Sparse Explanations for Gestational Age Prediction in Fetal Brain Ultrasound

Angus Nicolson [1 2] Yarin Gal [2] Alison J. Noble [1]

Abstract

Deep learning has been shown to be a powerful tool for modelling medical imaging tasks but its black box nature reduces its potential clinical use. Recently, ProtoPNet, a prototype-based interpretable deep learning model, has been shown to perform well for natural imaging tasks and some medical imaging datasets. Here we show that it can perform well on an example ultrasound task, a challenging modality to work with, but that the explanations provided are complex. We introduce a pruning method to increase the sparsity of the model, and hence it's explanations, providing a more interpretable model. A single parameter controls the level of pruning, allowing the performance/interpretability trade-off to be specified. At high levels of pruning, the model can be made to rely solely on positive reasoning.

1. Introduction

Gestational age (GA) prediction is vital for obstetric care, with key decisions relying on an accurate estimation. However, current methods involving ultrasound measurements are known to be less reliable later in pregnancy with estimation errors of ±3 weeks in the third trimester (Papageorghiou et al., 2016). This is an important issue in low and middle income countries where women can often arrive for their first hospital visit late into pregnancy (Uganda Bureau of Statistics, 2016).

Recent work has demonstrated the use of deep learning for ultrasound-based GA prediction (Lee et al., 2020). With high risk decisions (e.g. whether to perform an emergency caesarean) relying on an accurate GA, the black box nature of deep learning reduces its potential clinical use with no way to validate the model's response (Ching et al., 2018; FDA, 2019). Providing an explanation allows machine learning practitioners to debug the model and clinicians to trust the output if the explanation agrees with their own assessment. Recent work on interpretable models aims to alter the structure of deep learning models to provide explanations with their predictions (Rudin, 2018). These can be local (for a specific image/prediction) or global (average response of a model) (Doshi-Velez & Kim, 2017; Lipton, 2016). These approaches are desirable because we *design* the model to use these explanations rather than using a post-hoc method, like saliency, where the explanations may not be faithful (Adebayo et al., 2018).

Prototypical Part Network (ProtoPNet) (Chen et al., 2019), the model on which this work is based, classifies a test image by calculating its similarity to a set of sub-parts from within the training dataset and then weighting those similarities. This provides an explanation that is similar to how a clinician might make a prediction, e.g. "this fetus is 30 weeks because it looks like a 30 week fetus I have seen before". The model is globally interpretable because the prototypes and weights are fully accessible. It's local explanations consist of the prototypes most similar to a test image and their corresponding contributions to the output.

Although ProtoPNet encourages sparsity through regularising the fully-connected layer, this is not sufficient to arrive at simple explanations, with many redundant explanations still present. We propose enforcing a sparse model by pruning weights below some threshold, τ. This reduces model complexity, and therefore the explanations, by ensuring that less prototypes are relevant to each class and for each prediction.

The main contributions of our work are:

- We propose a novel method of increasing the sparsity of ProtoPNet and its explanations by pruning low weights from the fully-connected layer;
- We present the first interpretable deep neural network for ultrasound image analysis beyond simply applying post-hoc saliency methods;
- We demonstrate that ProtoPNet can perform comparably to a black box network for 2D ultrasound imaging.

2. Related Work

In previous work on ultrasound imaging, attention mechanisms or post-hoc saliency are the interpretability methods

[1] Institute of Biomedical Engineering, University of Oxford [2] OATML, Department of Computer Science, University of Oxford. Correspondence to: Angus Nicolson <angus.nicolson@eng.ox.ac.uk>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML). Copyright 2022 by the author(s).

utilised (Schlemper et al., 2018; Zhang et al., 2020; Qian et al., 2021; Salahuddin et al., 2022). By applying and then extending ProtoPNet we demonstrate the first interpretable-by-design deep neural network for ultrasound which goes beyond simply visualising important regions of the input image.

Multiple different changes to ProtoPNet have been suggested (Nauta et al., 2020; Rymarczyk et al., 2021b; Kim et al., 2021; Wang et al., 2021; Barnett et al., 2021; Rymarczyk et al., 2021a). The most relevant models to our work are ProtoPNet (Chen et al., 2019), ProtoTree (Nauta et al., 2020), ProtoPShare (Rymarczyk et al., 2021b) and ProtoPool (Rymarczyk et al., 2021a) as they also perform pruning, although each method only removes whole prototypes, as opposed to individual weights as in our work. ProtoTree uses a pruning method suitable only for decision trees. ProtoPNet removes prototypes which are similar to multiple classes, and therefore represent background patches, an assumption which does not hold for our ultrasound dataset. ProtoPool and ProtoPShare remove similar prototypes, reducing model redundancy and finding similarities between classes by sharing prototypes across classes.

3. Datasets

The ultrasound imaging is a subset of the Fetal Growth Longitudinal Study (FGLS) of the International Fetal and Newborn Growth Consortium for the 21st Century Project (INTERGROWTH-21st) (Papageorghiou et al., 2014; 2018). The gestational ages were binned in two week intervals to convert the task from a regression to a classification. There are 106,505 images from 3733 women. Each image is of the transthalamic plane. The dataset was randomly split patient-wise in a 80:10:10 ratio for the train, validation and test splits giving 85,033, 10,803 and 10,669 images, respectively. See Appendix A for more details.

We use CIFAR-10 (Krizhevsky & Hinton, 2009) as a proof-of-concept dataset to demonstrate that our pruning method is suitable across both natural and medical imaging datasets.

For both datasets, training images were resized to 244×244 and then heavily augmented to reduce overfitting by random application of the following operations: horizontal flip, skew, shear, elastic distortion and then application of a rotation, brightness jitter, contrast jitter, and additive Gaussian noise. Finally, a random crop and resizing to 224×224 was applied and, for INTERGROWTH-21st, the number of channels duplicated to convert the greyscale image to RGB.

4. Methods

4.1. ProtoPNet

A ProtoPNet (Chen et al., 2019) model consists of a convolutional network, f, a prototype layer, g_p, and a fully-connected layer, h. In our experiments the convolutional network is a ResNet-18 pretrained on ImageNet (He et al., 2016) followed by two 1×1 convolutional layers to reduce the number of output channels to 128.

The model makes a prediction by passing some input image, x, through the convolutional feature extractor to obtain a set of feature maps, $f(x)$, with shape $H \times W \times D$. The network learns m prototypes, $\boldsymbol{P} = \{\boldsymbol{p}_j\}_{j=1}^{m}$, which are vectors of shape $H_1 \times W_1 \times D$, where $H_1 < H$ and $W_1 < W$. In our experiments the feature maps are of shape $7 \times 7 \times 128$ and the prototypes $1 \times 1 \times 128$, so each prototype is a representation of some prototypical sub-patch of the image $\frac{1}{49}^{\text{th}}$ of its size. The prototype layer then calculates the L2 distance between each prototype and all 49 patches of the feature map. These distances are inverted to obtain a set of similarity scores and the maximum score for each prototype is then passed through the fully-connected layer. The maximal similarity scores can be seen as the likelihood that each prototype is present in the image and the weights in the fully-connected layer the importance of each prototype for each class.

To ensure each class will be represented, each prototype is pre-allocated a class. For ProtoPNet (Chen et al., 2019), 10 prototypes were used per class. In this paper we use only 5, as preliminary work showed equivalent performance. We denote $\boldsymbol{P}_k \subseteq \boldsymbol{P}$ as the subset of prototypes belonging to class $k \in \{1, \ldots, K\}$, where K is the number of classes.

The training protocol has three steps: (1) the prototypes and convolutional layers are jointly optimised using stochastic gradient decent (SGD); (2) the prototypes are pushed to the activations of the nearest (measured in feature space) image sub-patch of the train dataset - this provides the model its interpretability, as the prototypes can now be represented in image space by that sub-patch; (3) the rest of the network is frozen and the fully-connected layer is optimised using SGD. These three steps are repeated multiple times until the model converges.

The loss optimised during step (1) is composed of three parts: cross entropy, cluster loss and separation loss. Cluster loss encourages the model to have at least one training sub-patch with activations close to each prototype. This minimises the changes made to the prototypes in step (2) of training. Separation loss encourages the prototypes of different classes to be far apart in feature space. For a detailed explanation of the losses, architecture and training protocol see the ProtoPNet paper (Chen et al., 2019).

Let each individual weight within the fully-connected layer be denoted $w_{k,j}$, where k and j are the class and prototype indices, respectively. To initialise the fully-connected layer, for class k, we set $w_{k,j} = 1$ for all j with $\boldsymbol{p}_j \in \boldsymbol{P}_k$ and $w_{k,j} = -0.5$ for all j with $\boldsymbol{p}_j \notin \boldsymbol{P}_k$. This encourages the model to learn prototypes characteristic of class k and not of other classes. In step (3) the loss for h is cross entropy and a masked L1 loss. The mask is applied to regularise only weights from prototypes not of the same class as the output logit, i.e. the weights $w_{k,j}$ for all j with $\boldsymbol{p}_j \notin \boldsymbol{P}_k$.

4.2. ProtoPNet-CA

For INTERGROWTH-21st, the restriction that each prototype be relevant to a single class was removed. We name this model ProtoPNet Class Agnostic (ProtoPNet-CA). This is because the task is a regression task converted into a classification by binning, making the classes more similar to each other than in a standard classification task. Thus, it could be desirable for a prototype to be similar to multiple classes. We achieve this by removing the separation loss; the mask applied to the L1 regularisation and the cluster loss; and the class-dependent initialisation of the final layer. ProtoPNet (Chen et al., 2019) masked the L1 loss to reduce the level of negative reasoning present in the model, the argument being it is easier to interpret the model if solely positive reasoning is used. Therefore, to increase the relative levels of positive reasoning in ProtoPNet-CA, we initialised the final layer to a uniform distribution between 0 and 1.

4.3. Pruning

We introduce a novel method of pruning by simply setting each $w_{k,j}$ below some threshold, τ to zero, namely:

$$w'_{k,j} = \begin{cases} w_{k,j} & \text{if } w_{k,j} \geqslant \tau, \\ 0 & \text{if } w_{k,j} < \tau. \end{cases} \tag{1}$$

If all connections to a prototype are zero, the prototype is removed from the network:

$$\boldsymbol{P'} = \boldsymbol{P} - \{\boldsymbol{p}_j : j \in \{j : w_{k,j} = 0 \,\forall\, k\}\} \tag{2}$$

where $\boldsymbol{P'} \subseteq \boldsymbol{P}$ is the set of prototypes after pruning. The fully-connected layer is then trained for 15 epochs, keeping any weights at zero fixed to allow the model to adapt to the changes, while keeping the same level of sparsity.

4.4. Interpretability Metrics

We define the relevant prototypes for each class, $\boldsymbol{\mathcal{P}}_k$, as the set of prototypes whose weight connections are non-zero for that class:

$$\boldsymbol{\mathcal{P}}_k = \{\boldsymbol{p}_j : j \in \{j : w_{k,j} \neq 0\}\} \tag{3}$$

We propose a novel measure of model complexity specific to the ProtoPNet model: the mean number of relevant prototypes per class, r:

$$r = \frac{1}{K}\sum_{k=1}^{K} |\boldsymbol{\mathcal{P}}_k| \tag{4}$$

This is equivalent to the number of non-zero weights in the fully-connected layer divided by the number of classes. r gives an intuitive measure of the size of the model's global explanations as it is the average number of prototypes which affect each logit output. We also define r^+ and r^-, which are the subset of prototypes which have positive or negative weight connections, respectively. Similarly, we report the L1 loss on the subset of weights that are positive, L1$^+$, and negative, L1$^-$. These metrics allow us to track the relative level of positive/negative reasoning in the model.

5. Experiments

Our pruning method was applied to both ProtoPNet and ProtoPNet-CA models with τ ranging from $0.05 - 0.5$ in steps of 0.05. Table 1 summarizes the performance of a selection of these models and their blackbox counterparts for both CIFAR-10 and INTERGROWTH-21st. ProtoPNet-CA models were not trained for CIFAR-10 as we do not have the same justification to remove the class dependency as for INTERGROWTH-21st. The optimal τ is highly dependent on the unpruned trained model, as such, comparisons between models pruned to the same τ threshold are less relevant than comparisons at the same complexity (r or $L1$) or accuracy. See Appendix B for more extensive results.

CIFAR-10 Table 1 shows that we can prune to a high degree without losing classification accuracy. This high level of interpretability comes at a cost compared to the black box ResNet, with accuracies decreasing between $2.2-2.7\%$, but not compared to an unpruned ProtoPNet where a model with $\tau = 0.3$ has the same accuracy as the original model but almost half the L1 loss, 8% the r at just 3.9, and $34/50$ prototypes remaining. Our method has the ability to tune this trade-off with further increases in τ for a sparser, but less well performing model. For example, a τ of 0.4 gives an r of just 2.7 and an r^- of 0, indicating there is no negative reasoning in the model, with this additional interpretability coming at a cost of only 0.3% accuracy compared to the unpruned model.

INTERGROWTH-21st Mean average error (MAE) in days is reported by setting the centroid of each classification bin as the model's prediction. The lowest MAE achieved was by a ProtoPNet model (6.18) and the highest accuracy by ProtoPNet-CA (68.9), both improvements from their black box counterpart. As with CIFAR-10, pruning increases the sparsity of ProtoPNet for INTERGROWTH-21st while only suffering performance loss at higher τ values. For example, At $\tau = 0.1$, ProtoPNet shows no increase in

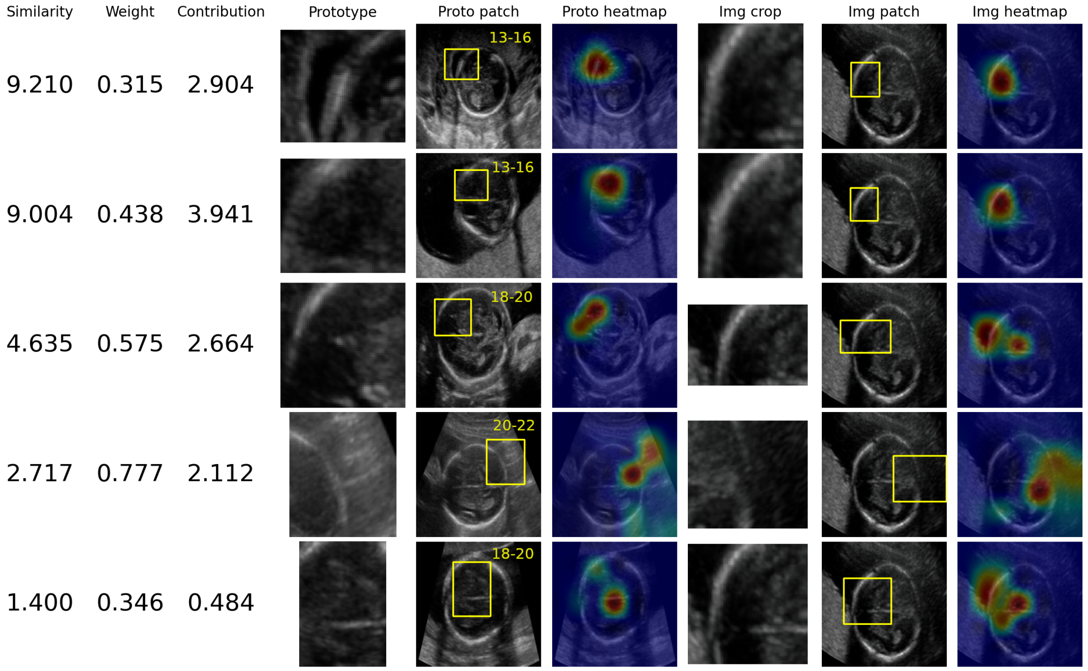

Figure 1. Local explanation of ProtoPNet-CA ($\tau = 0.25$) for the successful prediction of an 18-20 test image. All prototypes with a non-zero weight to the predicted class k are shown. Columns 1-3 are the similarity of the test image with the prototype, it's weight connection with class k and the product of the two. Columns 4-6 show the cropped training image representing the prototype, the bounding box containing 95% of the similarity and the similarity heatmap. Columns 7-9 contain the same information but for the test image. The prototypes' class is labelled in yellow in column 5. Note that the prototypes important for the prediction of class 18-20 are not restricted to solely that class.

MAE but r is reduced to a third of its value, whereas at $\tau = 0.20$ the MAE increases by 0.30 and r drops to 6.7 – 10% it's original value.

If interpretability is the primary aim, ProtPNet-CA can produce highly sparse models which rely almost solely on positive reasoning, although at a cost in MAE between $0.09 - 0.26$ compared to ProtoPNet. For a model with $\tau = 0.25$, an r^{-} of 0.2 and an L1^{-} of 1.5 were achieved compared to ProtoPNet's[1] values of 3.2 and 27.0, respectively. The ProtoPNet model actually has a higher level of negative reasoning than positive using L1 as a metric. The local explanation for this ProtoPNet-CA model for a fetus in the 18-20 weeks gestational age class is shown in Fig. 1. The explanation contains only 5 prototypes and all weights are positive, benefiting the ProtoPNet's *this looks like that* explanation style (Chen et al., 2019). Having fewer relevant prototypes assists human ability to understand the model. Although an unpruned ProtoPNet model works in the same manner, it is beyond human capacity to understand the relative importance of hundreds of different prototypes and their respective weights.

6. Conclusion

We presented an easy to implement solution to increase the sparsity of a ProtoPNet model by fixing low weight connections to zero, removing unused prototypes and retraining the fully-connected layer.

We demonstrated that our method produces simple explanations for a natural imaging and an ultrasound medical imaging task with only a small number of relevant prototypes for each class. This simplicity can come at a cost to accuracy at high levels of pruning, but this cost is small and can be fine-tuned by changing a single hyperparameter. Additionally, we introduced a modification to ProtoPNet (ProtoPNet-CA) that, after pruning, was shown to produce simpler models that relied almost solely on positive reasoning.

Future work could explore more complex forms of pruning where unimportant weights, rather than low weights, are pruned, or where our pruning method is combined with previous methods (Section 2). Additionally, the level of interpretability of the models should be measured using human experiments as recommended in (Doshi-Velez & Kim, 2017), particularly the assertion that models involving less negative reasoning are easier to interpret, and whether the prototypes have a semantic meaning.

[1] τ of 0.20 is used here for comparison because it is the largest value for which all classes have at least one weight connection.

Table 1. Classification metrics for ProtoPNet and Class Agnostic ProtoPNet (ProtoPNet-CA) models pruned at different thresholds (τ) and plain black box ResNets as comparisons. Models with a $*$ have been pruned. m is the number of prototypes remaining after pruning. r is the mean number of relevant prototypes per class. L1 and r are not shown for the plain ResNets as they are not measures of interpretability for black box models. Intergrowth MAE is in days. The best values for each dataset are shown in bold.

DATASET	MODEL	τ	$m\downarrow$	ACC $\uparrow$	MAE $\downarrow$	L1 $\downarrow$	L1$^+\downarrow$	L1$^-\downarrow$	$r\downarrow$	$r^+\downarrow$	$r^-\downarrow$
INTERGROWTH	RESNET-18	-	-	68.4	6.27	-	-	-	-	-	-
	PROTOPNET	-	65	68.8	**6.18**	63.4	32.9	30.5	65.0	31.8	33.2
	PROTOPNET*	0.10	63	68.8	**6.18**	57.1	26.9	30.2	19.0	10.5	8.5
		0.20	39	66.8	6.58	45.0	**17.9**	27.0	**6.7**	**3.5**	3.2
	PROTOPNET-CA	-	65	**68.9**	6.27	80.1	67.3	12.8	65.0	35.2	29.8
	PROTOPNET-CA*	0.20	46	68.2	6.33	53.4	48.7	4.7	12.1	11.4	0.7
		0.25	**34**	65.9	6.84	**40.0**	38.5	**1.5**	7.5	7.2	**0.2**
CIFAR-10	RESNET-18	-	-	**95.1**	-	-	-	-	-	-	-
	PROTOPNET	-	50	92.7	-	47.2	29.7	17.5	50.0	21.2	28.8
	PROTOPNET*	0.10	50	92.9	-	41.6	28.8	12.8	14.4	6.3	8.1
		0.30	34	92.7	-	27.0	26.5	0.5	3.9	3.8	0.1
		0.40	**25**	92.4	-	**22.9**	**22.9**	**0.0**	**2.7**	**2.7**	**0.0**

Acknowledgements

We would like to thank Lok Hin Lee and all anonymous reviewers for sharing their valuable feedback and insights. Angus Nicolson is supported by the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1).

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity Checks for Saliency Maps. *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., and Rudin, C. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*, 3 (12):1061–1070, 2021.

Chen, C., Li, O., Tao, C., Barnett, A. J., Su, J., and Rudin, C. This looks like that: Deep learning for interpretable image recognition. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P. M., Zietz, M., et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15(141), 2018.

Doshi-Velez, F. and Kim, B. Towards A Rigorous Science of Interpretable Machine Learning. 2017.

FDA. Clinical Decision Support Software Draft Guidance for Industry and Food and Drug Administration Staff. Technical report, 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Kim, E., Kim, S., Seo, M., and Yoon, S. XProtoNet: Diagnosis in Chest Radiography with Global and Local Explanations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 15714–15723, 2021.

Krizhevsky, A. and Hinton, G. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

Lee, L. H., Bradburn, E., Papageorghiou, A. T., and Noble, J. A. Calibrated Bayesian Neural Networks to Estimate Gestational Age and Its Uncertainty on Fetal Brain Ultrasound Images. In *Lecture Notes in Computer Science*, volume 12437 LNCS, pp. 13–22. Springer, 2020.

Lipton, Z. C. The Mythos of Model Interpretability. *Communications of the ACM*, 61(10):35–43, 2016.

Nauta, M., van Bree, R., and Seifert, C. Neural Prototype Trees for Interpretable Fine-grained Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 14928–14938, 2020.

Papageorghiou, A. T., Ohuma, E. O., Altman, D. G., Todros, T., Ismail, L. C., Lambert, A., Jaffer, Y. A., Bertino, E., Gravett, M. G., Purwar, M., et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet*, 384(9946): 869–879, 2014.

Papageorghiou, A. T., Kemp, B., Stones, W., Ohuma, E. O., Kennedy, S. H., Purwar, M., Salomon, L. J., Altman, D. G., Noble, J. A., Bertino, E., et al. Ultrasound-based gestational-age estimation in late pregnancy. *Ultrasound in Obstetrics and Gynecology*, 48(6):719–726, 2016.

Papageorghiou, A. T., Kennedy, S. H., Salomon, L. J., Altman, D. G., Ohuma, E. O., Stones, W., Gravett, M. G., Barros, F. C., Victora, C., et al. The INTERGROWTH-21 st fetal growth standards: toward the global integration of pregnancy and pediatric care. *American Journal of Obstetrics and Gynecology*, 218(2):S630–S640, 2018.

Qian, X., Pei, J., Zheng, H., Xie, X., Yan, L., Zhang, H., Han, C., Gao, X., Zhang, H., et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nature Biomedical Engineering*, 5:522–532, 2021. doi: 10.1038/s41551-021-00711-2.

Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence*, 1(5):206–215, 2018.

Rymarczyk, D., Struski, Ł., Górszczak, M., Lewandowska, K., Tabor, J., and Zieliński, B. Interpretable Image Classification with Differentiable Prototypes Assignment. 2021a.

Rymarczyk, D., Struski, Ł., Tabor, J., and Zieliński, B. ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, KDD '21, pp. 1420–1430, 2021b.

Salahuddin, Z., Woodruff, H. C., Chatterjee, A., and Lambin, P. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in Biology and Medicine*, 140:105111, 2022.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., and Rueckert, D. Attention Gated Networks: Learning to Leverage Salient Regions in Medical Images. *Medical Image Analysis*, 53:197–207, 2018.

Uganda Bureau of Statistics. Uganda Demographic and Health Survey 2016. pp. 625, 2016.

Villar, J., Altman, D. G., Purwar, M., Noble, J. A., Knight, H. E., Ruyan, P., Cheikh Ismail, L., Barros, F. C., Lambert, A., Papageorghiou, A. T., et al. The objectives, design and implementation of the INTERGROWTH-21st Project. *BJOG: An International Journal of Obstetrics Gynaecology*, 120(SUPPL. 2):9–26, 2013.

Wang, J., Liu, H., Wang, X., and Jing, L. Interpretable Image Recognition by Constructing Transparent Embedding Space. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 875–884, 2021.

Zhang, J., Petitjean, C., Yger, F., and Ainouz, S. Explainability for Regression CNN in Fetal Head Circumference Estimation from Ultrasound Images. In *Lecture Notes in Computer Science*, volume 12446 LNCS, pp. 73–82. Springer, 2020.

A. INTERGROWTH-21st dataset characteristics

A healthy cohort of women from 8 different countries (Brazil, China, India, Italy, Kenya, Oman, UK, USA) who had known gestational ages via agreement between biometry measurements at first scan and last known menstrual period (Villar et al., 2013). Each scan was done on the same model of scanner and using the same protocol. Multiple images per participant were obtained, with 4-5 visits at different gestational ages and a mean of 6 images per visit. Repeat images at a single time-point often differ only slightly in appearance.

In Figure 2 the gestational age distribution for our subset of the INTERGROWTH-21st dataset is shown, binned into the same classes as used for our experiments. The distribution is approximately uniform across 13 to 42 weeks with slightly less images at the extremes. The outermost classes were increased in size (13-16 and 38-42) to help with class imbalance.

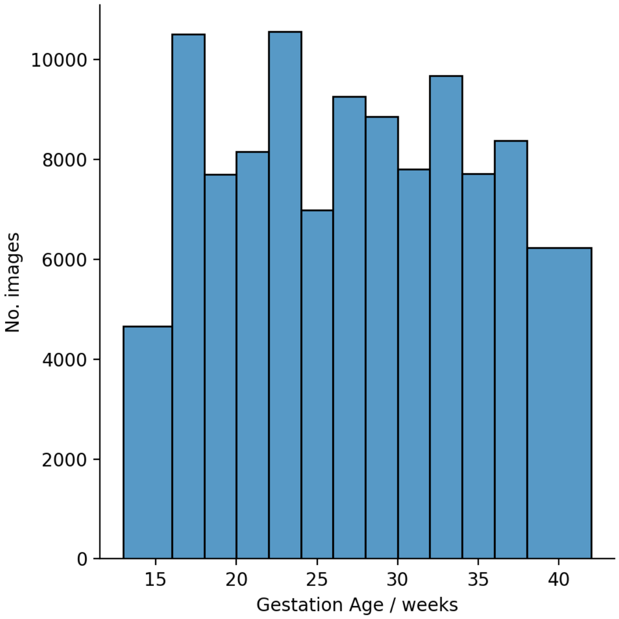

Figure 2. Age distribution for our subset of the INTERGROWTH-21st dataset binned as in the classification task

B. Effect of the tuning parameter

Figures 3 and 4 show the trend in interpretability metrics and performance of the ProtoPNet model for CIFAR-10 and the ProtoPNet-CA model for INTERGROWTH-21st as the prune threshold τ is increased. Although experiments were performed with τ reaching 0.5, each figure is cutoff earlier as the models would collapse when all the weights to a class were removed.

For CIFAR-10, Figure 3 shows that pruning successfully reduces the complexity of the model while retaining (and even slightly improving) its accuracy. Most notably, we can see a pruning threshold of 0.3 has an almost identical accuracy as the original model but almost half the L1 loss, 7% r. Note the changeover at $\tau = 0.2$, where the levels of positive reasoning within the model (r^+) overtake the negative reasoning (r^-).

For INTERGROWTH-21st, similar trends are observed in Figure 4, although the performance of the model seems to deteriorate more quickly as τ is increased compared to for CIFAR-10, and r^+ is always higher than r^-.

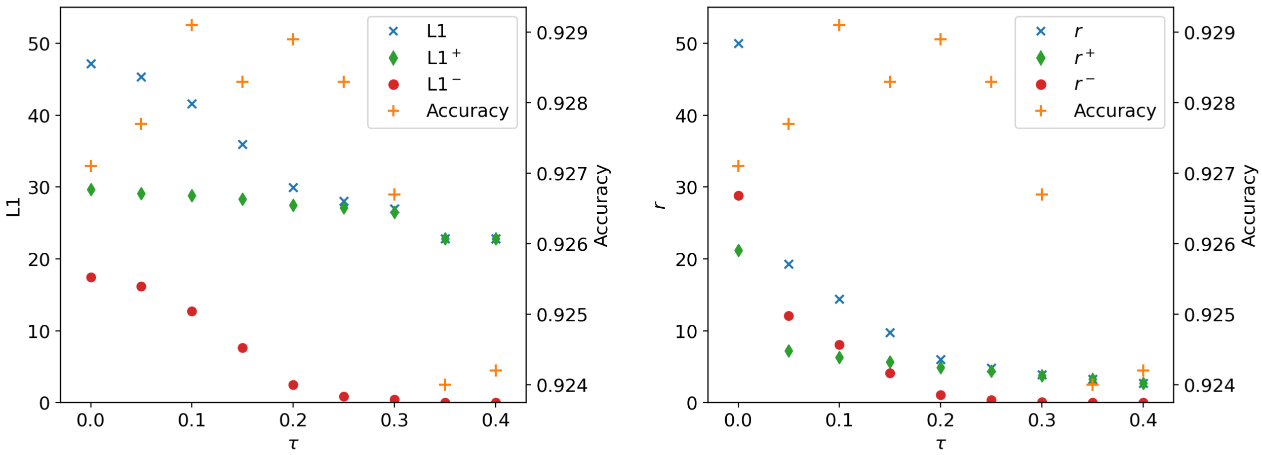

Figure 3. ProtoPNet results for CIFAR-10 as the pruning threshold τ is increased. Fully connected L1 loss (left) and mean number of relevant prototypes r (right) for all (blue cross), positive (green diamond) and negative (red circle) weights against pruning weight threshold τ. The model accuracies (orange plus) are shown on the right axis of each plot

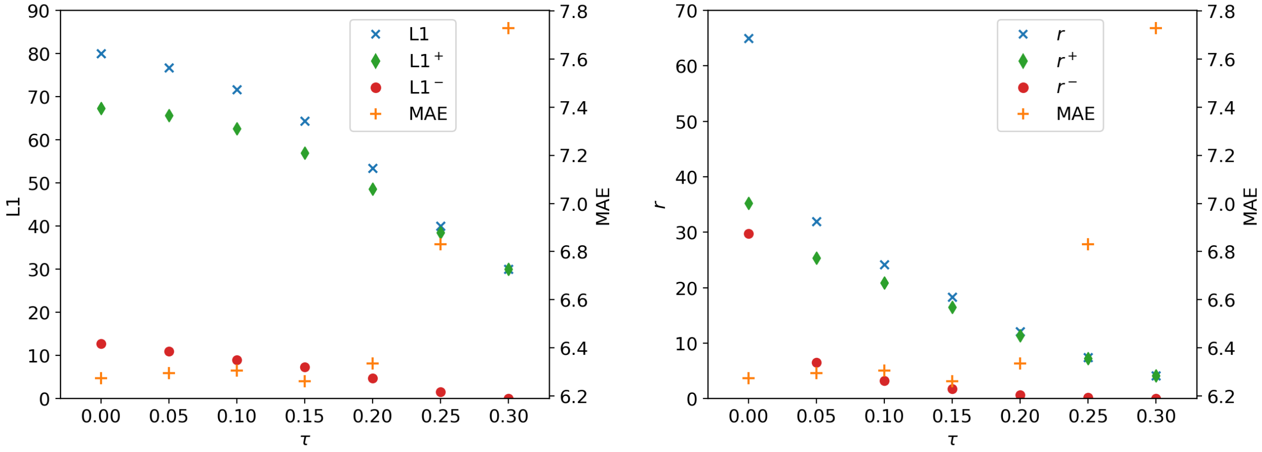

Figure 4. ProtoPNet-CA results for INTERGROWTH-21st as the pruning threshold τ is increased. Fully connected L1 loss (left) and mean number of relevant prototypes r (right) for all (blue cross), positive (green diamond) and negative (red circle) weights against pruning weight threshold τ. The model MAEs (orange plus) are shown on the right axis of each plot