

LADI v2: Multi-label Dataset and Classifiers for Low-Altitude Disaster Imagery

Supplementary Material

A. Appendix

A.1. Training details

In Table 1 we present all networks we trained and evaluated on the LADI v2 dataset in the initial phase of the model selection process, sorted by test mAP score.

A.2. v1 and v2 Comparison Class Mappings

In Table 2 we provide the class mappings used to compare the LADI v1 and v2 dataset performance.

A.3. Label matrices

In the main paper, we combined the label-label and label-event matrices for the training and validation splits for the interest of space since they have similar distributions. For transparency, we reproduce the label-label and label-event matrices for each individual split in Figure 1

A.4. Additional Performance Characterization

Incident hazard type analysis. The performance between the test and validation sets is comparable for most hazard types, except for fire (see Figure 2a). The difference in performance for fire incidents is likely due to the relative lack of fire data in the training and validation sets compared to the test set.

Geographic analysis. The classifier performs robustly across various geographies. Figure 2b shows the mAP of the classifier for states in the test set, which are those states in which CAP collected images for disasters in 2023. While only 10 states are represented in the test set, the classifier achieved an mAP between 80 and 96 for all of them. The state of Hawaii, with the lowest mAP at 80, had only one CAP mission in the test set, the August 2023 Hawaii wildfires. Relatively low performance is likely due to the relative lack of fire events in the training data, as well as potential differences due to geography. We thus caution practitioners and researchers against using models trained on LADI v2 for applications not well represented in the training set and recommend supplemental data collection and training.

A.5. Semantic Similarity Analysis

CLIP (Contrastive Language-Image Pretraining) [2] is a vision language model pretrained on 400 million image-text pairs from the internet. The model is given a batch of images and captions, and is trained to pair the associated image to its respective caption. In doing so, the model learns to align the encoded image representation to the respective encoded caption text representation. As a result, images

with similar textual descriptions tend to be closer together in the CLIP image embedding space, and dissimilar images are further apart. We use this property to characterize the distribution of our validation and test sets below.

We attempt to quantify “out-of-sample-ness” by using distance in CLIP space [2]. CLIP embeddings align images with similar textual descriptions, such that images with semantically similar content will be nearby in CLIP space even if they are not necessarily visually similar in their pixel representations. In this way, we can use distance in CLIP space as a proxy for semantic similarity between two images, where similar images are closer in CLIP space. We use the Euclidean distance between normalized vectors as the distance metric, $d(\theta) = \sqrt{2(1 - \cos \theta)}$, where θ is the angle between the two vectors. For each image in the validation and test sets, we compute the CLIP distance between it and its nearest neighbor in the training set. We also compute the L^1 norm of the error vectors (the difference between the post-sigmoid/pre-threshold prediction and ground-truth vectors) for each image in the validation and test sets.

We visualize the joint distribution of the L^1 error norm and distance to nearest training point for each image in the validation (blue) and test (orange) set in Figure 3. Kernel density estimates of the marginal distributions are visualized along the top and right hand axes. We can see that the test set is on average further away in CLIP space and has larger error norms. There appears to be a positive relationship between the distance to the nearest training example and the average error norm, as well as the variance in the distribution of the error norm. This approach could be used to characterize how out-of-sample a given set of images is, as well as estimate the potential expected degradation of performance associated with that distribution shift.

A.6. Vision-Language Model Prompts

A.6.1. LLaVA-NeXT Prompts

We prompted LLaVA-NeXT for each label individually. For a given image/label, LLaVA-NeXT saw a generic introduction followed by one of the following bullets:

Respond to the following question as accurately as possible with a ONE WORD yes/no answer. ONLY RESPOND WITH ONE WORD, 'YES' OR 'NO'. Question:

- Does this image contain bridges? Answer with one word, 'yes' or 'no.'
- Does this image contain buildings? Answer with one word, 'yes' or 'no.'

model	test mAP	val mAP
bit-50	0.890315	0.895946
swinv2-large-patch4-window12to16-192to256-22kto1k-ft	0.869659	0.881908
swinv2-large-patch4-window12-192-22k	0.859774	0.886702
vit-large-patch16-384	0.856639	0.869559
swin-tiny-patch4-window7-224	0.845096	0.850194
vit-large-patch16-224-in21k	0.834595	0.854939
deit-base-patch16-224	0.833991	0.847348
resnet-50	0.819502	0.825781
vit-base-patch32-384	0.797757	0.815523
vit-base-patch16-224-in21k	0.759668	0.711914
vit-base-patch32-224-in21k	0.738049	0.728242
mobilenet_v1_1.0_224	0.730085	0.647575
vit-huge-patch14-224-in21k	0.721056	0.651463
efficientnet-b0	0.712821	0.607476
swin-large-patch4-window7-224-in22k	0.705249	0.561424
mobilenet_v2_1.0_224	0.699678	0.614432
resnet-152	0.695796	0.743922
focalnet-base	0.695531	0.566548
convnextv2-large-22k-224	0.691524	0.565504
deit-base-patch16-224	0.640672	0.625217

Table 1. Performance comparison of initial selection of 20 models, sorted by test mAP performance

Original Class	Source Dataset	Mapped Class(s)
flood	v1	flooding, damage
rubble	v1	damage, debris
misc_damage	v1	damage
building	v1	building
road	v1	road
bridges_any	v2	
buildings_any	v2	building
buildings_affected_or_greater	v2	building, damage
buildings_minior_or_greater	v2	building, damage
debris_any	v2	damage, debris
flooding_any	v2	flooding, damage
flooding_structures	v2	building, flooding, damage
roads_any	v2	road
roads_damage	v2	road, damage
trees_any	v2	
trees_damage	v2	damage
water_any	v2	

Table 2. The class mappings established between the LADI v1 and v2 labels and the condensed label set for Section ??

- FEMA defines four levels of damage: affected, minor, major, and destroyed. A building is affected if damage is mostly cosmetic. A building has sustained minor damage if the damage is repairable and non-structural. A building has sustained major damage if the damage is structural or if it is significant damage that requires extensive repairs. A building is destroyed if it cannot be repaired. Does

this image contain any buildings which have sustained an "affected" level of damage or greater? Answer with one word, 'yes' or 'no.'

- FEMA defines four levels of damage: affected, minor, major, and destroyed. A building is affected if damage is mostly cosmetic. A building has sustained minor damage if the damage is repairable and non-structural. A building

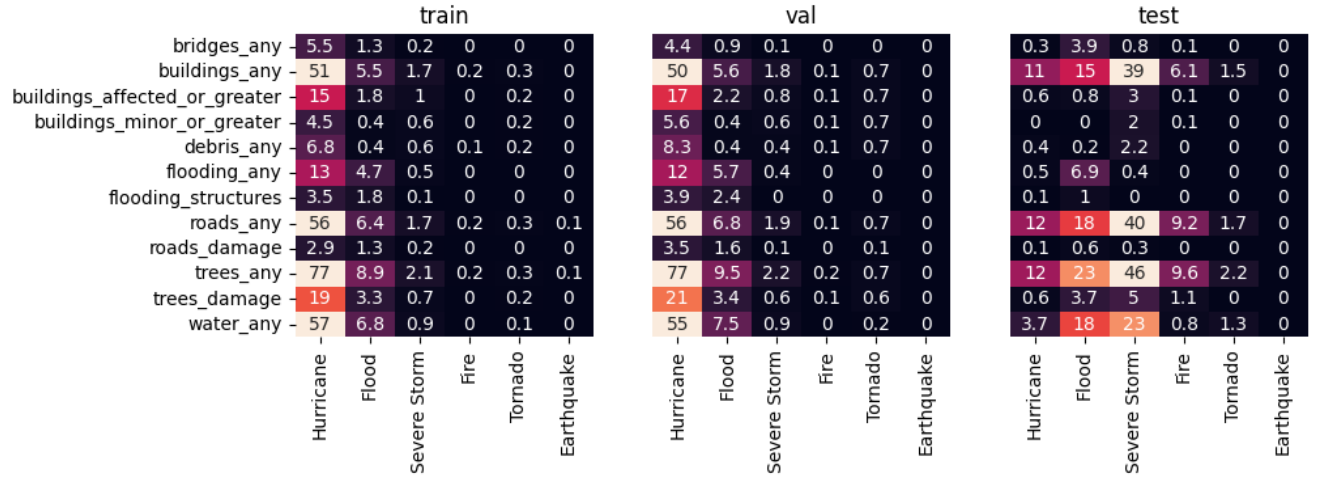
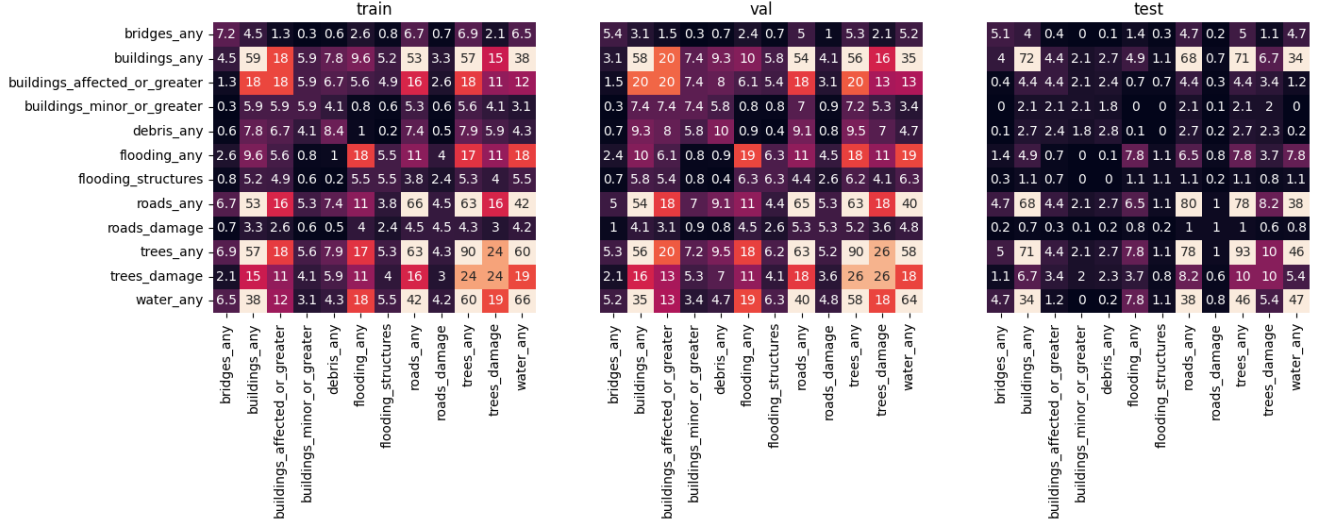


Figure 1

has sustained major damage if the damage is structural or if it is significant damage that requires extensive repairs. A building is destroyed if it cannot be repaired. Does this image contain any buildings which have sustained an "minor" level of damage or greater? Answer with one word, 'yes' or 'no.'

- Does this image contain debris? Answer with one word, 'yes' or 'no.'
- Does this image contain flooding of any structures or land? Answer with one word, 'yes' or 'no.'
- Does this image contain flooded structures? Answer with one word, 'yes' or 'no.'
- Does this image contain roads? Answer with one word, 'yes' or 'no.'
- Does this image contain damaged roads? Answer with

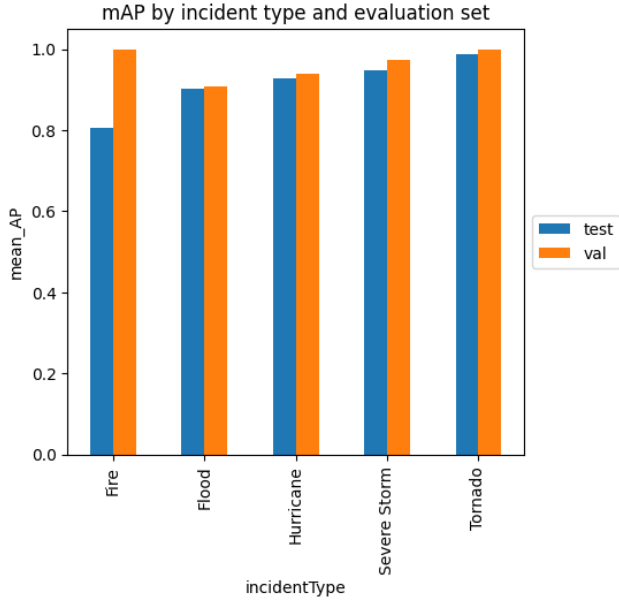
one word, 'yes' or 'no.'

- Does this image contain trees? Answer with one word, 'yes' or 'no.'
- Does this image contain damaged trees? Answer with one word, 'yes' or 'no.'
- Does this image contain water? Answer with one word, 'yes' or 'no.'

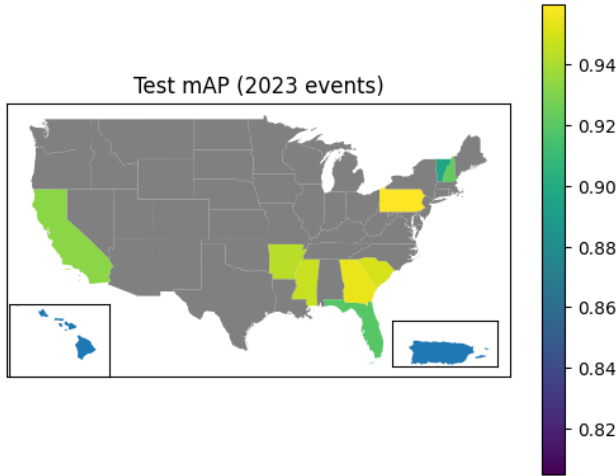
We did not encounter any difficulties with improperly formatted responses with either LLaVA-NeXT or GPT-4o

A.6.2. GPT-4o Prompt

The batch size limit in GPT-4o's batch processing was an obstacle to evaluation. The number of tokens which can be queued for processing by GPT-4o is capped, meaning that batches must be submitted in serial. The size of each



(a) mAP by event type for validation and test sets.



(b) mAP by state on test set.

Figure 2. Characterization of classifier performance by event type and location.

batch is also limited, meaning that our test set had to be split into more than 50 batches. As batch processing is not guaranteed to start immediately on submission, evaluating the set took multiple days.

To minimize the number of requests, we gave GPT-4o a prompt asking it to emit a structured output answering each question by filling in a JSON object:

Answer questions about this image by setting the 'answer' values in the following JSON data structure to boolean values. The data structure should NOT contain any 'null' values when you are done. Respond with ONLY the completed data structure: {"bridges_any":

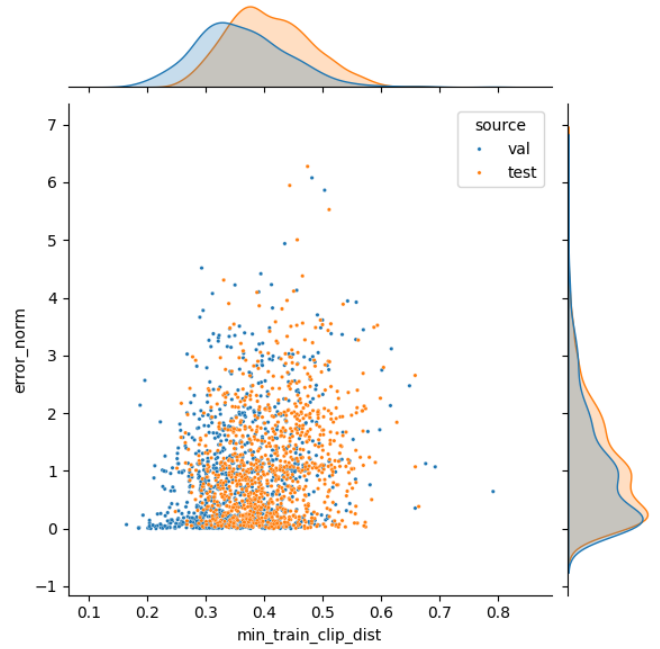


Figure 3. Error vector L^1 norm vs. the distance from a point in an evaluation set to its nearest neighbor in the train set in CLIP space. Validation data is plotted in blue and test data in orange.

{"question": "Does this image contain bridges? Answer with a boolean true/false.", "answer": null}, {"buildings_any": {"question": "Does this image contain buildings? Answer with a boolean true/false.", "answer": null}, {"buildings_affected_or_greater": {"question": "FEMA defines four levels of damage: affected, minor, major, and destroyed. A building is affected if damage is mostly cosmetic. A building has sustained minor damage if the damage is repairable and non-structural. A building has sustained major damage if the damage is structural or if it is significant damage that requires extensive repairs. A building is destroyed if it cannot be repaired. Does this image contain any buildings which have sustained an 'affected' level of damage or greater? Answer with a boolean true/false.", "answer": null}, {"buildings_minor_or_greater": {"question": "FEMA defines four levels of damage: affected, minor, major, and destroyed. A building is affected if damage is mostly cosmetic. A building has sustained minor damage if the damage is repairable and non-structural. A building has sustained major damage if the damage is structural or if it is significant damage that requires extensive repairs. A building is destroyed if it cannot be repaired. Does this image contain any buildings which have sustained an 'minor' level of damage or greater? Answer with a boolean true/false.", "answer": null}, {"debris_any": {"question": "Does this image contain debris? Answer with a boolean true/false.", "answer": null}, {"flooding_any": {"question": "Does this image con-

tain flooding of any structures or land? Answer with a boolean true/false.”, “answer”: null}, “flooding_structures”: {”question”: “Does this image contain flooded structures? Answer with a boolean true/false.”, “answer”: null}, “roads_any”: {”question”: “Does this image contain roads? Answer with a boolean true/false.”, “answer”: null}, “roads_damage”: {”question”: “Does this image contain damaged roads? Answer with a boolean true/false.”, “answer”: null}, “trees_any”: {”question”: “Does this image contain trees? Answer with a boolean true/false.”, “answer”: null}, “trees_damage”: {”question”: “Does this image contain damaged trees? Answer with a boolean true/false.”, “answer”: null}, “water_any”: {”question”: “Does this image contain water? Answer with a boolean true/false.”, “answer”: null}} Your entire response should be a JSON object, so it will start with ‘{’ and end with ‘}’

A.7. Environmental Impacts

The models were trained on the TX-GAIA supercomputer [1] using single nodes with two NVIDIA Tesla V100 GPUs. An estimated 1000 kWh of energy was used to train the models, including runs for architecture search, pretraining, and hyperparameter optimization. Using the estimated Carbon Use Efficiency (CUE) of the system’s host facility—the Massachusetts Green High Performance Computing Center (MGHPCC) [3]— of 0.03 kg CO₂/kWh, this corresponds to an emissions of ~ 30 kg CO₂ in the training of the models. Similarly, using the Water Usage Efficiency of MGH-PCC [3] of ~ 1.7 L/kWh, we estimate ~ 1700 L of water used.

References

- [1] TX-GAIA (Green AI Accelerator). Available: <https://www.top500.org/system/179603/>. 5
- [2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1
- [3] Prateek Sharma, Patrick Pegus II, David Irwin, Prashant Shenoy, John Goodhue, and James Culbert. Design and Operational Analysis of a Green Data Center. *IEEE Internet Computing*, PP(99):1–1, 2017. 5