Stanford eCorner

Raw Data + Human Insight = High-Quality AI

Alexandr Wang, *Scale AI*

26-01-2022

URL: https://ecorner.stanford.edu/clips/raw-data-human-insight-high-quality-ai/

Scale AI founder and CEO Alexandr Wang examines the role of human annotators in building high-quality AI and machine learning systems. He explains the systems and technology his company has built to empower human annotators to transform raw data into useful annotated data.

Transcript

- One of the fundamental beliefs 00:00:05,957 that we have at Scale is that there's sort of a few resources that go into producing really high-quality machine learning systems.. If you think about what are the raw ingredients, one of those raw ingredients is compute, one of those raw ingredients is data, and then the last raw ingredient is sort of like human insight, so to speak, and sort of the process of producing annotated data is combining raw data and human insight together and sort of mixing those ingredients together.. And all AI systems of the future are going to be heavily reliant on human insight, even if we have incredible advancements in the technology, etc., because, at the end of the day, at minimum, we're going to need oversight in these AI systems to make sure that they're built, they're producing the results that we would expect, and producing the results that we think make sense and whatnot.. And so, we think this problem, one of the core problems of machine learning is, how do you make sure that for every problem that you might wanna solve using machine learning or you might wanna have a machine learning system do, that you're able to effectively get human insight to sort of power that use case.. And the way we think about it is twofold.. I think first is, for folks who are trained up, you want those individuals to have as much leverage as possible.. So you want, and this is somewhat circular, but it becomes very important, you want machine learning systems to do as much of the work as possible and then, really, have humans spend all of their effort on almost, quote, unquote, the edges or on the really high-judgment work that's required that enables the utmost quality while making them the most efficient in their roles.. So that's the first piece.. And the second piece is sort of this general education problem around how do you enable people to be experts, and how do you train them most efficiently? And one of the almost funny side effects of solving this problem at scale is that I think we have one of the, probably the more interesting edtech systems out there, in which we have systems internally which track all the different kinds of edge cases might be one way to talk about it, but all the different kinds of nuances to some of these data problems, right? This cilantro versus parsley example could be one of them, and the different ways in which you can tell, or detecting a melanoma versus other kinds of growths on skin.. So, there's all these different nuances, and we track all these different nuances within our internal systems, and we are constantly trying to understand, for each of the annotators, which nuances, or which edge cases are they performing well at and they understand super well, and which ones might be tripping up, and then proactively serving them materials and content and examples that help them elucidate these cases that they might not understand super well..

And so, I really think about it as an edtech system which enables people to really quickly grok and understand all the nuances of the data.. And humans are incredible pattern recognition, right? If you think about what it takes to go through medical school and you talk to folks who've gone through medical school, it's an incredibly long process of just sort of continued pattern recognition.. And that's sort of that same process where we try to distill for the annotators in our system to enable them to do great work...