CLINICALLY VALIDATING EXPLANATIONS FOR FETAL BRAIN ULTRASOUND

Angus Nicolson, Elizabeth Bradburn, Yarin Gal, Alison Noble - University of Oxford
angus.nicolson@eng.ox.ac.uk angus.nicolson.com

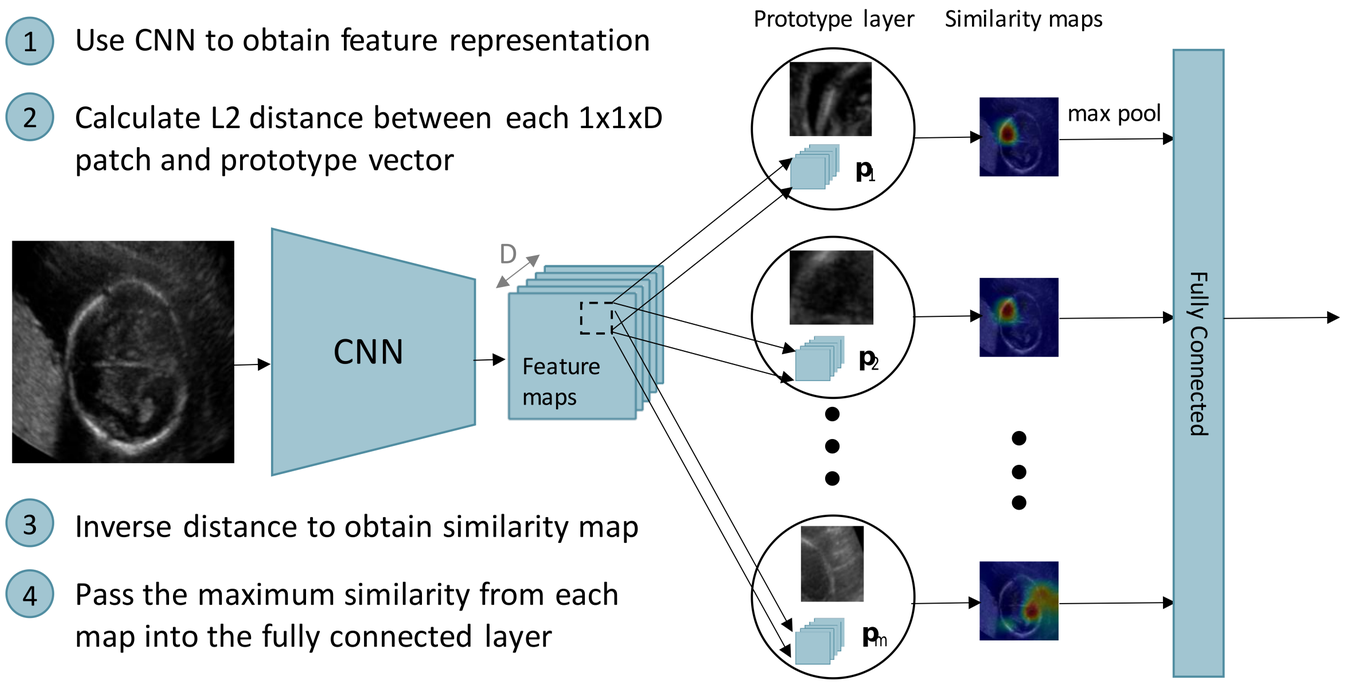
Abstract

Deep learning has been shown to be a powerful tool for modelling medical imaging tasks, but its black box nature reduces its potential clinical acceptability. Recently, ProtoPNet, a prototype-based interpretable deep learning model, has been shown to perform well at the task of gestational age estimation from fetal brain ultrasound. In the current work, we demonstrate the utility of that model through a preliminary reader study with one clinician, where, for a sample of 64 images across a 13-42 week age range, the clinician mean average error (MAE) was reduced from 23.3 days to 9.2 days when a clinician had access to model predictions. The clinician MAE reduced by a further 0.8 days when model explanations were displayed. In future work, we aim to repeat the study with multiple clinicians to gain further insight into how the explanations influence human decision-making and trust in the model.

This Looks Like That

Prototypical Part Network (**ProtoPNet**) [3] 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".

1. Use CNN to obtain feature representation
2. Calculate L2 distance between each 1x1xD patch and prototype vector
3. Inverse distance to obtain similarity map
4. Pass the maximum similarity from each map into the fully connected layer



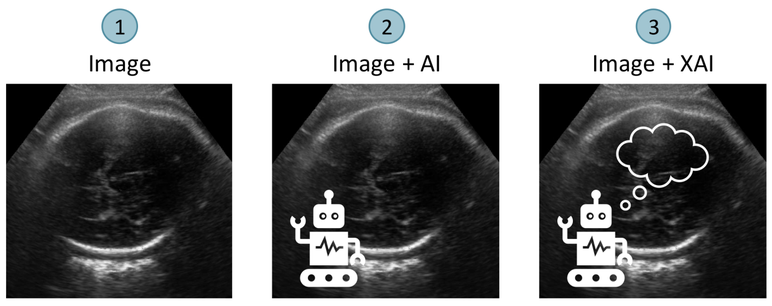
Gestational Age

We use data from the Fetal Growth Longitudinal Study (FGLS) of the International Fetal and Newborn Growth Consortium for the 21st Century Project (**INTERGROWTH-21st**) [1]. There are $106,505$ 2D fetal brain ultrasound images from 3733 women. We binned the ages in two week intervals to convert the task from a regression to a classification. Only 64 images are used for our test set as clinician time is difficult to obtain.

Study Design

Recently, we have shown that interpretable models can be used to estimate gestational age from fetal brain ultrasound [2] but how do these models affect clinician behaviour in practice? And do the explanations improve performance?

Our study is split into three phases where the clinician is asked to estimate gestational age with successively more information:



In phase 1, the clinician is also asked which features contributed to their estimate, allowing us to determine the most likely sources of information the model could be using.

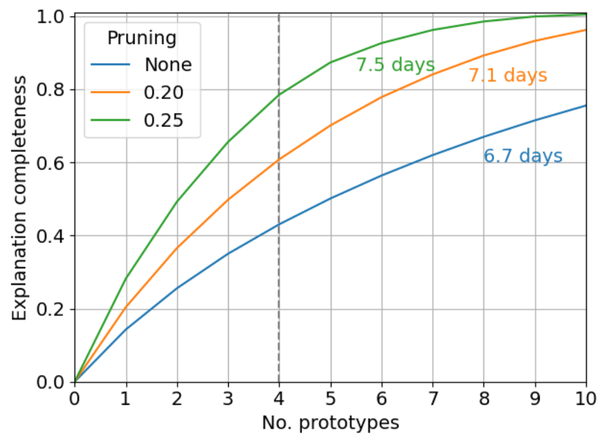
We aim to have 30 trained sonographers perform the study and are currently applying for ethics approval. This poster contains preliminary results from a single clinician.

Pruning

We simplify the ProtoPNet model, and it's explanations, through **pruning**. This is done by simply setting each weight in the fully-connected layer, $w_{k,j}$, below some threshold, τ to zero, namely:

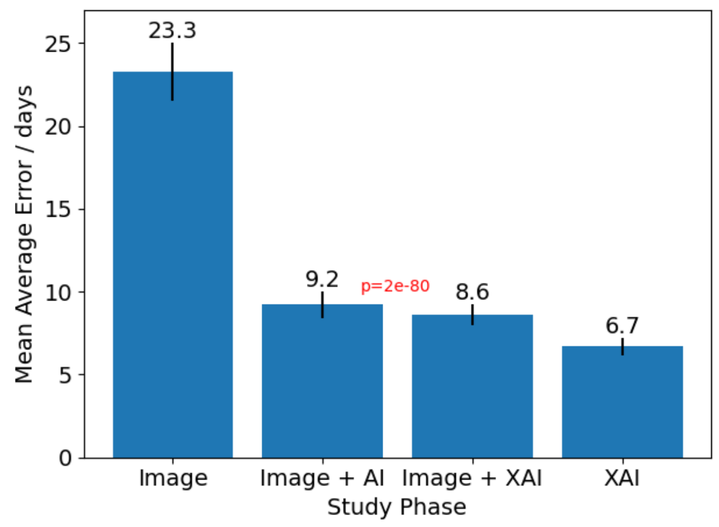
$$w'_{k,j} = \begin{cases} w_{k,j} & \text{if } w_{k,j} \geqslant \tau, \\ 0 & \text{if } w_{k,j} < \tau. \end{cases}$$

By pruning the model we can reduce the number of prototypes in each explanation, simplifying the cognitive load on the clinician.



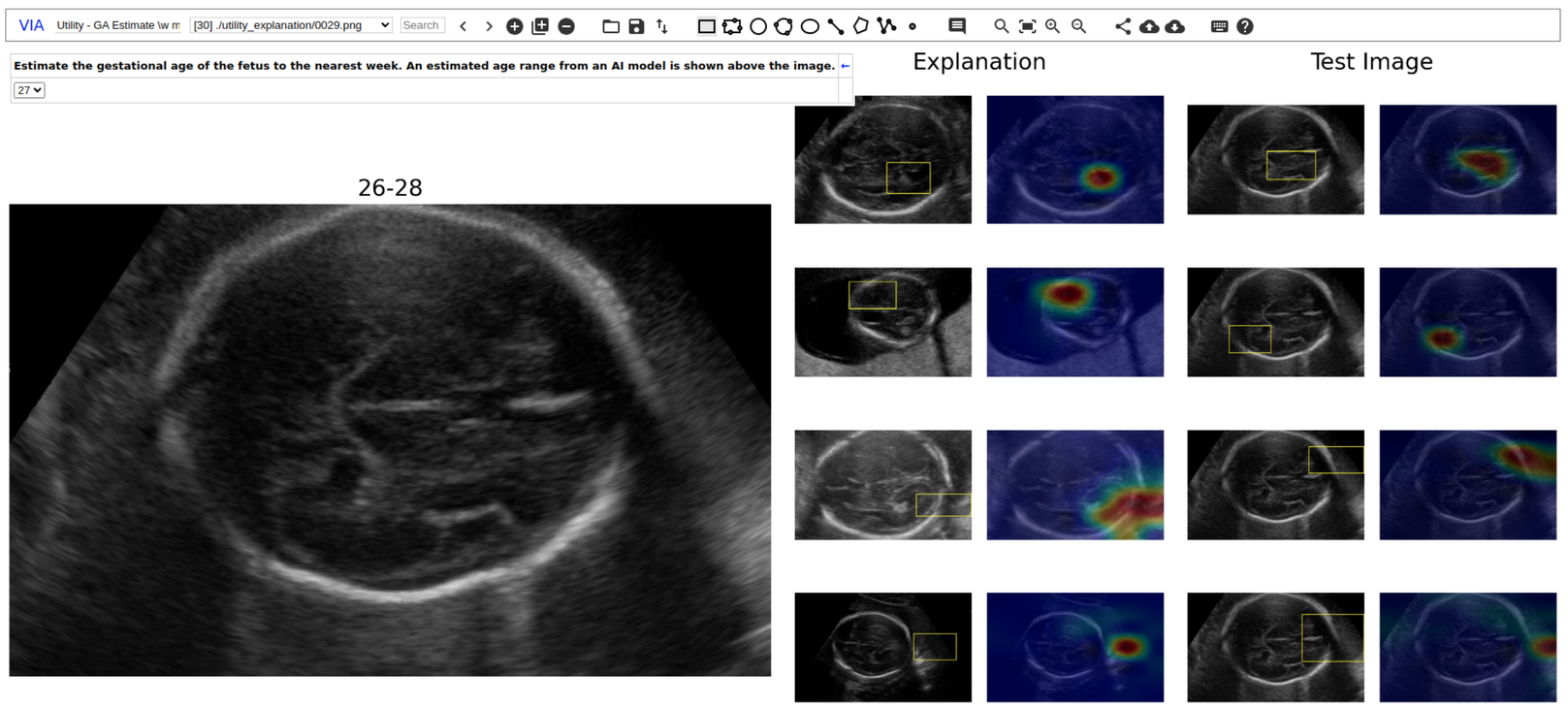
Preliminary Results

Preliminary results with a single clinician suggest the algorithm greatly improves their ability to estimate gestational age without performing biometry.



In the complete study we will use a model pruned with $\tau = 0.25$ so that the displayed prototypes account for 79% of the explanation, as opposed to only 43% for the unpruned model in this work.

VGG Image Annotator (VIA)



Screenshot of phase 3 of the study. On the left is the test image with the model's prediction above. On the right is the model explanation. Each row is a different prototype with the highest contributing prototypes at the top. The first two columns show the similarity heatmap and the bounding box containing 95% of the similarity for the prototype and the right two columns show the same information but for the test image. The software used to create the survey is the VGG Image Annotator (VIA).

References

[1] Papageorghiou, A. T., Ohuma, E. O., et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *The Lancet*, 2014.

[2] Nicolson, A. Yarin, G., Noble, A. Sparse Explanations for Gestational Age Prediction in Fetal Brain Ultrasound, *ICML, IMLH Workshop*, 2022.

[3] C. Chen, O. Li, et al. This looks like that: Deep learning for interpretable image recognition, *NeurIPS*, 2019.

Acknowledgements

A. Nicolson is supported by the EPSRC Center for Doctoral Training in Health Data Science (EP/S02428X/1).