Google Research

Agile Modeling: From Concept To Classifier in Minutes

Otilia Stretcu*, Edward Vendrow*, Kenji Hata*, Krishnamurthy Viswanathan, Vittorio Ferrari, Sasan Tavakkol, Wenlei Zhou, Aditya Avinash, Enming Luo, Neil Gordon Alldrin, MohammadHossein Bateni, Gabriel Berger, Andrew Bunner, Chun-Ta Lu, Javier A Rey, Giulia DeSalvo, Ranjay Krishna, Ariel Fuxman

Email: {otiliastr, kenjihata, afuxman}@google.com Github: https://tinyurl.com/52xfhmpw

Introduction

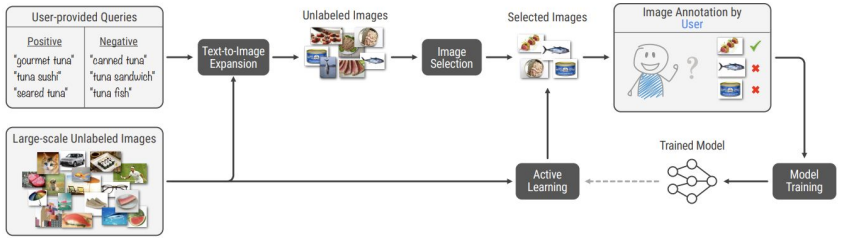
Agile Modeling Problem

- Today, crowd workers label the majority of ML data.
- However, there exist subjective concepts where decisions may be difficult for the crowd to emulate.
- This highlights the need for user-centric approaches to develop real-world classifiers for these concepts.
- **Agile Modeling** introduces the process of turning any visual concept into a computer vision model through a realtime user-in-the-loop process.

Our contributions

- Formulation of the Agile Modeling problem.
- Real-time prototype built upon image-text co-embeddings that trains models better than zero-shot in 5 minutes.
- We compare models trained with labels from real users versus crowd raters. We find that the value of a user increases when the concept is nuanced or difficult.
- Verify the results of the user study with a simulated experiment of 100 more concepts in ImageNet-21k.

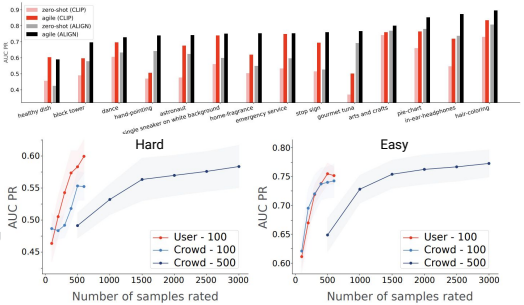
Prototype System



Results with Real Users

Setup

- 14 volunteer participants were each assigned a different concept.
- Each participant used our Agile Modeling system, starting with no labels and an unlabeled pool of 100 million images drawn from LAION 400M.
- We compare accuracy after 6 rounds of 100 image labeling as compared to zero-shot accuracy.
- We compare "hard" versus "easy" concepts by dividing the 14 based on their zero-shot performance.



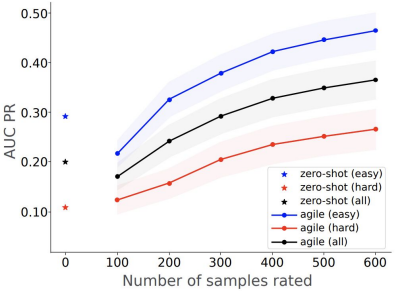
Results

- **In just 5 minutes of work, the resulting model outperforms zero-shot by 5% AUC-PR and continues to improve with each labeling round.**
- In an average of 24 minutes, users are able to train high accuracy models beating zero-shot by 16% AUC-PR.
- By using pre-extracted image-text embedding and small models, we reduce the active learning image selection step to just 1 minute and model training to 23 seconds.
- On "hard" concepts, models trained with users (User-100) outperform models trained with crowd raters, even when 5x more ratings are obtained from the crowd (Crowd-500).

Results with ImageNet-21k

Setup

- Fifty "easy" concepts are randomly selected from ImageNet-1k.
- Fifty "hard" concepts were randomly chosen from 546 classes with the following criteria:
  - 2-20 hyponyms, to ensure visual variety.
  - More than one lemma, to ensure ambiguity.
  - Not an animal or plant, which have objective descriptions.
- Apply the Agile Modeling framework with the ImageNet-21k training set as the unlabeled data pool, and the validation set for evaluation.
- Ground-truth class labels simulate a user providing ratings.
- We use the class name and its corresponding WordNet description as positive text phrases in the text-to-image expansion step.



Results

- We see a similar trend to our user experiments, with significant improvements over zero-shot baselines as well as continued improvement with each active learning round.
- The zero-shot baseline differed significantly between the "easy" and "hard" concepts with scores of 0.29 and 0.11, respectively.
- The equivalent of 30 minutes of human work yields a 20% boost in AUC PR over the zero-shot baseline.