WeRateDogs Image And Tweet Analysis

I am trying to answer the following questions when I analyze this data

1. Would rating, retweet and likes (favorites) change according to the age of the dogs in the tweets?
2. How does the image classification performance for tweets with images change, with respect to the source of the tweet?
3. Which class labels are confused most for misclassified images?
4. Are there a relationship, or correlation, between favorite and retweet counts of tweets? Do people like the tweets they retweet, especially for WeRateDogs?

I will communicate my findings in four sections of this report.

Would rating, retweet and likes (favorites) change according to the age of the dogs in the tweets?

I computed the following statistics in my analysis.

	retweet_count		favorite_count		rating_numerator		rating_denominator	
	mean	median	mean	median	mean	median	mean	median
age								
doggo	6466.433735	2961.0	17820.228916	11519.0	11.771084	12.0	10.0	10.0
pupper	2557.550218	1270.0	7474.397380	3428.0	10.820961	11.0	10.0	10.0
puppo	7124.875000	3230.0	22715.125000	16057.0	12.041667	12.0	10.0	10.0

The distribution of retweets and favorite counts are skewed for all age groups. The distribution of the ratings on the other hand are not skewed for all age groups. Looks like middle aged dogs are most popular. They have the highest rating on average, retweeted by most people, and their pictures are liked most. Based on the same metrics, young dogs are the least favorite, which is interesting in my point of view. Most people would expect the opposite.

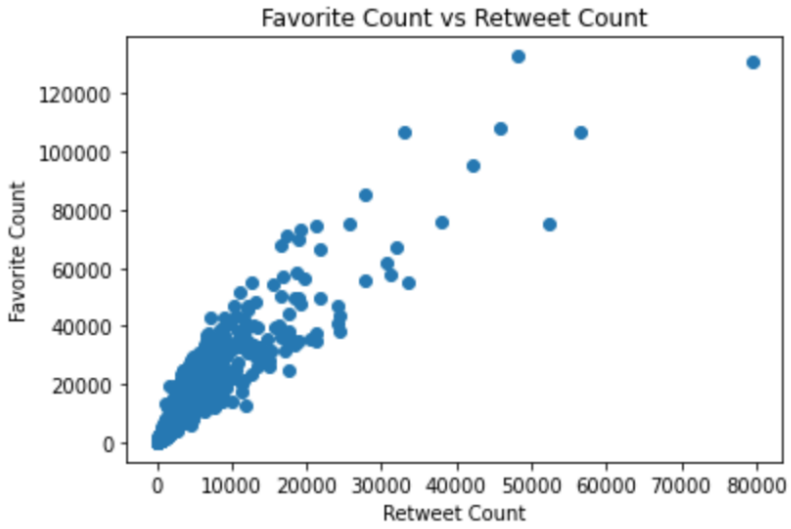
How does the image classification performance for tweets with images change, with respect to the source of the tweet?

Across all tweets from different sources, our image classifier is pretty confident with respect to the first predicted class. However, there are pretty a lot of predictions in all three top classes where the predictions are not actually dogs.

Which class labels are confused most for misclassified images?

If we focus particularly on the tweets sourced from the iPhone (as this is the most popular source of tweets), the misclassified images can be analyzed for the top 1 most confident class. The top 3 most confused classes are seat_belt, teddy and web_site. This is something to tackle further by analyzing the actual images of dogs who are labeled like this. It is interesting to observe this.

Are there a relationship, or correlation, between favorite and retweet counts of tweets? Do people like the tweets they retweet, especially for WeRateDogs?



We can see this relationship through a plot. Looks like there is a positive linear correlation between the retweet and favorite counts for WeRateDogs tweets. This suggests that when people retweet a dog picture, they also actually like the picture itself. We can also predict one variable from the other using linear regression.