Embracing Human Subjectivity in Computer Vision

Introduction. Computer vision research builds datasets that assign a single human label to images. Models are often trained under the implicit assumption that this label is indisputable. I am interested in challenging this assumption by exploring how **data and machine learning systems can better reflect and engage with human subjectivity** across a wide variety of real-world problems.

[1/3] ***Learning from noisy human labels.*** **Researchers often want to build systems that automate human judgment tasks for improved accuracy and efficiency.** However, in tasks across many domains, humans produce different judgements due to differing domain knowledge, annotation habits, or perceptual tendencies.

I worked on building a custom segmentation model for a kidney artifact segmentation task in a pathology lab. Because this task was highly specialized, there were no suitable existing datasets. Instead, my collaborator and I built an annotation interface and facilitated in-house annotation collection. Because the kidney structures of interest are small and numerous, as well as visually similar to false-positive artifacts, annotators often made errors and disagreed with each other. Our challenge was to **build a reliable model from 'unreliable' data**.

Our system used both classic computer vision and deep learning methods. First, the model segments the membrane along which structures lie. Then, another model segments structures within overlapping "windows" along the membrane, which are ultimately combined into a unified segmentation. Our approach substantially outperforms standard models: by structurally breaking apart the problem, annotator errors are isolated and can be automatically double-checked; further, they do not propagate as far during learning. This system is currently being evaluated as an assistive tool in a clinical setting.

[2/3] ***Faithfully representing human subjectivity.*** While working on the kidney segmentation project, I saw the need for human-oriented model design and became interested in systems that faithfully capture and leverage human subjectivity rather than working around it.

Current approaches work around the 'problem of subjectivity' by interpolating an uncertainty distribution from existing samples. However, the distributions produced by these models are a) cognitively difficult for humans to interpret as judgment-making tools and b) have not been shown to have clinical significance/meaning.

In response, we **reconceptualize segmentation annotations to explicitly include human-provided uncertainty**. Confidence Contours (CCs) are a simple and easily interpretable modification of standard annotations. Rather than providing a singular segmentation, annotators annotate both a region of high confidence and additional areas of lower confidence (a "range" of possible segmentations). I ran user studies to collect 3.6k annotations from 45 participants through a custom annotation interface I built. Then, I analyzed and trained models on the user annotations, finding that CCs reduce disagreement and increase information coverage compared to standard annotations. Moreover, I interviewed medical experts and found that CCs provide more useful judgment information in practice. This work shows that

letting human annotators represent uncertainty produces more useful models. I led the project, wrote the paper, and revised it with my mentor and PI (accepted to HCOMP '23).

[3/3] Uncovering and embracing human subjectivity. As the preceding two projects show, **human subjectivity is an essential part of human judgment tasks** and cannot be excluded from "ground truth" labels. In my next project, I sought to demonstrate this not only in specific medical domains but also in a foundational computer vision task, image captioning.

Currently, computer vision models train image captioning models using web images associated with text descriptions. Work in psychology and linguistics[1] shows that cultural background and language affects individuals' visual perception and expression. I worked to uncover whether these perceptual differences may have manifested in vision datasets and models by using scene graphs and other representations to study the semantic content mentioned across descriptions for the same image across multiple languages. We find **substantive differences in the visual information captured by annotations in different languages** in both datasets and models, and that models trained on multilingual content generally perform better than monolingual models. Our work highlights the need to examine and diversify annotators' cultural and linguistic backgrounds in datasets. More generally, it **demonstrates that embracing diverse modes of perception can improve visual understanding.** I led the project, wrote the paper draft and revised it for submission to ICLR 2024 with my mentors and PIs.

Conclusion. In graduate school, I plan to continue studying human subjectivity in machine learning systems, with an even stronger focus on the social sciences, particularly cognitive science and philosophy. I believe building stronger connections between these two areas is essential to building more robust and usable models.

[1] See:
Edward T Hall. *Beyond culture*. Anchor, 1976.
Richard E. Nisbett and Takahiko Masuda. "Culture and point of view." Proceedings of the National Academy of Sciences of the United States of America, 100:11163 – 11170, 2003.
Jakob Prange and Nathan Schneider. "Draw mir a sheep: A supersense-based analysis of german case and adposition semantics." KI-Kunstliche Intelligenz , 35(3-4):291–306, 2021.