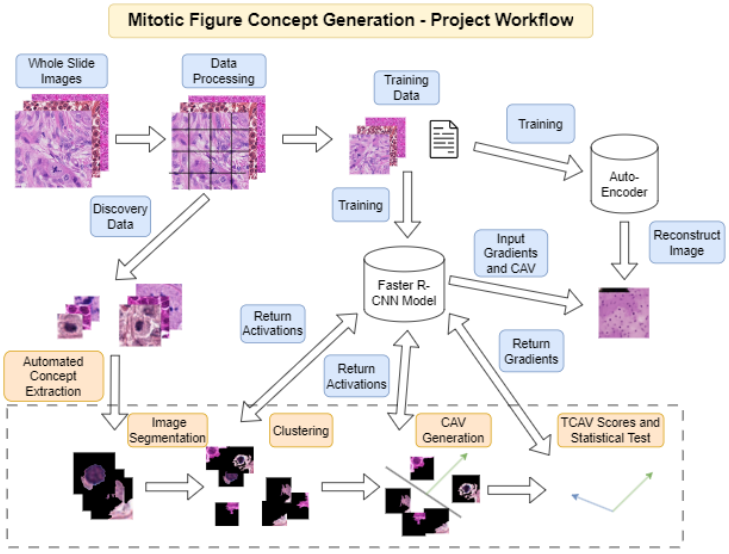

Fig. 2. A figure depicting the project workflow.

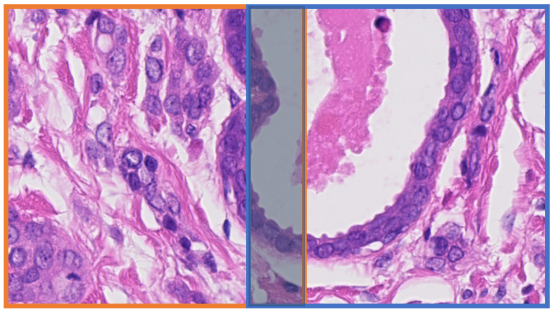

Fig. 3. A figure showing the overlap between two adjacent tiles. This avoids cutting off annotations that fall within this overlap (darker) region.

size of 2 images was used due to memory constraints. The weights were frozen for the initial convolutional layer and first bottleneck layer to match the official PyTorch implementation. There are additional actions that could be taken with respect to the learning rate, frozen weights and the dataset, though model accuracy is not the focus of this paper.

The explainability portion of this project is based on the Automated Concept Extraction [50] implementation. This process involves segmenting the images we want to derive concepts for, clustering these to find visually similar image segments, using these clustered image segments as potential concepts and testing if they have an influence on the models prediction. We will delve into the specifics of these individually.

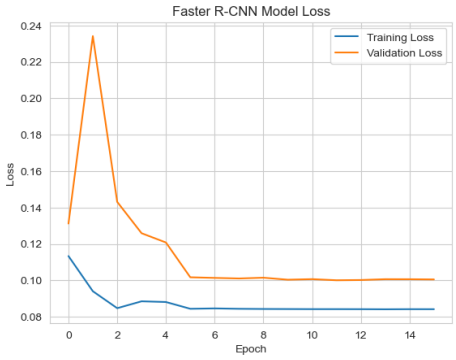

Fig. 4. Plot of the training and validation loss for the Faster R-CNN model.

MAP	0.5688
MAP 50	0.841
MAP 75	0.6119
MAR 1	0.527
MAR 10	0.7425
MAR 100	0.7434

Table 1. Faster R-CNN Metrics

3.1 Image Segmentation

Considering the visual concepts used to detect mitotic figures by experts I decided to make use of a multi-level approach to segmenting the annotations. The motivation was to capture information regarding the context the mitotic figure occurred in, such as the distance from neighbouring cells as well as the visual attributes that mitotic figures themselves display. The segmentation tools within scikit-image [55] were used for creating image patches from both the cropped annotations and annotations with context.

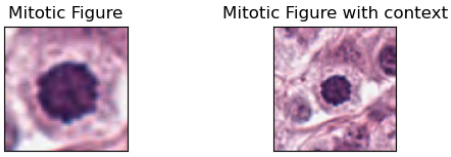

Fig. 5. A figure to show the two different types of images used for concept discovery.

3.2 Segment Clustering

These image segments were passed through the model to obtain the activations so that clustering could be performed. Incremental Principal Component Analysis **incremental˙pca** was required to ensure the activations could fit in memory. The scikit-learn [56] implementation of K-Means clustering [57] and IncrementalPCA was used. Due to compute limitations I was unable to run repeated experiments with varying numbers of clusters. Considering the number of visual features of a mitotic figure I decided to set the number of clusters to be relatively high (n=50) so that there was a possibility of a concept being repeated as opposed to missed.

3.3 Random Sample for Random experiments

In order to create a concept activation vector we need a random concept consisting of randomly sampled image segments. This is essential as we need to train a linear classifier to differentiate between the potential concept and the random concept. We can use the random samples as a way of allowing multiple CAVs to be trained, in addition to a random concept CAV to test out discovered concepts against. We can perform a two-sided t-test [58] to determine if the distribution of TCAV scores for our discovered concept is statistically different from the random concept. This will act as a means to ensure that any results we get are statistically significant and explain some visual feature in a more meaningful way than a random sample.

3.4 CAV Generation

There were several hundred images in some clusters and generating the activations for all of these across different tissues takes a long time, so I decided to sample a small portion of 50 images from each to create a concept against 50 random images. A linear model from scikit-learn [56] was trained using stochastic gradient descent [59]. The learned coefficients were extracted and used as the concept activation vector.

3.5 TCAV Scores

The TCAV scores were found by calculating the gradients for a class with respect to the bottleneck layer they were generated in for each detection. The dot product of each of these gradients with the CAV tells us what direction they are pointing relative to one another. The gradient points in the direction to decrease the class probability, so a negative value means the CAV points in the direction that causes an increase in the class probability. The fraction of all the detections in which the CAV has a positive influence is taken to be the TCAV score. Once this has been carried out for every CAV generated for a concept against the random counterparts we can take the distribution of scores and compare them against the scores from the random concept. Making use of a two-sided t-test [58] allows us to determine if the difference between the distributions is significant. I made use of 25 random samples, and so each concept has 25 CAVs generated against 25 different random samples.

3.6 Gifsplanation

In order to ensure that the CAVs were meaningful in the space an autoencoder would operate in I made use of the ResNet [52] backbone from the Faster R-CNN [16] model and froze the weights. There is a drawback with this approach, the encoder for our autoencoder cannot be fine-tuned for our specific task and will likely be less performant than if the encoder and decoder trained in tandem. The decoder weights were initialised using the Kaiming initialization [60]. The

decoder architecture [61], [62] consisted of deconvolution bottlenecks to upsample the data using transpose convolutional layers. Due to hardware limitations the decoder could only construct an output image of size 200x200, which was smaller than the original image of size 512x512. In order to account for this I scaled the original image down to the 200x200 output and used the mean squared error loss as a reconstruction loss. A stochastic gradient descent [59] optimizer was used with a learning rate of 0.5.

The motivation for the autoencoder was to create a means to visualise the effect that the CAV has in the latent space. This will allow experts to ensure that the visual concept that the CAV encapsulates aligns with current practices and is not some spurious correlation. In this way, the reasoning behind predictions can be determined as CAVs that have an influence can be displayed to determine what features the model is relying on to make predictions. The method I propose for doing this involves adding a proportion of the gradient initially to determine if the change can be localised. Continuing on from this, I propose that the CAV and scaled-gradients are multiplied together element-wise. This would allow for information in the CAV to be carried forward only if it was present within the gradients.

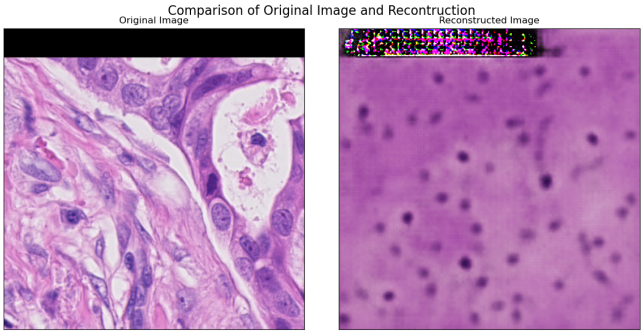

Fig. 6. Comparison between the original images and those reconstructed from the autoencoder.

4 Results and Discussion

The primary results of the project are the returned concepts and both their TCAV scores and p-values from the two-sided t-test [58]. These can be plotted to discovered trends across the various tissue types in each of the bottleneck layers and to ensure that the use of the two-sided t-test produced reasonable values. We can see from the plots for Cutaneous Mast Cell Tumour and Canine Lymphoma that the values with an average TCAV score seem to be heavily correlated with the random concept, which is shown by the high p-value assigned to these concepts. This is to say that the concepts that are considered to be mildly

influential within the model are correlated with the random concept and are considered similar. This helps to justify that the concepts that are highly influential and statistically different from the random concept are actually meaningful and are not influential just by sheer coincidence.

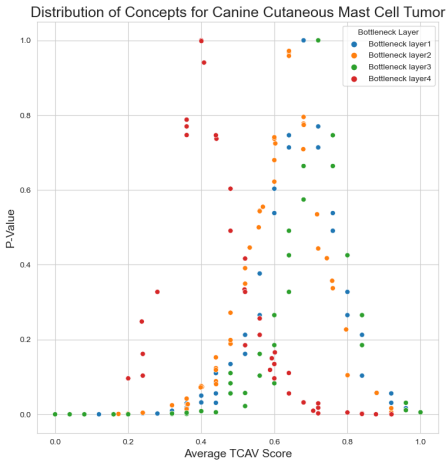

Fig. 7. Scatterplot of cutaneous mast scores and p-values.

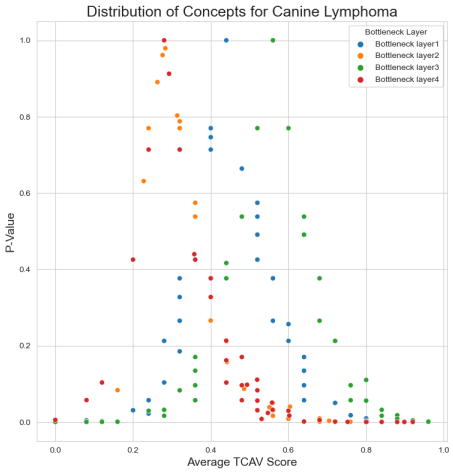

Fig. 8. Scatterplot of lymphoma scores and p-values.

We can see that the concepts within the final bottleneck layer 4 (Red markers) are scoring lower on average than the lower layers. This is due to the increased information preserved through the use of principal component analysis. In a similar fashion, they seem to be less correlated as we move to more influential concepts. This helps to signify that the approach we used has some significance in finding concepts that explain the predictions, now we can consider what these concepts may be.

The above is a sample of segments from one of the most influential potential concepts found. There seems to be no specific visual feature that I can discern from these images. There is a clear mix of segments containing the nucleus of the cell, context in the immediate vicinity around the nucleus and further off again in seemingly unrelated corners. As part of my research I composed a questionnaire consisting of samples of images from these concepts I found to be influential. This questionnaire was shared with contacts within my INTRA company, Deciphex. Unfortunately, due to time constraints, I was unable to receive results back from these questionnaires. Regardless, the preliminary analysis through visual inspection from an untrained eye can discern no notable patterns in the clusters.

The Gifsplanation approach produced some disappointing results. The reason for this was that the model did not have sufficient capacity to produce full reconstructions. The reconstructions produced were low quality resolution with

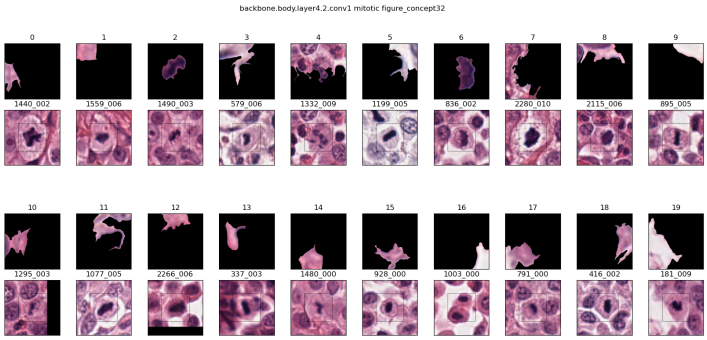

Fig. 9. Samples of patches from an influential cutaneous mast concept.

no pores or texture for the tissue. Regardless, the methods of adding gradients and CAVs to the activations resulted in some interesting images.

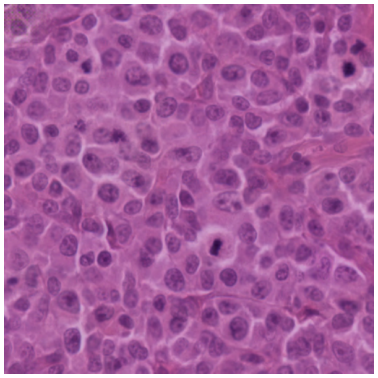

Fig. 10. Reconstructed image with a small portion of the gradient (0.02) added to activations.

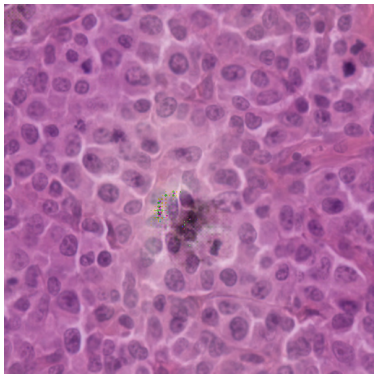

Fig. 11. Reconstrcuted images with a large portion of the gradient (0.048) added to the activations.

We can see that introducing a portion of the gradients to the activations does cause changes in the region relating to a detection, though shifted slightly to the left. I believe if the autoencoder was capable of creating high fidelity reconstructions the results would have been more promising.

We can see that implementing some of the CAV into the image does not make the results more clear. The changes in the image become less localised and seem to spuriously affect different regions in the image, though only mildly. This approach has not worked as expected, and I believe the point of failure was the autoencoder used.

Regarding the limitations of the project, they were primarily hardware based. I had issues with running both the models and automated concept generation methods as a result of limits being hit on both GPU memory and system RAM. These resulted in compromises such as settling for an adequate model and averaging across channels for the activations for the patches used to find clusters of potential concepts. The generated CAVs only consisted of a subset of the images as a result, and this hampered the results found.

There were also limitations with regard to interpreting the results, due to the specialised nature of the field. I don't have the formal training to identify mitotic figures or their causal factors. It is for this reason that I composed a questionnaire to gather the opinions of experts regarding the generated concepts, though considering the mix of image types I believe that the set of images is impure. Making use of the results would require redefining sets of concepts, meaning that the results are essentially a starting point.

5 Conclusion

Key points to take away from this project are that automated approaches for concept extraction seem to be useful as a starting point for concept-based explainability, but ultimately struggle with the intricacies of small, complex features such as those of a mitotic figure. There is some form of explainability occurring which is corroborated by the combination of the TCAV scores and p-values. There were also meaningful findings when adding the gradients to the activations passing through the network. The changes were localised, though less so with the CAVs. I believe with a better capacity model trained with consideration for the task the results would be more promising and the CAVs could become apparent. Preserving more of the variance in the activations for carrying out the clustering process would have resulted in better CAVs in the lower layers.

There were many shortcomings in the project and looking at it retrospectively I can point out flaws in my methods. There could have been more consideration regarding the model and training procedure to be used, namely that of the autoencoder. Considering the memory limitations I was unable to increase the size of the model and unable to change the encoder as the CAVs would lose meaning in that space. The solution would be to jointly train the encoder for both tasks, updating based on the loss for Faster R-CNN and then for the reconstruction loss. The lack of hyperparameters tuning and failure to stratify the training data also contributed to the less than optimal performance, though this time was put into developing the explainability aspect of this project.

If I had more time I would consider sampling images from within the influential concepts to create new CAVs. This could be done iteratively to find the images which are most influential and would allow for the CAV to be iteratively improved. There was also scope to consider the similarity of concepts between the different tissues using the cosine similarity. This would allow for insights regarding the presence of similar concepts across different tissue types.

6 Acknowledgments

I would like to thank my supervisor, Dr Alessandra Mileo, and her PhD student, Bulat Maksudov, for the helpful discussions and guidance throughout this project.

References

[1] OpenAI. "Introducing chatgpt." (Nov. 2022), [Online]. Available: https://openai.com/blog/chatgpt (visited on 04/28/2023).

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[3] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018. DOI: 10.1109/ACCESS.2018.2870052.

[4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020, ISSN: 1566-2535. DOI: https://doi.org/10.1016/j.inffus.2019.12.012. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253519308103.

[5] M. R. Islam, M. U. Ahmed, S. Barua, and S. Begum, "A systematic review of explainable artificial intelligence in terms of different application domains and tasks," *Applied Sciences*, vol. 12, no. 3, 2022, ISSN: 2076-3417. DOI: 10.3390/app12031353. [Online]. Available: https://www.mdpi.com/2076-3417/12/3/1353.

[6] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare Journal*, vol. 6, no. 2, pp. 94–98, 2019, ISSN: 2514-6645. DOI: 10.7861/futurehosp.6-2-94. eprint: https://www.rcpjournals.org/content/6/2/94.full.pdf. [Online]. Available: https://www.rcpjournals.org/content/6/2/94.

[7] A. Lashen, A. Ibrahim, A. Katayama, *et al.*, "Visual assessment of mitotic figures in breast cancer: A comparative study between light microscopy and whole slide images," *Histopathology*, vol. 79, no. 6, pp. 913–925, 2021. DOI: https://doi.org/10.1111/his.14543. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/his.14543.

[8] A. Bychkov and M. Schubert, "Constant demand, patchy supply," *The Pathologist*, Feb. 2023. [Online]. Available: https://thepathologist.com/outside-the-lab/constant-demand-patchy-supply.

[9] S. S. A. Zaidi, M. S. Ansari, A. Aslam, N. Kanwal, M. Asghar, and B. Lee, "A survey of modern deep learning based object detection models," *Digital Signal Processing*, p. 103 514, 2022.

[10] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[11] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 2005, 886–893 vol. 1. DOI: 10.1109/CVPR.2005.177.

[12] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–8. DOI: 10.1109/CVPR.2008.4587597.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2016. DOI: 10.1109/TPAMI.2015.2437384.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[15] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, 2015.

[17] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[18] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[19] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[20] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[21] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "You only learn one representation: Unified network for multiple tasks," *arXiv preprint arXiv:2105.04206*, 2021.

[22] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," *arXiv preprint arXiv:2207.02696*, 2022.

[23] W. Liu, D. Anguelov, D. Erhan, *et al.*, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, Springer, 2016, pp. 21–37.

[24] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[25] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[26] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[27] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, Springer, 2020, pp. 213–229.

[30] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[31] Z. Salahuddin, H. C. Woodruff, A. Chatterjee, and P. Lambin, "Transparency of deep neural networks for medical image analysis: A review of interpretability methods," *Computers in Biology and Medicine*, vol. 140, p. 105 111, 2022, ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2021.105111. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010482521009057.

[32] M. Pocevičiūtė, G. Eilertsen, and C. Lundström, "Survey of xai in digital pathology," *Artificial intelligence and machine learning for digital pathology: state-of-the-art and future challenges*, pp. 56–88, 2020.

[33] B. H. van der Velden, H. J. Kuijf, K. G. Gilhuijs, and M. A. Viergever, "Explainable artificial intelligence (xai) in deep learning-based medical image analysis," *Medical Image Analysis*, vol. 79, p. 102 470, 2022, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2022.102470. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522001177.

[34] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[35] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, Springer, 2014, pp. 818–833.

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

[37] 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.

[38] 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.

[39] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PLOS ONE*, vol. 10, no. 7, pp. 1–46, Jul. 2015. DOI: 10.1371/journal.pone.0130140. [Online]. Available: https://doi.org/10.1371/journal.pone.0130140.

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[41] L. S. Shapley, "17. a value for n-person games," in *Contributions to the Theory of Games (AM-28), Volume II*, H. W. Kuhn and A. W. Tucker, Eds. Princeton: Princeton University Press, 1953, pp. 307–318, ISBN: 9781400881970. DOI: doi:10.1515/9781400881970-018. [Online]. Available: https://doi.org/10.1515/9781400881970-018.

[42] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje, "Not just a black box: Learning important features through propagating activation differences," *arXiv preprint arXiv:1605.01713*, 2016.

[43] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[44] R. C. Fong and A. Vedaldi, "Interpretable explanations of black boxes by meaningful perturbation," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3429–3437.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[46] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[47] J. P. Cohen, R. Brooks, S. En, *et al.*, "Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest x-rays," in *Medical Imaging with Deep Learning*, PMLR, 2021, pp. 74–104.

[48] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, *et al.*, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*, PMLR, 2018, pp. 2668–2677.

[49] M. Graziani, V. Andrearczyk, and H. Müller, "Regression concept vectors for bidirectional explanations in histopathology," in *Understanding*

and Interpreting Machine Learning in Medical Image Computing Applications: First International Workshops, MLCN 2018, DLF 2018, and iMIMIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16-20, 2018, Proceedings 1, Springer, 2018, pp. 124–132.

[50] A. Ghorbani, J. Wexler, J. Y. Zou, and B. Kim, "Towards automatic concept-based explanations," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[51] M. Aubreville, C. Bertram, K. Breininger, S. Jabari, N. Stathonikos, and M. Veta, "Mitosis domain generalization challenge 2022," Mar. 2022. DOI: 10.5281/zenodo.6362337.

[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[53] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, Springer, 2014, pp. 740–755.

[54] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[55] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, *et al.*, "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, e453, Jun. 2014, ISSN: 2167-8359. DOI: 10.7717/peerj.453. [Online]. Available: https://doi.org/10.7717/peerj.453.

[56] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[57] S. Lloyd, "Least squares quantization in pcm," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982. DOI: 10.1109/TIT.1982.1056489.

[58] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[59] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951. DOI: 10.1214/aoms/1177729586. [Online]. Available: https://doi.org/10.1214/aoms/1177729586.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[61] J. P. Cohen, *Torchxrayvision*, https://github.com/mlmed/torchxrayvision/blob/master/torchxrayvision/autoencoders.py, 2020.

[62] A. Pasquali, *Autoencoders*, https://github.com/AlexPasqua/Autoencoders/blob/0084497b3b4ed6217006dfa386479300980444d5/src/training_utilities.py, 2021.