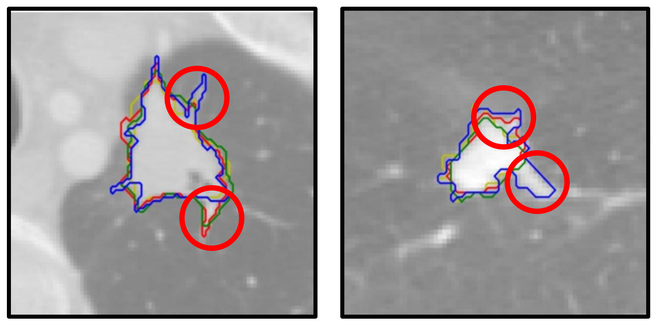
Confidence Contours: Uncertainty-Aware Annotation for Medical Semantic Segmentation

Read the paper!

Andre Ye
Jim Chen
Amy Zhang

1 Problem

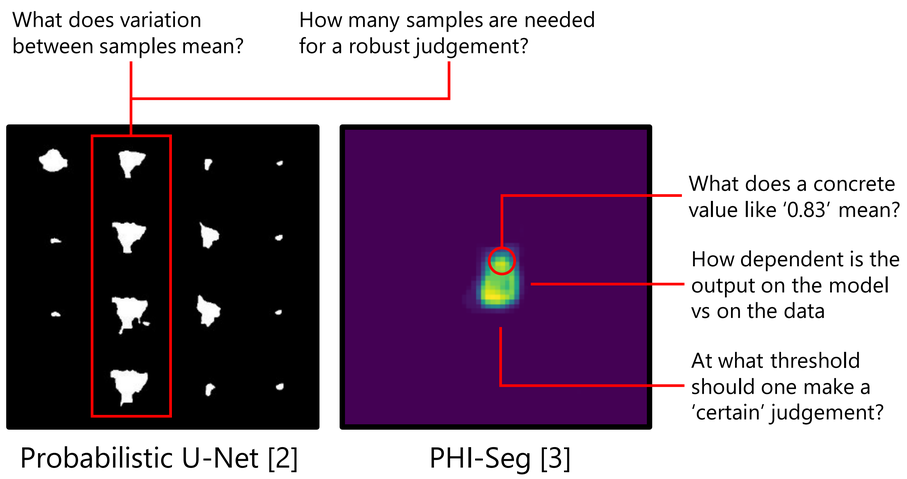
Medical imaging problems often feature structural uncertainty. Existing models train on singular segmentation maps, which show no uncertainty info.



[1]

= regions of high structural uncertainty; disagreement among annotators over inclusion or exclusion of region

Existing work in uncertainty-aware semantic segmentation **attempts to model uncertainty from multiple 'certain' inputs** by modifying the model while still training on singular segmentation masks.



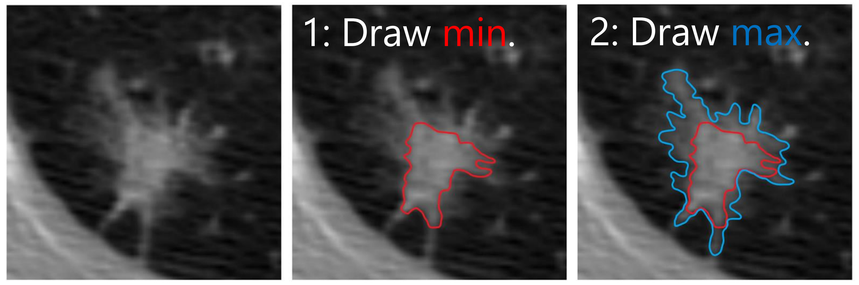
These outputs are **not clearly interpretable by humans**, **contingent on arbitrary parameters,** and **unreliable for medical judgement-making.** [4]

> **Model-centric approaches disconnect uncertainty representations from human judgement**, even though uncertainty is, at root, a tool for human decision-makers.

[1]: LIDC dataset. [2]: Kohl et a. 2019. [3]: Baumgartner et al. 2019.
[4]: Jungo et al. 2020, Ng et al. 2020.

2 Solution

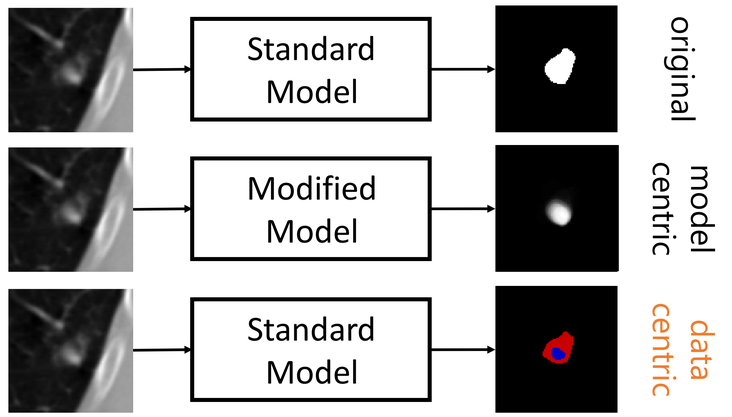
Confidence Contours takes a data-centric approach in which uncertainty is marked **directly by annotators** rather than inferred after the fact.



Training models on CCs requires no model modifications, compared to singular segmentation tasks.

Models trained on CCs produce uncertainty maps which directly correspond to human annotators' uncertainty judgements.

No black-box uncertainty inferences!



3 Experiments & Results

Recruited 45 participants to annotate 600 images across two datasets (LIDC – lung nodule segmentation, FoggyBlob – synthetic), with 3 singular and 3 CC annotations each.

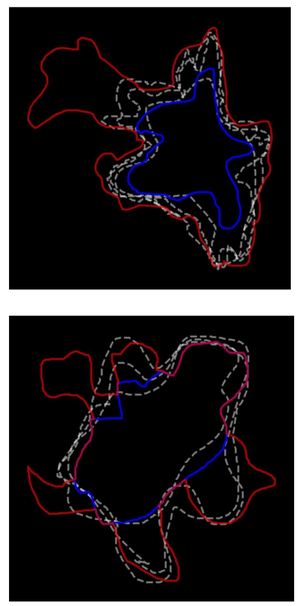
Annotators do not find CCs significantly more burdening to annotate

- ≤1.3 difference between mean singular and CC task load across all NASA TLX dimensions (10-point scale)
- Average 27 sec for singular annotation and 44 sec for CCs

CCs have statistically significantly higher representative capacity than singular annotations for bounding multiple singular annotations.

Mins reduce cross-annotator disagreement compared to singular annotations, from 0.72 to 0.60 (*) for LIDC.

CCs expand the amount of annotated information in masks. The max is on average 25.6% larger than a singular annotation for LIDC and 17% larger for FoggyBlob.



Dotted white: singular annotations. Blue: single min. Red: single max.

CCs not only bound the range of variation across standard annotations, but the max extends the annotation range.

Trained 156 models across 4 architectures. Observed no significant differences in performance. **Standard segmentation models are as capable of learning CCs as singular annotations**.