Deep Learning Model Uncertainty in Medical Image Analysis

Dušan Drevický, Oldřich Kodym
Department of Computer Graphics and Multimedia
Faculty of Information Technology, BUT

Motivation

Deep convolutional neural networks (CNNs) achieve super-human results in image analysis but their outputs lack reliable information about the uncertainty of their predictions which prevents their wide-spread adaptation in medicine.

In this work we:
- Propose and train a CNN for the task of automatic cephalometric landmark localization on skull X-rays. This is usually done by a dentist manually and is a time-consuming and tedious process.
- Design and implement uncertainty metrics accompanying the trained models which are able to provide estimates of how certain the network is of its predictions (i.e., how much we should trust the predicted landmark positions).

Solution: Landmark Localization

- CNN was trained on a dataset [1] of 200 X-ray cephalograms (annotated with 19 landmark positions).
- Training done via heatmap regression (annotated landmark location is used to create a 2D heatmap with a Gaussian spike at that location).

Solution: Uncertainty Measures

We proposed three models with corresponding uncertainty measures:
1. Baseline is a CNN based on U-Net [2]. It uses the value of the maximum heatmap activation as a measure of how uncertain the CNN is with its prediction (we assume lower values indicate higher uncertainty).
2. Ensemble is an ensemble model of 15 Baseline models based on the work of Lakshminarayanan et al. [3].
3. MC-Dropout is a CNN based on U-Net but additionally contains dropout layers which randomly remove some weights from the network both during training and evaluation (making the CNN stochastic so that multiple evaluations of the same image do not produce the exact same output). Based on the research by Gal et al. [4].

For both Ensemble and MC-Dropout models the prediction mean is used as the final predicted landmark position and the prediction variance (of the ensemble members and Monte Carlo samples respectively) as the uncertainty measure. We assume that as variance increases so does the models’ uncertainty.

Experiment 1: Can Uncertainty Measures Detect Skull Rotation?

Correlation between skull rotation and the mean uncertainty measures. Uncertainty increases together with rotation for all measures. Note that we expect lower heatmap activation values as uncertainty increases and therefore a negative correlation coefficient for this measure.

Experiment 2: Can Uncertainty Measures Detect Deformed Data?

Correlation between elastic distortion magnitude applied to the test set and the corresponding mean uncertainty measures. Uncertainty increases together with distortion magnitude for all measures.

Conclusion

- Models achieve performance close to state-of-the-art on the studied landmark localization task.
- Uncertainty measures were able to reliably detect data unsuitable for automatic evaluation.
- Research conducted in cooperation with TESCAN 3DIM and is planned to be implemented as a feature in their medical diagnostic software.

References