The ‘Cheap’ Decisions That Can Affect Image Synthesis

Originally published at https://arquivo.pt/wayback/20240203191557oe_/https://blog.metaphysic.ai/the-cheap-decisions-that-can-affect-image-synthesis/

January 13, 2023

Some of the most important decisions in artificial intelligence research are being made by people, mostly in the United States, earning on average $2 an hour, and routinely trying to cheat a system that rewards efficiency over integrity.

These are the crowdsourced, *ad hoc* freelance workers of Amazon Mechanical Turk (MTurk), currently the most popular and affordable resource for cash-strapped computer vision researchers looking to evaluate tasks as diverse as image disentanglement in text-to-image models; the image quality of Stable Diffusion output; and testing for bias in new latent diffusion systems – among a slew of similar cases familiar to any regular reader of computer vision or general machine learning research papers.

Once limited to under-budgeted studies using campus students, it’s now practically routine for researchers to validate the results of novel or incrementally-improved image synthesis frameworks and architectures at greater scale through MTurk – using remote workers, who (since 2012) are required to be located in the United States.

Major academic institutions (such as Berkeley) offer guidelines to researchers for using the system, and the consistent stream of studies outlining both ethical and practical issues with MTurk don’t seem to be causing much reflection on the practice, because it’s a very cheap service.

The People Upstream

If you follow the code far enough upstream, the decisions that power the actions of generative image systems are eventually human in some way. For instance, even though the web-scraped image/caption pairs that power the formidable generative capabilities of Stable Diffusion were evaluated algorithmically through OpenAI’s Contrastive Language-Image Pre-training (CLIP), even this apparently human-free operation was *personally* judged and validated (by *‘five different humans’*, according to the official paper, which gives no further details of this modest study group).

So in the case of web-scraped data that will ultimately form training sets for AI systems such as Stable Diffusion, there are always at least three significant human influences: a) the people who originally captioned (or at least *named*, i.e., *‘bear.jpg’, ‘hotgirl.webp”*) the uploaded images, who will usually have done so for SEO purposes rather than to help machine learning systems understand the relationships between pictures and words; b) the human annotators who label subsets of datasets which are too vast and expensive to hand-label in their entirety, and whose work is used as a ‘control set’ for automating the labeling of the majority of the data; and c) the human race itself, whose predominant passions (*sex*, *money*, *hot women*, *beauty*, *horror*, *shock*, *music*, *conspiracies, weapons*, etc.) are so out of proportion to common daily experience that they become statistically over-emphasized in the data, and in the output from systems trained on that data.

Therefore the net influence of casual human evaluation on data that contributes to image synthesis systems is not negligible, and the people tasked as the ‘front line’ against unbalanced data are fallible, under-managed and relatively little-understood – even though, in the case of the dominant MTurk, they’re now essentially writing history itself, for, on average, far less than minimum wage.

The Challenge of Triage

To address at least part of the problem, a new collaboration from the University of Notre Dame, KAIST and the University of Minnesota has proposed a method to predict when labeling is likely to ‘go wrong’ among human annotators, by considering their demographic make-up, and developing systems capable of predicting the annotators’ outcomes.

Titled *Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information*, the new work proposes a *disagreement predictor*, for cases when a value judgement conflict emerges between at least three annotators who have given differing responses to the same data, during a trial.

Datasets	Text	Annotation Distribution	Disagreement Label
SBIC	"Abortion destruction of the nuclear family contraceptives feminism convincing women to wait for children damaging economy so youth cannot leave the nest ramping up tensions between sexes all serves one primary goal to lower the population."	A1 (age: 32, politics: liberal, race: white, gender: woman) votes for inoffensive A2 (age: 34, politics: liberal, race: white, gender: woman) votes for inoffensive A3 (age: 29, politics: mod-liberal, race: hispanic, gender: woman) votes for offensive ⟶ Aggregated Label: **inoffensive**	Binary: 1 Continuous: 1/3
SChem101	"It's okay to have abortion."	A1 (age: 30-39, education: high school, race: white, gender: woman) votes for people ocassional think this A2 (age: 40-49, education: grad, race: white, gender: man) votes for controversial A3 (age: 30-39, education: bachelor, race: white, gender: man) votes for common belief A4 (age: 21-29, education: high school, race: white, gender: woman) votes for controversial A5 (age: 30-39 , education: bachelor, race: hispanic, gender: woman) votes for controversial ⟶ Aggregated Label: **controversial**	Binary: 1 Continuous: 2/5
Dilemmas	1st action: "refusing to do a survey on the credit card reader while paying with cash at the Office Max." 2nd action: "saying my bf has no right to dictate who I tell about my abortion."	1 annotator votes for the first action is less ethical while 4 others vote the second action is less ethical ⟶ Aggregated Label: **2nd action is less ethical**	Binary: 1 Continuous: 1/5
Dynasent	"Had to remind him to toast the sandwich."	4 annotators believe it's negative while one think it is neutral ⟶ Aggregated Label: **negative**	Binary: 1 Continuous: 1/5
Politeness	"Where did you learn English? How come you're taking on a third language?"	5 annotators politeness scores are 5, 13, 9, 11, 11 with the maximum of 25. ⟶ Aggregated Label: **impolite**	Binary: 0 Continuous: 0

Examples of disagreement from five different datasets evaluated in the new work, with 'A' standing for 'Annotator'. Source: http://export.arxiv.org/pdf/2301.05036

There are two scenarios in which this kind of triage can lead to a non-optimal result: one is where a minority opinion happens to be better-informed or more credible than those outvoting it; and the other is where the schema of the study or tendency of the supervising researchers automatically discards any rounds in which the participants could not reach easy agreement; and this usually occurs because of lack of resources to deal with the conflict.

The first of these cases could be exemplified by two respondents voting *yes* to the question 'Is Baltimore a safe place to live?', and a third voting *no*. If the first two don't live anywhere near Baltimore, and perhaps have never been there, while the third respondent is a lifelong Baltimore resident, it's arguable that the wrong voice got silenced. Even though all possible responses are naturally subjective, the downvoted respondent likely had additional information backing up their answer.

Therefore the system developed by the researchers is capable of predicting disagreement where both the data to be analyzed and some basic demographic details of the researchers are known. The framework is capable of predicting disagreement in a generic manner as well, without the benefit of any insight into the respondent's demographics.

The 'Easy Way'

Regarding the second aspect: though the system is intended to 'shed light on various applications of data annotation', it could also effectively act as a way for researchers to anticipate and avoid roadblocks to fruitful (or at least theory-validating) survey results, either by shaping the format, style or content of the test data until conflict predictions drop adequately, or by choosing researchers whose demographics are more likely to reduce the number of conflicts.

Though this is not the intention of the work, it could be an undesirable collateral effect. In any case, it's a scenario almost unique to the MTurk era, since pre-internet polling was far more likely to gather generically similar groups into the respondent pool – or at least to be undertaken by people with some broad geographical or ideological connection – which is the kind of trait that's hard to obtain in crowdsourced respondents, who are disparately located, and who have a tendency to game the system for financial reward.

That 'failure teaches more than success' is even more applicable in science than in life; but this is a hard truth for the under-funded AI researcher. In late 2021, a paper from Google Research, which examined the influence of individual and collective participant identities on the quality of crowdsourced responses, stated:

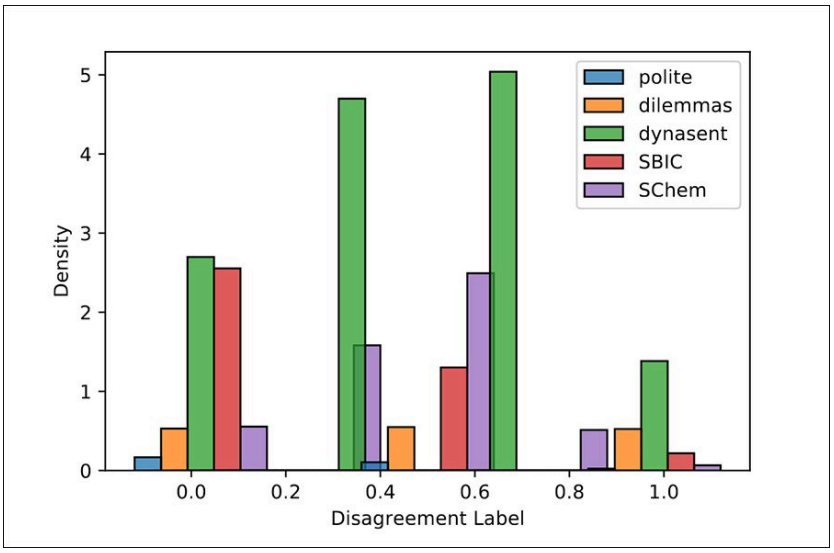
'[The] notion of "one truth" in crowdsourcing responses is a myth; disagreement between annotators, which is often viewed as negative, can actually provide a valuable signal. Secondly, since many crowdsourced annotator pools are socio-demographically skewed, there are implications for which populations are represented in datasets as well as which populations face the challenges of [crowdwork].

'Accounting for skews in annotator demographics is critical for contextualizing datasets and ensuring responsible downstream use. In short, there is value in acknowledging, and accounting for, worker's socio-cultural background — both from the perspective of data quality and societal impact.'

Approach and Tests

For the new work, the researchers modeled annotation disagreement using annotators' demographic information as additional inputs in the pre-trained language model RoBERTa. The model was fine-tuned with the Adam optimizer at a learning rate of 1e-5 on Adam's default hyperparameters. For text classification, the training was undertaken at batch size 8 overt 15 epochs.

The benchmark datasets used were *Social Bias Inference Corpus* (SBIC); *Social Chemistry 101* (SChem101); *Scruples-dilemmas*; *Dyna-Sentiment*; and *Wikipedia Politeness*.



The disagreement distributions found via the model across the five datasets.

Though the schema was initially calibrated to quantify 'controversy' as a contributing quality for disagreement and consensus, the authors note that participant demographics are a notable potential contributor to outcomes. They therefore incorporated both *group* and *personal* demographic information, based on the available data for the datasets.

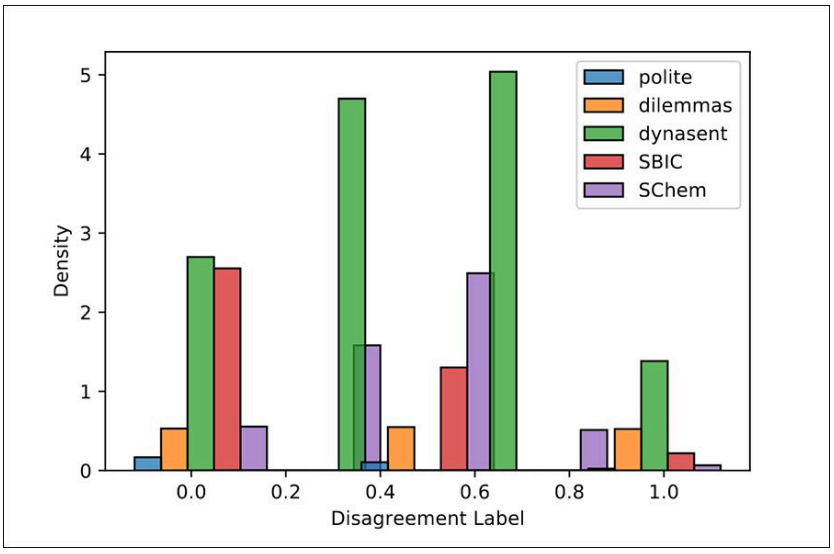
Prior to training, these two strands of data (the data itself, and characteristics of the annotator) were concatenated both as formulaic and natural language sentences. From the paper:

'[We] propose two different ways with specific templates: (1) Templated format and (2) Sentence format. Templated format represents the category and value of each demographic information in a separate sentence, then concatenate all of them with the given text.

'For example, if one annotator is 36 years white woman, this demographic information is converted to "Age: 36, Color: white, Gender: women", then concatenated with the original sentence in case of the text with person demographic.

'On the other hand, sentence format represents the demographic information with a natural sentence, e.g., the annotator is a 36 years old white woman., then concatenate it with the original sentence.'

With these real cases trained, the researchers went on to use the model with artificial (i.e., fictitious) data, concerning non-existent gender and ethnicity types, thus obtaining predicted disagreements in accord with the median results from the real data.



Disagreement prediction using simulated demographic information on two of the benchmark datasets, with varying shapes and colors indicating diverse disagreement labels as per the legend.

The purpose of the benchmarking across the five datasets was to establish whether the system can differentiate between cases where the demographic makeup of the group constitute a primary potential roadblock, and where, instead, the material itself is so charged or controversial that to account disagreement to demographic differences would not be reasonable.

For text-only results (i.e., without a demographic factor), continuous agreement, also known as *soft disagreement*, achieved better prediction than binary disagreement (also known as *hard disagreement*), across most of the datasets.

	Binary Label		Continuous Label	
Datasets	F1 (↑)	MSE (↓)	F1 (↑)	MSE (↓)
SBIC	61.5	0.309	**66.0**	**0.086**
SChem101	0.0	0.905	**52.3**	**0.056**
Dilemmas	0.0	0.330	**34.2**	**0.165**
DynaSent	**74.9**	0.361	11.8	**0.114**
Politeness	55.9	0.490	**56.8**	**0.110**

Results of binary and continuous disagreement prediction across the tested datasets.

When also considering demographic information, the researchers found that personal-level demographics improved the prediction accuracy better than group demographics, and they observe:

'One potential reason is that the annotator's level of demographics may imitate the annotation process that each annotator labels the text without knowing each other. And also because concatenating personal level demographics can be considered as oversampling that group-level setup can not.'