Milestone 2 Team Data Pets Documentation

Ivan Shu, Xiang Bai, Yuxin Xu, Sean Gao

Background

Austin Pets Alive (APA) is an association of pet owners with a database of available pet photos. The goal will be to build a reusable application, design, and framework that can be implemented in any animal welfare nonprofits to connect future pet owners with pets. The outcome of the project will be to build a full featured application for the APA.

Problem Definition and Project Scope

For this Pets project, our group's focus will be to create a user friendly tool to match potential dog loving adopters/owners to dogs available for adoption. As stated above, the end goal is developing a nice web application that can harness the data and be used with features/deliverables that will be helpful for the matching process.

The core problem we are trying to solve is to help future dog owners find a dog who is a good fit for their lifestyle and family environment. First we help the user search for dogs based on certain features such as size, color, and breed or the images users upload. Secondly we will connect the dog with the user by allowing the user to chat with a persona of the dog. The user can ask this virtual dog any question about it -- its breed characteristics, or any general questions about puppies and dogs.

Data Science Pipeline Logistics

All the data is stored in google cloud platform, and models will be run and trained in Colab. We plan to containerize both front-end and back-end applications using docker and deploy the app using Kubernetes (more details are to be added in milestone 3). To accelerate training process, we resized the images, created TFRecords, generated predicted embeddings and stored them in our GCP bucket.

Exploration Data Analysis

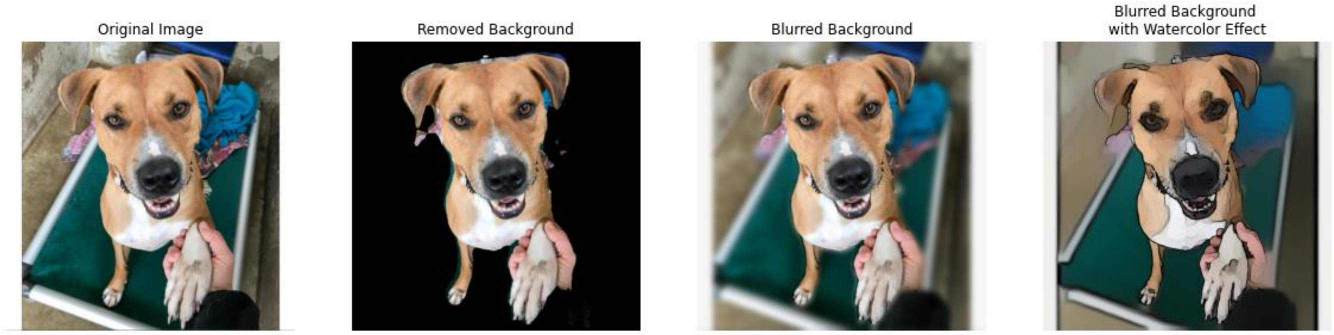
APA makes available a repository for its animals that is roughly ~17k dog records and close to 40k if you include cats. For the ~40k pets there looks to be ~140k photos. Our dataset consists of several csv files [Dataset Link], including dogs metadata, dogs images, website memos, as well as dogs questions and answer datasets.

Proposed Solution I: Computer Vision

To help the matching process, we leverage several computer vision models to accomplish a few specific tasks. As an adoption center, when adopting new dogs from users, a function will be available for some pre-processing of the images.

- Remove noisy background from the uploaded dog picture using DeepLabv3 plus.
- Allow to choose and add new backgrounds/effects.
- Enhance the image if the solution of the uploaded picture is not ideal.

1. Removing Old and Add New Background with Different Effects



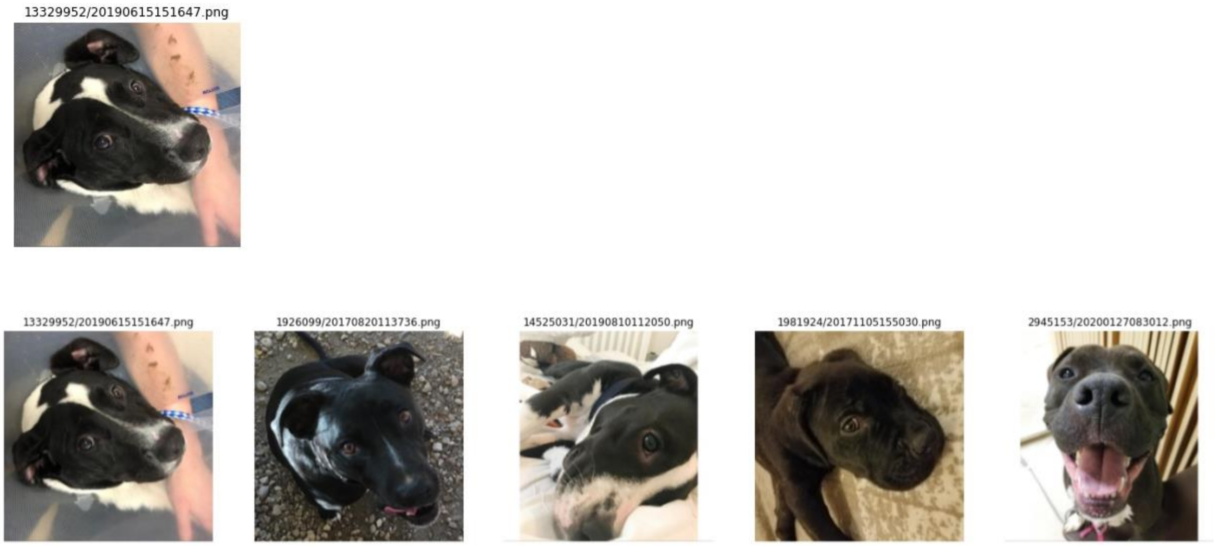
One the very left shows the original image, and the second image was after the background was removed by DeepLabv3 Plus. The third and forth images are different backgrounds users or adoption sites can choose when donating or adopting new dogs.

On the other hand, to allow for users searching for their interested dogs, some other tasks listed below need to be considered.

- Create embeddings using EfficientNet for all the dog images in the dataset.
- Search top 5 similar images using Facebook AI Similarity Search (FAISS).
- Segment and remove background before embedding search by using DeepLabv3 plus.

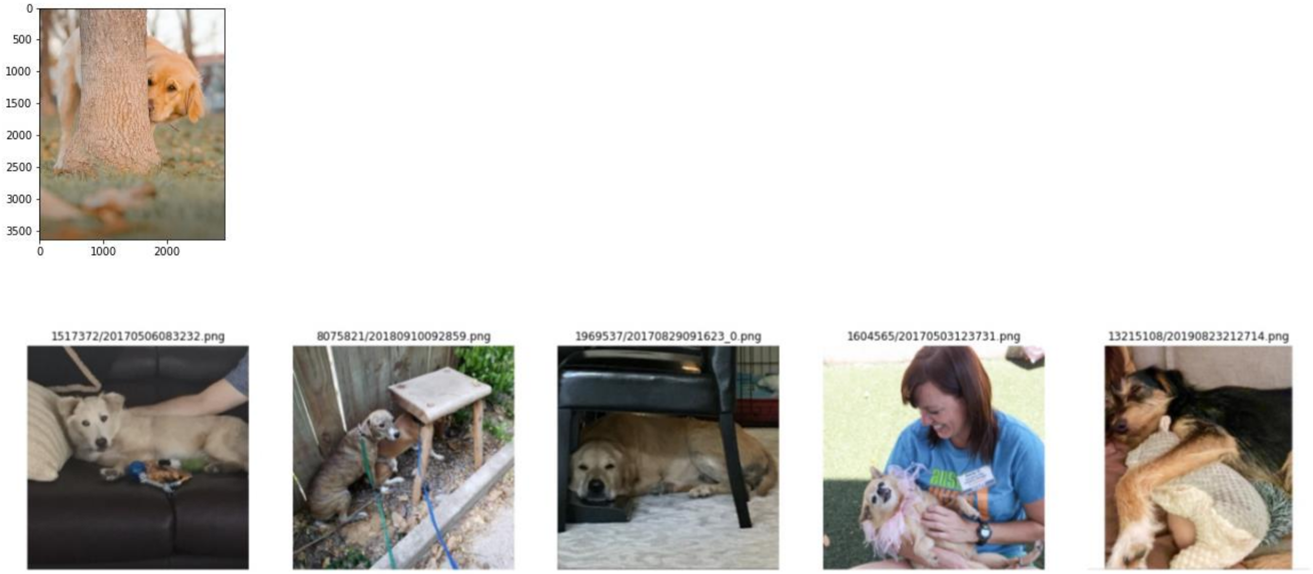
One way to search similar images given an input is to create image embeddings and look for similar ones in that space. To generate embeddings, we used EfficientNet, B0 model [1], as shown to have better performance with less parameters. Then we used an embedding search algorithm developed by Facebook, FAISS to find the top 5 similar images [2]. See figure 2 for some results.

Figure 2. Example Matched Images By Using EffecientNet and FAISS Embedding Search


13329952/20190615151647.png
13329952/20190615151647.png
1926099/20170820113736.png
14525031/20190810112050.png
1981924/20171105155030.png
2945153/20200127083012.png

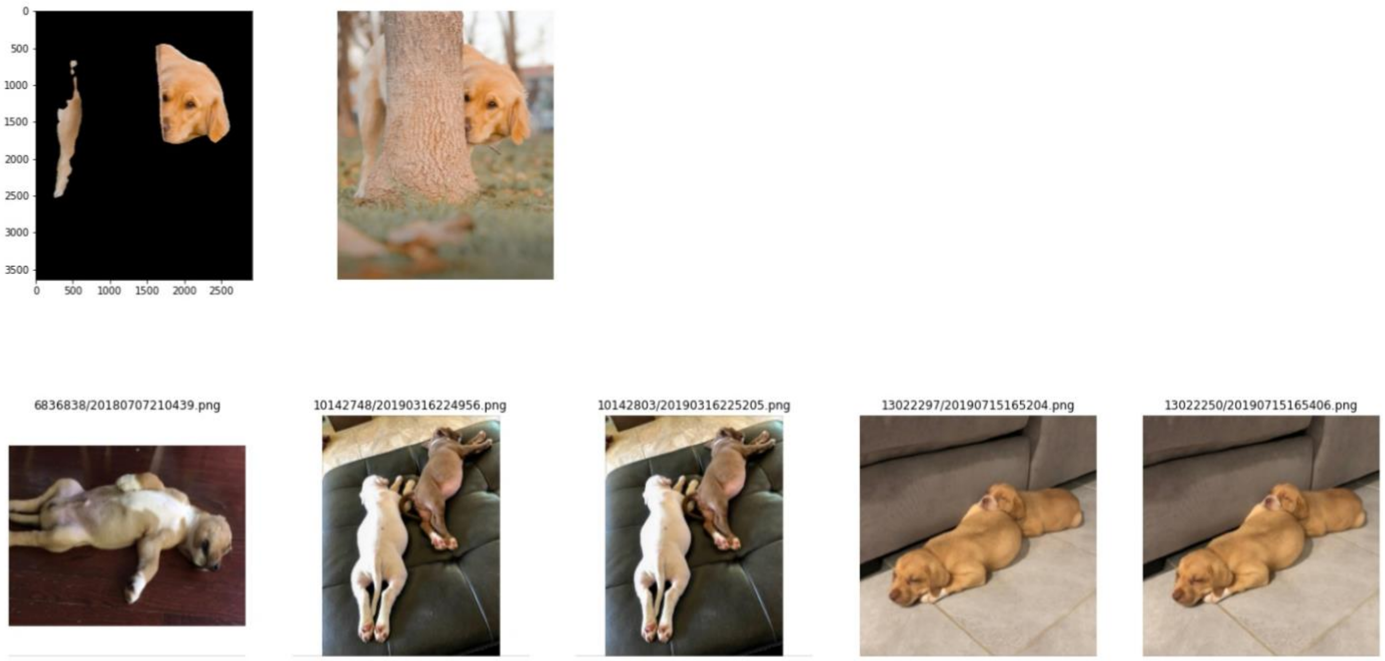
The top row shows the input image and the second row shows the top 5 matched images using FAISS. However, one problem we encountered is that when generating embedding for an input image that has noises or features not related to dogs, the matched images contain irrelevant features as well. See figure 3. Note: the main steps of model training were referenced by Rashmi's notebook during 2021 ComputeFest.

Figure 3. Example Input Images That Contains Dog-Irrelevant Features


0
500
1000
1500
2000
2500
3000
3500
0
1000
2000
1517372/20170506083232.png
8075821/20180910092859.png
1969537/20170829091623_0.png
1604565/20170503123731.png
13215108/20190823212714.png

Here, we can see that when the inputted image has a large dog-irrelevant feature (tree), the matched images contain dog-irrelevant features as well. For example, the matched images contain arm, chairs or human. This causes the searchd pictures to contain less information about the dog, which led to unideal results. To bypass this constraint, we introduced another pre-trained model, DeepLabv3 plus to pre-segment the image before embedding search [3].

Figure 4. Pre-segment Input Images Before Embedding Search Similar Images

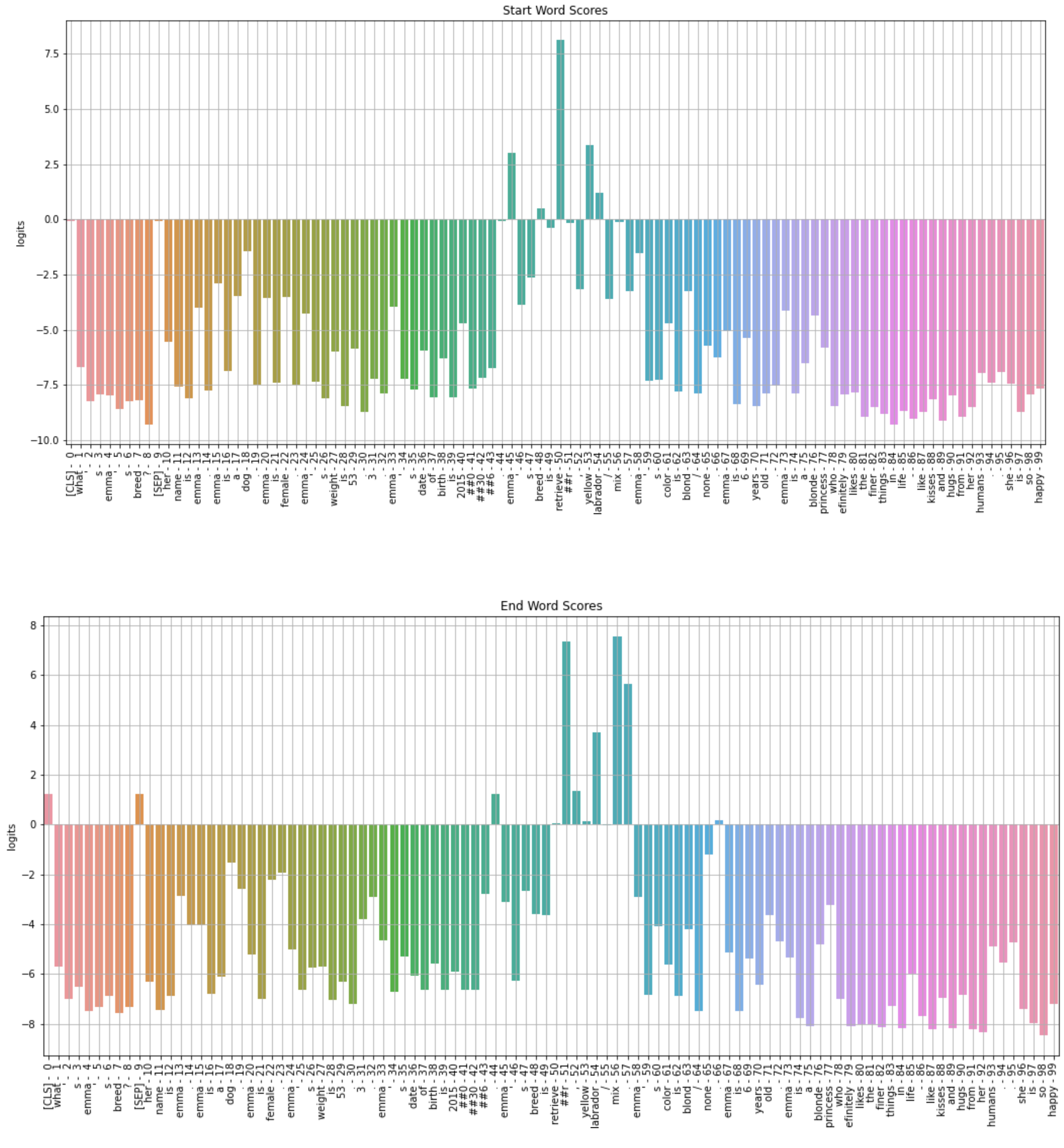

0
500
1000
1500
2000
2500
3000
3500
0
500
1000
1500
2000
2500
6836838/20180707210439.png
10142748/20190316224956.png
10142803/20190316225205.png
13022297/20190715165204.png
13022250/20190715165406.png

After the image was segmented and the background was removed, the matched images contained more features or information related to dogs, such as age, or fur color. One caveat is that the matched images might contain multiple dogs because of the segmentation.

Proposed Solution II: Natural Language Processing

The next task is to create a persona of the dog that the user has searched and clicked and allow users to ask more information regarding the dog by directly chatting with the persona. To implement a baseline model, we used BERT question and answering model, which was trained on the Stanford Question Answering Dataset (SQuAD) [4]. We will provide sample questions and reference text to the model and let the model predict the start and end token of a "span" of text which acts as the output answer. We followed the main steps from Shivas' notebook in 2021 ComputeFest [5] and the tutorial by Chris McCormmick [6].

Figure 5. Probability scores for start and end tokens predicted by BERT with an example question



In one example, we asked the question, “what’s Emma’s breed” and fed it together with a reference text. The model correctly predicted with the correct span of text in the reference, which is "retriever , yellow labrador / mix". To further visualize the probabilities of each start and end tokens, we plotted the results in figure 4. Both “retriever” and “mix” are predicted with highest probabilities for either start and end tokens.

Limitation and Next Step

One limitation that caught our attention is that it takes around 8 to 12 seconds for DeepLab to generate and predict image embedding. This could be a huge hindrance to user experiences as multiple requests are made. To account for this, we are looking for a more lightweight model to segment images. Potential models under consideration are PointRend(Kirillov et al., 2020), BiSeNetV2(Yu et al., 2020).

As for language models, we noticed by only using BERT, it can't provide answers outside its reference text, meaning not being able to generate language. This might pose some problems that when users interact with the persona, they might not have a natural feeling of having an actual conversation. With this, on the next step, we want to switch to a GPT-based model, which can generate new conversation text. In the meantime, we are also searching for another pre-trained language generating model that is trained on a broader conversation dataset, but lightweight. Currently we only found DistilGPT-2. We'll keep searching for more options.

Last but not least, we are planning on diving into app design for the next few weeks. Specifically, use Node JS for the front-end and Oracle for the back-end database. Then use Docker to containerize both applications and deploy it to the GCP using Kubernetes.

Timeline for the remaining project

- Week 10/26 - 11/8:
  - Wrap up Modeling (Computer Vision + Natural Language Processing)
  - Research on finding lightweight pre-trained models
- Week 11/9 - 11/20
  - Focus on UI and App Design
- Week 11/21 - 11/31
  - Buffer Time and Project Wrap Up