Label Diversity Flagging:

The Algo:

Calculate the average number of label types in a project and if a task has a below average number of label types then flag it for review.

Short-Term Improvements:

This system has a lot of room to grow to include new features and become more refined. For instance, the limit right now is the average, but we can modify this limit to be many different values (i.e. median, a std off of the norm, variance, etc.) that will affect the frequency of flags and the accuracy of the data. I'd imagine a better limit could be made using data Scale collects regarding accuracy. It could also be of value to assume if a task has too many label types it could be incorrect.

There's potential for excessive or too little flagging in the event that the dataset being input is extremely diverse - for instance, if a lot of rural street images are alongside a lot of urban street images. For this project specifically, if there were a lot of rural images (little sign variety) alongside a lot of urban images (high sign variety) then this flag may become less effective in a case with a large variance in number of label types by task.

Long-Term Improvements:

It would be quite cool to implement a basic deep learning algo that can determine if an image is 'rural','suburban' or 'urban' before this flag is raised in order to establish three unique limits - one for each type of environment. This description assumes the use of traffic sign data but the principle of environment dependent limits seems useful. It is reasonable to assume that the average number of label types in a city would be higher than on a highway through farmland. Having multiple, unique limits would make this a more effective tool. There would be trade offs here in terms of performance but depending on the size of the dataset it could be useful. A shorter-term fix for this problem would be to allow users/customers to indicate if a dataset is urban/suburban/rural before labeling or have our labelers include this as a feature in their labeling. This could then be used as training data for a deep learning algo.