ML4Seismic Partners Meeting 2023 A New Seismic Fault Label Uncertainty Dataset: Insights from Expertise, Certainty, and Consistency

Jorge Quesada, Chen Zhou, Mohit Prabhushankar, and Ghassan AlRegib





Introduction

Where does annotator disagreement stand in the interpretation pipeline?

Proposed Uncertainty Framework



Introduction

Where does annotator disagreement stand in the interpretation pipeline?



Introduction Multiple annotators can disagree during data labeling







Introduction When not addressed, annotator disagreement hinders downstream performance



Noisy labels



5 of 19

[ML4Seismic 2023] | [Jorge Quesada] | [November 8th, 2023]

Fleisig, A., Abebe, R., Klein, D. (2023). When the Majority is Wrong. *arXiv preprint arXiv:2305.06626v* Wang, S., Li, C., Wang, R., ... & Zheng, H. (2021). Annotation-efficient deep learning for automatic medical image segmentation. *Nature Communications*





Introduction How is annotator disagreement generally dealt with?





6 of 19

[ML4Seismic 2023] | [Jorge Quesada] | [November 8th, 2023]

Mostafazadeh Davani, A., Diaz, M., Prabhakaran, V. (2022). Dealing with Disagreements. Transactions of the Association for Computational Linguistics





Introduction Where does annotator disagreement stem from?







Introduction Novelty of Our Work

A dataset comprising labels across multiple levels of *expertise* and *confidence* Expertise Confidence Fault (uncertain) **Experts Practitioners** No fault (certain) Novices Fault (certain)





Summary of our dataset We leverage Amazon Mechanical Turk for the labeling process

- 400 images, divided into 20 batches
- For each batch, 2 images are repeated 3 times for quality assessment
- 2 bonuses:
 - Number of images, promotes full dataset completion
 - Consistency, promotes thorough labeling

Current contribution:

• 1 expert, 8 practitioners







Experimental design - Platform What is Amazon Mechanical Turk (MTurk)?

MTurk allows for a distributed outsourcing of virtual tasks

- Crowdsourcing marketplace
- Distributed workforce can perform tasks virtually
- Data validation and research, survey participation, content moderation, etc

2	Summer 1	22	
Requesters have tasks they	MTurk Marketplace	Workers want to earn	
need to be completed	anala nanasipase	money and work on interesting tasks	





Experimental design - Labels Three label categories enable flexible annotations

- We present 1 image per HIT, and consider 3 fault label categories for different certainty levels:
 - Fault (certain)
 - Fault (Uncertain)
 - No fault (Certain)

- Other schemes considered:
 - Multiple imgs per HIT
 - Sliders and text input boxes for certainty





Experimental design - Batches

Task is divided in 20 batches for easy navigation

- We divide the dataset into 20 batches, with 20 unique images each
- Each batch also contains 3 copies of 2 redundant images for quality assessment, totaling 6 QA images per batch
- Total batch size --> 24 imgs

			1	10000		
i secolul	WY MARY Description	100	-	Br. and		Annual States
- yernen	WILLIAM DE LES		84.449	in the		And a state
Daskasjon	SCALE, 2023/024038/008/099742/9_0107w0w1xd1+400e-80ea-Coloradiation		\$1.00	Se ap		*
 Anazon Repeater Inc RVD Team 	following bestrag, please sprom. This is a cartie to test some things we are working on	5	80.01	Rm age	-	Serge & West
# CR	DestAul.aph-TextII		\$3.01	Smiaph	-	Arrest & West
Anazzar Responder Inc Tamain	APRON Pattern	238	33.02	(in apr		
i sentat	Guilding feedback 0x1-04-0323	tő.	\$9.00	20m apr	-	-
weedute	US Preactive Did-24-3023 - Trianslamy Experiment Remainstant Repart	29	80.00	22n api	-	-
Arrester Requester Inc Tartain	MotorTrainingPostmeTests 2023-19-29-192229	1	\$9.12	Ziriği	-	
Ansaton Requester free - Tarrata	Motion/Training/Pgoild-in/Incation/2023-10.24-152244	.6	99.02	-281 apr	-	-
0 sheepplich	Label (his perkipite) builts in service insigns (R20)	30	80.00	28n api	Provine	Anise & Miles
D allees galook	Later the prological tauta in sevenic mapes \$22	23	\$5.00	279(agi)	Press	Anners West
0.00	Ontriving and Transfill	155	80.01	31m api	Design	Arrest & West
Explorer Holdstrong Greyerally	TEST TITLE S	10	\$1.00	38mapi	-	Animal & Work
II alless gatech	Laber the protopical faults in second integers 7/02	21	88.00	38n age	-	Annal & Web
C Explorer Weilsburg Leversty	THET TITLE 2	4	\$1.00	30m apr	Press	Arrest & More
dien prich	Label the geoingical faults in element images R20	17	\$3.00	39m apt	-	Surger & West
atten prinds	Label the protoplical basis to service images \$522	21	81.00	41m.4pi	Present	Accept & West
U other painch	Label the geological faulte transmic trages 400	21	90.00	41m apr	Report.	Accept & More
ProbalPase	Perform a service burgers direct to order ProductPrints inffit: Requires a Ma		\$25,10	4210 1000	-	Annual & Mark



12 of 19





Experimental design - Instructions

Concise guidelines aim to allow novices to annotate effectively

- We provide an instructional video in the task website and inside the layout, with the following details:
 - Fault definition
 - Sample image and label meaning
 - Platform usage
 - Payment scheme





13 of 19



Experimental design - Payment

Concise guidelines aim to allow novices to annotate effectively

- Base pay:
 - Reviewed existing mturk tasks and literature to arrive at initial pay rate
 - Adjusted pay rate for average time after internal task completion
- Number of image bonus:
 - Prorated bonus applied to final section of dataset to motivate completion
- Consistency bonus
 - Internal self-agreement metric for quality assessment







Insights from current annotators - Speed Domain expert is generally faster than intermediate users

Expert is significantly faster than intermediate users

- Average labeling time: 75s
- Expert average time: 30s







Insights from current annotators - Label usage

Annotators use less labels over time

Most people use only two labels or less (on average)

- Most people use only two labels on average
- Expert uses generally less labels than practitioners







Insights from current annotators - Certainty Confidence oscillates throughout task

Annotators with the most exposure make more confident labeling

- Confidence taken only for the 2 fault labels
- Most experienced annotators
 label more confidently







Agreement measurements

Intra and inter annotator agreement varies significantly

Expert is very distanced from other annotators as well as himself



Intra-annotator disagreement



Inter-annotator disagreement



18 of 19





Conclusion

Cross-expertise and multi-confidence dataset

- Uncertainty insights can help machine learning community build better models and methodologies to account for annotator disagreement
- A better understanding of the expertise gap can lead to more efficient fault labeling pipelines, reduced expert workload, and better fault detection models
- Our experimental design also constitutes a valuable resource for the seismic community to harness crowdsourcing platforms for efficient data labeling and annotation

For more OLIVES content, please visit:





