**////Title: Teaching Algorithms to Caption Ultrasound Images**

**////Standfirst:**

Medical professionals require years of training before they can describe ultrasound images of developing foetuses. Dr Mohammad Alsharid and colleagues from the Institute of Biomedical Engineering and Nuffield Department of Women’s and Reproductive Health at the University of Oxford suggest that this task could one day be carried out by machine learning algorithms. In their latest study, the team showed how neural networks, trained by the expert knowledge of real sonographers, could convert subtle features within the images into accurate, readable captions.

**////Main text:**

Ultrasound imaging is one of the most widely used techniques for monitoring foetus development inside the womb. It works by emitting high-frequency sound waves from a probe on the surface of the skin, and then detecting the waves reflected back to the probe from anatomical features within the growing foetus. Since these waves arrive at the probe at different times, depending on the distance they travelled before being reflected, this approach allows doctors to precisely identify where foetal organs are located.

In the UK, all women are offered a detailed ultrasound scan roughly 20 weeks into their pregnancy – allowing doctors to identify any potential birth defects or medical issues that may urgently need attention. Currently, every ultrasound scan must be analysed by trained sonographers, who draw on years of experience to describe extremely subtle features in ultrasound images.

…

A key role of sonographers is to provide precisely detailed captions of ultrasound images, which explain their content clearly for those with non-expert knowledge. Since the anatomy of a growing foetus is so complex, these captions must encompass a rich vocabulary of nouns, verbs, and adjectives. Inevitably, this process is costly and time-consuming; and due to human error, there is no guarantee that even the most experienced sonographers will make accurate assessments 100% of the time.

In many other fields of research, similar problems are being addressed using ‘machine learning’ algorithms. If they are implemented correctly, these algorithms can perform complex analyses more effectively than even their most highly-trained human counterparts.

So far, the use of machine learning in describing medical images is relatively unexplored. The issues with this approach are twofold: not only is it incredibly challenging for these algorithms to match the expert knowledge of trained medical professionals; but when the health of a developing foetus and its mother are both at stake, any damaging oversights caused by computational errors would be ethically inexcusable.

Being able to describe foetal ultrasound images may also be useful in remote settings, where there may be a lack of expert sonographers to adequately explain what is shown on the screen. Effectively, the trained algorithm could be used to communicate medical information to a layperson in the same way an expert would.

…

To meet the critically stringent requirements of this task, Dr Alsharid and his colleagues turned to a branch of machine learning techniques called ‘neural networks’. They are named so because their behaviour is somewhat inspired by the behaviour of neural networks in the brain. In a human brain, neurons are continually forming connections with each other, and transmitting information in the form of electrical signals. As we learn new information, neurons then alter the strengths of these connections, depending on how closely our previous responses to the information reflected an ideal response.

To mimic the speech of sonographers as they scan, Dr Alsharid and his colleagues drew on two types of neural network. The first type is a ‘convolutional’ neural network, which is built for handling images, while the second is a ‘recurrent’ neural network, which is designed for text. These algorithms can work by either generating novel text descriptions from images, or by storing descriptions beforehand, and using them to describe objects in images – before the words are combined into readable sentences.

…

To reduce the computational complexity of this process, Dr Alsharid and his colleagues merged the outcomes of convolutional and recurrent neural networks together in a novel way. This ensured that ultrasound image features and their associated text could be jointly learned in a single step, producing annotations that were perfectly matched with their associated visual features.

The researchers trained their neural networks using a set of 2,240 video frames, taken from a real full-length, second-trimester ultrasound scan. These images were accompanied by captions in the form of transcribed voice recordings, provided by an expert sonographer. Afterwards, the combined neural networks were applied to generate captions for video frames that had never been encountered before.

…

To measure the accuracy of these captions, a medical expert in the team was asked to assess how closely the generated words reflected their own assessments, when describing anatomies including the foetal spine, abdomen, head, and heart.

The neural network’s performance was not perfect, but it was able describe ultrasound images with anatomically relevant captions using words that would be used by sonographers themselves. However, for images of the head and abdomen, which require particularly broad vocabularies to describe, the captions didn’t always match the descriptions of real sonographers. For an image depicting an abdomen, for example, the caption mentioned the visible stomach; however, when this image was acquired, the real sonographer was talking about a rib, which was also visible in the same image. It is important to note though that both descriptions were anatomically relevant to the image.

Describing the heart proved to be more challenging compared to other anatomical structures. In real ultrasound scans, the heart must be viewed from multiple angles, and can be identified by its characteristic heartbeat. As a result, this issue could potentially be solved through future improvements to the technique – where merged neural networks could be trained to caption moving images in real time. Encouragingly, the algorithms already displayed a fairly strong performance when analysing features that don’t require a diverse range of words to describe accurately, such as for the spine.

…

While there is still some way to go before the team’s method can compete with the assessments of real sonographers, their results present promising first steps towards a highly advanced approach of ultrasound analysis.

In the future, the Oxford team hopes to adapt their algorithms to express a broader range of vocabulary, and to handle dynamic video clips that change over time. If achieved, this could allow the approach to account for how a developing foetus kicks and rolls, how their heart beats, and how the ultrasound probe is moved to view foetal anatomy from different angles.

…

This SciPod is a summary of the conference proceedings ‘Captioning Ultrasound Images Automatically’, from the International Conference on Medical Image Computing and Computer-Assisted Intervention. [doi.org/10.1007/978-3-030-32251-9\_37](https://doi.org/10.1007/978-3-030-32251-9_37)

For further information, you can connect with Dr Mohammad Alsharid at [mohammad.alsharid@eng.ox.ac.uk](mailto:mohammad.alsharid@eng.ox.ac.uk)

This research is part of the Pulse Project that has received funding from the European Research Council under the European Union’s Horizon 2020 research and innovation programme – grant number 694581.