Modeling Noise in Recall Related to Perception of Categories

Meena Rajan, Srilaya Bhavaraju, Priya Pillai

Abstract

In this work, we attempt to model how humans recall information as it related to their perceptual understanding of categories. We first create a dataset of cat and dog face images, where each face is constructed with particular feature parameters. We then ask subjects to classify a fixed set of images as either cat or dog. Following this, we fit two separate multivariate Gaussian models to both the cat and dog feature vectors. Finally, we model two separate Markov chains for each animal, where the first subject in both chains is shown a face with features that are ambiguous for a fixed period of time. The subject in one chain is told the face is a cat, whereas the subject in the other chain is told the face is a dog. The subject is then asked to recreate the face and the resultant image is passed to the next subject in the chain. We present a Bayesian model to understand the noise present in the recollection of the face at each step of the chain. Overall, we were able to experimentally determine how perception of categories affects noise in recall.

Introduction

Categorical effects in cognition and perception are present and similar across various domains. These effects can be observed in domains from speech sounds and colors, to faces and even learning artificial categories. In each domain, the results of categorical effects have been found to be qualitatively similar, having enhanced between-category discriminability and reduced within-category discriminability [1]. It it often thought that biases contribute to the perceptual patterns that account for categorical effects and that these patterns can be influenced by learned categories as well (such as implicit categories that arise from specific distributions of the shown examples). However, the reasons and mechanisms behind the connection between categories and perception are not well understood [1]. In this work, we attempt to model noise in recall as it relates to the perception of categories composed of multi-dimensional features when we introduce bias into the recall.

More specifically, this work considers perception related to a dataset of two categories we created ourselves. The dataset consists of dog and cat faces, each of which is created using a feature vector of size 13 (the creation of the dataset is explained in more detail under ”Dataset Creation”). We first attempt to obtain a ”ground truth” of what the ideal feature parameters that make up a cat and dog are, respectively. In order to do this, the first part of our experiment consisted of generating images with random feature parameters and asking participants to classify each image as a cat or dog. Using these classifications, we fit two separate multivariate Gaussian distributions, one for each category of cat and dog. We obtain a posterior mean representing the ”ideal” feature values that define a particular category from these distributions. We also obtain probability density functions for each category that we can sample from for later parts of the experiment.

Next, we select an image that was found to be relatively ambiguous during the experiment (seemed to be equally likely to be classified as cat or dog). Using this image, we simulate two separate Markov chains, each consisting of five different people. For one chain, we show the first subject the ambiguous image for a fixed period of time and tell them it is a cat. We ask the subject to recreate the image and record the resultant feature values. We use the resultant image as the first image shown to the next participant in the chain and repeat the process. We repeat the same for the other chain except we tell each participant that the image is a dog instead.

We model the noise in each subject's recall as a Bayesian model. The noise is modeled taking into account recreation of the image and the differences in the feature values. We also evaluate the results of the chain, expecting the feature values of each chain to converge to the mean values found from the fitted multivariate Gaussian distributions.

Interestingly, we found that the chains converged to the extremes of our distributions rather than the means. However, we also found that the noise in recall can be meaningfully modeled using Bayesian model. More specifically, we found that the amount of noise in different features can reveal information about which features are more important and memorable for categorical perception.

The code for this work can be found here: `https://github.com/priyappillai/stress_interp`

Related Work

The influence of categories on perception is well known in domains ranging from speech sounds to artificial categories of objects. [2] describes the categorical perception of speech sounds, noting that between-category discrimination for listeners of stop consonant sounds is nearly perfect. Similar patterns have been observed in the perception of colors [3], facial expressions [4], familiar faces [5], and the representation of objects belonging to artificial categories learned over the course of an experiment [6].

In several studies, it has been found that novel categories can form throughout the course of an experiment and that these categories can affect perception during the experiment [1]. In [2], Liberman suggest that this learning component can take two forms: 1) Acquired distinctiveness involving enhanced between-category discriminability, 2) Acquired equivalence involving reduced within-category discriminability. In this work, we explore the categorical perception that can occur through the course of an experiment and how it relates to between-category discriminability.

These phenomenon have been supported by studies on categorization training in color perception and auditory perception of white noise [7]. The results of other studies suggest that categorizing stimuli along two dimensions can lead to acquired distinctiveness [6], whereas similarity ratings for drawings that differ along several dimensions have shown acquired equivalence in response to categorization training. It is thought that such effects involve changes in the underlying stimulus representations [9]. In this work, we explore multi-dimensional features are they relate to distinguishing between two different categories.

Furthermore, several studies have demonstrated that categories for experimental stimuli are learned quickly during an experiment without explicit training [1]. In [10], learned categories of subjects for unfamiliar face continua seemed to correspond to the endpoints of the continuum. Additionally, implicit categories have been used to explain why subjects often bias their perception toward the mean value of set of stimuli in an experiment [1]. In [11], it is argued that subjects form an implicit category that includes the range of stimuli they have seen over the course of an experiment and that they use this implicit category to correct for memory uncertainty when asked to reproduce a stimulus. Under their assumptions, the optimal way to correct for memory uncertainty using this implicit category is to bias all responses toward the mean value of the category, which in this case is the mean value of the set of stimuli [1]. [1] presents a Bayesian analysis in the context of speech perception, using a similar structure and approach to previous work accounting for bias in visual stimulus reproduction. We present a Bayesian analysis representing noise in recall in the context of learned categories. The categories are learned because we ourselves create the cat and dog images based on a fixed set features (rather than using real images of cats and dogs), but participants may also draw on their learned outside knowledge of cats and dogs during the experiment.

Dataset Creation and Experimental Setup

To create the dataset, initial work was done to draw and analyze a number of different references of cartoon cats and dogs. Once this work was done, we hand-drew variation of across a number of features to determine which features produced the most salient changes amongst people. We piloted these hand-drawn versions on ourselves and a small number of colleagues to distinguish visually at what points these features became a ”cat” and at what point those features became a ”dog”, as well as which features were most important for us to model. We decided to pick features that were specifically Boolean rather than continuous so that we could model the distribution as a multivariate Gaussian.

We split up the face into 8 component parts-the outer edge of the face, the ears, the eyes, the nose, the snout, the mouth, the whiskers, and the fur. The outer edge of the face was simplified to an ellipse rather than a more complicated or defined shape as this would allow more understandable modification during the Markov Chain process. We considered including tufts of hair on the cheeks, chin, and top of the head as separate variables (as we noticed difference in responses given changes in this feature specifically), but decided to exclude these due to the additional difficulty of the code were we to include them as anything other than Boolean variables. This parameter was listed as ”Face Aspect Ratio”. From initial cursory analysis, we expected taller faces to correspond with dogs and inversely wider faces to correspond with cats.

Ears were a particularly difficult feature to model, as an extremely common method of distinguishing cat versus dog is whether the ears point down or up, a Boolean-like variable. We accounted for this by changing the angle at which the ear appear to come from the head, visually implying the ears were attached at the back of the head. This parameter was listed as ”Ear Angle”. Other parameters we found may affect judgement of the ears were roundness of the ears (”Ear Point”, modeled by increasing radius of a circle at the ear tip), length of the ears (”Ear Length”), ear width (”Ear Tip Angle”, as it was modeled by the angle at the tip), and the specific orientation of the ear relative to the horizontal and the center (”Ear Orientation”, model by a parameter that would rotate the ear). We expect ears pointed up, pointy, short, and skinny to correspond to cats and the opposite for dogs. Ear orientation and ear angle both combine to define how much the ears point up.

The eyes were modeled by plain black circles rather than the more biologically accurate sclera, iris, and pupil as it was a simplification we found frequently in cartoon images and as we needed to limit the number of tested variables. The variables we decided to incorporate were the distance between the eyes (”Eye Distance”), the aspect ratio of the eyes ("Eye Aspect Ratio"), and eye size ("Eye Height", as eye size is a combination of eye height and eye aspect ratio). We expected skinny, large, close together eyes to correspond to dogs, and the opposite for cats.

The nose, snout, and mouth were all greatly simplified for the purposes of this experiment. The nose was modeled as a triangle pointing down, and we allowed its size to be changed with one variable ("Nose Size"). We initially included a second ellipse on the face around the nose and mouth to indicate a snout, but found that this frequently corresponded with identifying the animal as a dog, and could not determine a method of making this a continuous variable. The mouth was modeled as a line down from the nose, plus two arcs coming from that line. These were sized according to the nose size to reduce the number of modifiable variables. We expected larger nose sizes to correspond to dogs.

The whiskers were modeled as a set of three lines coming from each side of the snout. As we found most cartoon images of dogs included no whiskers or solely dots, we considered including a parameter for number of whiskers, but decided against it as we could not make it a Boolean variable. We instead simply focused on the length of the whiskers ("Whisker Length"). We considered also including a parameter for the line weight of the whiskers, but found that there was not enough resolution in the drawing to provide a near continuous variety of line weights. We expect cats to correspond with long whiskers and dogs to correspond to short whiskers.

The fur was modeled as a single color throughout the entire image. While we did consider including various fur patterns, as we found that fur pattern could have a strong effect on perception of cat versus dog, we were unable to find a method of depicted these that would be continuous. For the fur color, to limit the colors to colors that would realistically be fur colors, we chose a specific orange colored hue, then allowed for variation in the lightness and saturation ("Fur Saturation" and "Fur Lightness"). This provided a reasonable variety of colors including black, white, grey, brown, orange, and tan, under a continuous distribution. We generally expect dogs to be labeled with darker fur colors and cats with lighter fur colors, and cats to be labeled with more orange colors than dogs, but we are very uncertain that these variables will actually form a normal distribution.

An interesting point we noted in pilots is that the variable parameters that most commonly matched "cat" versus "dog" were far more extreme in the cartoon representations than they are in real life images of cats and dogs. For example, while dogs do have whiskers in real life, short whiskers are associated with dogs. Similarly, cats ears are rarely as pointy as the assumed mean of roundness of cat ears would imply. Eye shape was also particularly different-while both cats and dogs have round eyes in real life, cartoon images with taller eyes tended to imply dog more frequently. This is particularly interesting as cat eyes are associated with having tall pupils.

One difficulty in generating these images was that, given the feature set we determined, the conceptual path between "cat" and "dog" is non-linear. Frequently, we found labels such as "otter," "bear," or "seal" being applied to the images we showed individuals. Additionally, while we modeled each variable for "cat or "dog" as having one mean per variable on a multivariate Gaussian distribution, it is highly likely that there are multiple peaks, dependent on the specific breeds of dogs and cats people consider. For example, while most dogs are considered to have ears that point down, if other features match a breed that matches with ears that point up (such as German shepherds, corgis, or pomeranians), the face may be more classified as a dog regardless of the ear feature implying a cat. We hypothesize that this breed specific effect would be stronger with fur patterns included.

To get a set of random images, we individually found reasonable maximums and minimums for each parameter, then drew each feature from a uniform random distribution along each feature. To reduce complexity, we discretized variables to be rounded to the nearest . We then generated a set of 50 random images per participant, and had them label each image as either cat or dog. We had 10 participants, giving us 500 labelled images. After this, we tested their recall given labels by showing them an image labeled ”cat” or ”dog”, taking it away, and having them redraw the image from memory.

Modeling Features as Multivariate Gaussian Distributions

Once we obtain the classifications of images with randomly generated feature parameters, we fit multivariate Gaussian distributions to each class. We separate all feature vectors that were classified as cat `cat-data` from those classified as dog `dog-data`. We create multivariate Gaussian distributions for each category, where the mean and covariance for the cat distribution is taken as the mean and covariance of `cat-data` and the mean of covariance of the dog distribution is taken as the mean and covariance of `dog-data`. Next, we sample approximately $50,000$ samples from each distribution and use these samples to obtain a posterior mean estimate for the ”true” mean of each distribution using `pymc3`. The posterior mean is estimated using the samples as observed values.

The cat face obtained from the posterior mean is shown below.

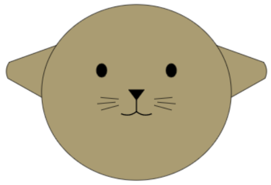

Figure 1: The posterior mean of the Multivariate Gaussian of a cat face

The dog face obtained from the posterior mean is shown below.

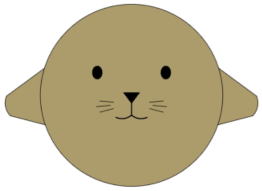

Figure 2: The posterior mean of the Multivariate Gaussian of a dog face

While there are notable differences between the cat face obtained from the posterior mean and the dog face obtained from the posterior mean, one may expect more significant differences. Noticeable differences include the angle of the ear, which is much lower for a dog than that of a cat. Furthermore, the whisker length for the dog face is slightly less than that for the cat. Furthermore, the fur for dog image is a bit darker and more saturated than that for the cat image.

The lack of stark differences between the two images may be attributed to the feature parameters of the random images generated. Because the parameters were generated completely randomly, most of the images generated may have had around the same parameter values for fur-lightness and fur-saturation. This can possibly be controlled for in future experiments by requiring a certain percentage of randomly generated images to have values for fur-lightness and fur-saturation to be within particular ranges. Furthermore, many of the features are restricted to values within a relatively small range. Therefore, the variance may have been relatively small and differences in value for those features may not have been immediately evident.

Additionally, we assumed all features to be independent from one another. However, it may be the case that certain feature values should depend on each other. For example, having correlations between the values for nose size and whisker length may be more representative of cat and dog faces in the real world. This may have resulted in participants having more definitive classifications of certain images rather than hesitating before classification.

However, the features that are noticeably different between the two images appear to make sense in terms of the features humans may be more likely to attribute to either category. For example, the ear angle for the cat image is is higher than that for the dog image. This indicates that humans tend to associate higher, upright ears with cats and lower, hanging ears with dogs. We may expect that humans are more likely to observe dogs with low, hanging ears as dogs than cats over their lifetimes. Similarly, they have probably observed more cats with high, upright ears than dogs with similar ears. Furthermore, observations of cats and dogs over the subjects' lifetime experiences may present more evidence for cats having longer whiskers than dogs. Additionally, the subjects' overall may have observed more dogs with dark fur than cats.

We also analyzed the correlation coefficients between the feature variables from the collected data to determine any strong correlations. For the dog data, there were no particularly strong correlations found between any of the feature variables. The largest correlation coefficient value found was 0.125 between the feature variables ear-point and eye-height. For the cat data, there were were also no particularly strong correlations found between any of the feature variables. The largest correlation coefficient value found was 0.151, which was between the feature nose-size and whisker-length. Given our own observations and prior knowledge, we think that nose-size and whisker-length may have a reasonable correlation. Given more data, stronger correlations between the feature variables may have been found.

Bayesian Modeling of Noise in Recall

Now that we fit a multivariate Gaussian to our cat and dog data, we wanted to do some interesting analysis about how this knowledge can impact people's recall ability.

Recall Problem Data Collection

Specifically, the problem or phenomenon we are interested in exploring is how people's recall of a stimulus is affected by categorizing the stimulus. In order to analyze this, we presented people with an image - a fairly ambiguous image - and then told the subject that the image was either a cat or a dog. We allowed the subject to study the image for three seconds and then we asked them to recreate the image. We wanted the subject to be able to accurately recreate the image and for us to be able to read the parameter values for the image created. Thus, we created an UI with sliders that allowed the subject to change the value of the different parameters with ease (shown below). Each of the sliders varies from that parameters min value to its max values.

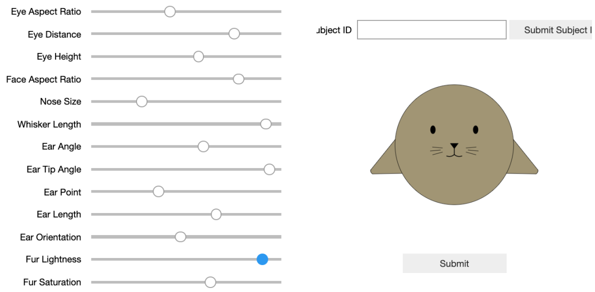

Figure 3: The UI design to allow subjects of the Markov Chain experiment to modify the drawing. Each slider only allowed specific step sizes, depending on the range of the variable.

We were initially planning on just showing the subject an image and then having them simply draw what they could recall. However, after trying that out on a couple subjects, it became clear that this wasn't a good plan. It was pretty hard for the subject to draw the image from memory. It was also hard for us to then extract the feature vector from this image. Thus, we decided to go ahead and take the time to implement this slider UI that allowed the subject to vary the different parameters and see the image changing. This seemed to give the subjects a lot more confidence in their recall image and also allowed us to get the exact feature vector of the image created. It is also important to note that when we show subjects the UI, all the sliders default to their middle value and from there the subjects can move them in whichever direction they desire.

Now that we had a UI that allowed us to effectively and accurately gather information, we ran this experiment. We showed 10 people an image and classified the picture as either cat or dog. We then had them use this UI to redraw the image they saw.

Qualitative Results

Looking at the data, it becomes clear that the labelling of the image seems to skew the recall towards whatever category we labelled the image as. When we showed an image and labelled it as a cat, the image the subject then drew from memory skewed more cat-like. On the other hand, if we showed an image and labelled it as a dog, the subject redrew an image that looked more dog-like.

For example, we showed the following image to two different subjects:

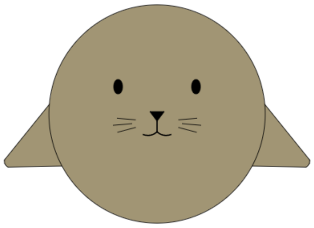

Figure 4: The original ambiguous image shown to subjects 1 and 2.

For subject 1, we labelled the image as a cat. We then asked the subject to redraw the image and they created the following:

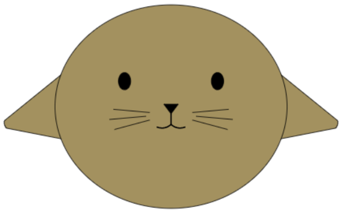

Figure 5: The recall image created by subject 1.

As you can see, the recall is fairly good but the re-created image does appear to have more cat-like features, such as longer whiskers and higher ear angle values (which is in accordance with the values obtained from the posterior mean of the multivariate Gaussian distributions). By ”cat-like”, we mean values of features that we identified to be associated with cats with our multivariate Gaussian distributions. It is quite interesting to see which features changed and how. This is another way to evaluate what features and what values for those features humans tend to associate with cats.

The most prominent features that changed seem to be face-ratio, whisker length, eyes, and ear angle. Specifically, as we could see from our Gaussian for cats, humans tend to associate cats with wider faces and here we can see that the subject made the face slightly wider. In addition, the subject also made the whiskers longer as whiskers are more associated with cats rather than dogs. The subject also moved the ears slightly higher, which makes sense as our Gaussian showed that people think of cats as having higher, pointy ears. The eyes also seemed to have gotten a bit bigger as well as wider. Although this may say something about people’s perception of cats, it is more likely that this was more dependant on the face. The subject, in this case, actually changed the face aspect ratio first and then went back and moved the eyes in this way. So, this change is more likely due to them trying to fit eyes into this wider face.

We then showed the same image to a different subject but now labelled it as a dog. This was the image that subject 2 recreated:

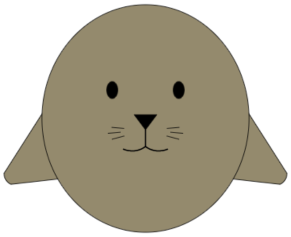

Figure 6: The recall image created by subject 2.

Again, we can see that labelling the image as a dog seems to have skewed the subject to produce something more ”dog-like.” Looking at this image, the most prominent features that changed are face ratio, whiskers, and nose/snout. In contrast to subject 1, subject 2 made the image longer with shorter whiskers which, based on our Gaussians, were representative of people’s idea of a dog. The ears also drooped a bit more and the eye got closer together. This time the subject made the whiskers shorter as shorter whiskers make the image appear more dog-like to the subject.

When we compare the values of the different parameters, they actually end up being fairly close to that of the original image. But a slight change in multiple parameters together work to make the image appear more like a cat or a dog to the subject. Overall, labelling the image did skew people’s ability to recall and redraw the stimulus they were presented with.

Modeling Noise

Having obtained qualitative results, we then sought to model this phenomenon within the framework of Bayesian inference. In this model, we are trying to model the amount of noise created by recall as well as formalize the impact of labelling the image on recall. Thus, we will consider the problem where a subject needs to recover a target (image) from a noisy stimulus (memory recall), given that the target was sampled from a Gaussian.

We define the following variables: the recalled image S, the presented target image T, category c, category variance $\sigma_c{}^2$, and noise variance $\sigma_s{}^2$. The target T is produced by sampling from the Gaussian representing category c with mean $\mu_c{}^2$ and variance $\sigma_c{}^2$. The target distribution is therefore:

$$T|C \sim N(\mu_c{}^2, \sigma_c{}^2)$$

However, subjects cannot directly recover/recall T due to noise in the memory recovery process. Instead they are only able to recover a noisy image S that is normally distributed around the target image with noise variance $\sigma_s{}^2$:

$$S|R \sim N(T, \sigma_s{}^2)$$

If we integrate over T, it yields:

$$S|c \sim N(\mu_c{}^2, \sigma_c{}^2 + \sigma_s{}^2)$$

So, under the given assumptions, the image that the subject is able to recall is normally distributed around the category mean with a variance that is the sum of the category and noise variances.

Given a specific category, for example cat, the experimenter can try to infer the target image T given the recovered image S knowing that the image is from category c (the image is of a cat). Bayes' rule gives:

$$p(T|S,c) \propto p(S|T)p(T|c)$$

Using equation 1 and 2, this can be simplified into the following [1] (details about this math in Appendix of reference [1]):

$$p(T|S,c) = N(\frac{\sigma_s{}^2 S + \sigma_s{}^2 \mu_c}{\sigma_s{}^2 + \sigma_c{}^2}, \frac{\sigma_c{}^2 \sigma_s{}^2}{\sigma_c{}^2 + \sigma_s{}^2})$$

This gives us a Gaussian distribution whose mean is in between the stimulus S and the category mean $\mu_c{}^2$. The posterior probability distribution can be summarized by its mean (the expectation of T given S and c):

$$E[T|S,c] = \frac{\sigma_s{}^2 S + \sigma_s{}^2 \mu_c}{\sigma_s{}^2 + \sigma_c{}^2}$$

Thus, the best guess of the target is a weighted average of the recalled image (S) and the mean of the category the image is labelled with, where the weighting is determined by the ratio of category variance to noise variance. This quantifies the idea of ”skewing” images due to a labelling. The term μ_c pulls the recall of the target image T towards the category center causing subjects to recall an image that is more cat or dog-like depending on the labelling of the image.

Fitting Noise

Based on the data collected and the equations described above, we tried to find a way to approximate noise in our data set. Specifically, we are trying to determine how much noise exists in our subjects' recall step. We wanted to try to solve for what value of $\sigma_s{}^2$ most closely modeled what we saw in our experiments/data.

As mentioned before, we had 10 subjects, each of whom were shown an image for three seconds that they later had to redraw. Each subject performed this recall task on only a single image. The reason we had each subject only perform this task once is because we thought there was a possibility that doing more would skew the data as their recall of previous images may influence what they redraw. So, although this is not a lot of data to use to fit noise, it was the most we were able to collect.

For each of our subjects, we knew both T and S (feature vectors) and so were able to use this to solve for the noise variance. We assumed that $E[T|S,c] = T_{actual}$ and calculated the noise variances that would most closely fit the data we collected. In the end we got a good estimation for noise in our data-set:

Table 1: Noise Variance

Parameter	Min-Value	Max-Value	Variance
Eye Aspect Ratio	0.5	1	0.174
Eye Distance	50	80	7.902
Eye Height	10	15	1.423
Face Aspect Ration	0.75	1.33	0.112
Nose Size	10	20	2.68
Whisker Length	1	60	20.54
Ear Angle	-50	50	18.639
Ear Tip Angle	40	60	6.208
Ear Point	0	80	7.772
Ear Length	80	120	12.85
Ear Orientation	0.3	0.5	0.059
Fur Lightness	30	80	11.4809
Fur Saturation	0	60	18.484

These are pretty interesting results and it is useful to see which parameters had the least variance and which had the most variance. The parameters that had the least variance are face aspect ratio, ear point, and ear angle. This could indicate that these are the features that subjects tend to pay the most attention to. These could simply be considered the ”major” features in that, upon looking at an image, they are the first ones that a subjects notes. This seems quite likely as they are all fairly noticeable features - particularly face aspect ratio and ear angle. After finishing the experiment, the subjects said that they often could not remember a lot of specifics but they could remember the shape of the face - that it was round, or wide, or long. The also said they could immediately place the ears or at least the general location of the ears - at the top of the head, halfway down, all the way down, etc. They also said that they spent a bit of time studying the ears as they are a major feature and so they felt as though they remembered ear point and placement fairly well.

The features that seemed to have the most variance were eye aspect ratio and whisker length. After talking to the subjects, it seemed that these were high variance for different reasons. Subjects said that they could not really recall much about the eyes except generally if they were larger or smaller. So eye aspect ratio was quite difficult to remember other than just what seemed to fit within the face shape they re-created. Whiskers, on the hand, was more deliberate. Whiskers always seemed to vary towards larger if the image was labelled as a cat and smaller if the image was labelled as a dog. Subjects said that for the most part, they didn't really pay attention to whiskers. But, upon re-creation, they put whiskers mostly based on the classification of the image. So whiskers seem to have had high variance due to the classification of the images.

It would have been really interesting to also run the same experiment without the classification aspect to see if any of our variance values changed significantly but unfortunately we did not have the time/subjects to do so.

Markov Chain

Now that we have found a good estimation for the noise variance, we wanted to run a similar experiment but with a Markov Chain. We wanted to see if people's recollection would eventually converge to something that appeared quite cat-like/dog-like given that the images were labelled as a cat/dog.

Experiment

We began with an ambiguous image labelled as a cat, which we showed to subject 1. We allowed the subject to study the image for three seconds. Then we took away the image and asked the subject to use our UI to recreate the image. We then took this recreated image and gave it to subject 2. Subject 2 studied this image - which was again labelled as a cat - and then was made to recreate it. The recreated image was given to subject 3 and so on. We used five people in total for this chain.

We then did the exact same thing except now we labelled the image as a dog. We started from the same ambiguous image as before (now labelled as a dog) and did this chain with five people again.

Results

This is the output from the cat chain:

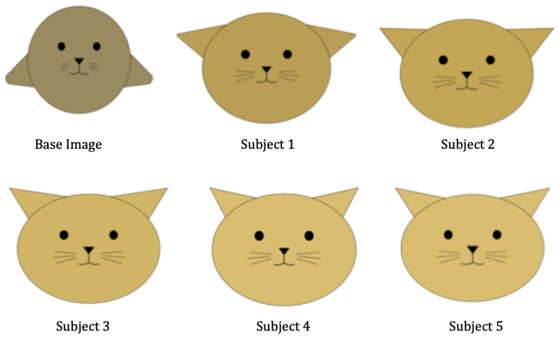

Figure 7: The full Markov chain given 5 participants of images generated when all were given the label ”cat”

This is the output from the dog chain:

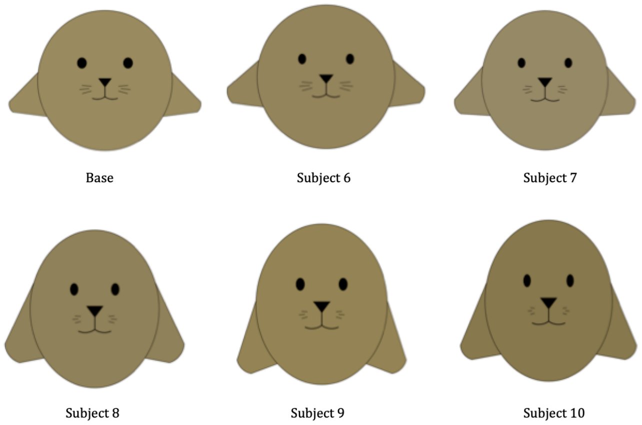

Figure 8: The full Markov chain given 5 participants of images generated when all were given the label ”dog”

The results are very impressive - even within 5 iterations, we can see the image becoming very cat-like and very dog-like. For the cat chain, the first subject skewed quite drastically to a cat. It is very likely that this subject simply did not perform well and just re-drew an image that they thought was cat-like. So, we don't want to place too much emphasis on that first iteration as it does seem more severe than expected. However, if we look only at the iterations after that, we can still see that consecutive subjects made the image even more cat-like. Analyzing specific features, we see the whisker length increased each time in the recall stage. Ears and face shape also seemed to be common features that skewed more cat-like at each iteration. Even if we disregard the first subject as having performed quite poorly, we can see that the following subjects made the ears higher and pointier at each iteration. The subjects also made the faces wider at each iteration. The color also changed fairly dramatically but this is probably because subject 1 changed the color and then from there the color varies slightly.

On the other hands, the dog chain is a bit more gradual but it still does end up being something that strongly resembles a dog. Again, the same features seem to be the ones changing - face shape, ears and whiskers. The face shape gradually becomes longer and longer and the whiskers become shorter and shorter. The ears also droop more and more with each iteration. In the dog chain, the nose also gets bigger at each iteration.

It is pretty interesting that the same features seemed to go to their extremes for the cat and dog chain. Although whiskers started being at around a medium length (30), the final image of cat had whiskers close to their maximum length (58) and the final image of dog had whiskers close to their minimum length (6). Similarly, despite the face shape staring fairly circular, by the end of the cat chain it was quite wide and by the end of the dog chain it was very long. Ears also went very pointy and at the top of the head in the cat chain and droopy and more rounded in the dog chain.

It is quite remarkable that even within 5 iterations, the image could change this dramatically. It says a lot about how much the label was able to influence the subject's recall. The chains do seem to converge to people's general perception of a cat and a dog. The chains also show results that are much more drastic that the results we obtained from our Gaussians. This is probably because for our Gaussians, the subjects had to label cat or dog for images that were fairly ambiguous and so there were many images that were not our ideal image of a cat or dog that got labelled as such. However, in this experiment, subjects had a lot more power and their preconceived notions of cats and dogs were able to impact what they redrew giving us quite drastic results.

Conclusion and Future Work

We were able to experimentally determine the importance of a number of various features on categorization of cartoon images into ”dog” and ”cat” using Bayesian analysis to determine the mean of a multivariate Gaussian for each category, and then used a human Markov Chain process to determine how variation moved images towards the prototypical dog or cat. While the Bayesian analysis did not provide particularly obvious means for every feature for dogs and cats, it did separate some features. The features it was unable to separate may not be Gaussian distributions. The Markov Chain process produces clearer results, and strongly implies the potential of this methodology to analyze the more complicated landscape of the distribution of ”cat” versus ”dog”.

This research could be extended with analyzing how various time gaps affect the noise in the Markov Chain model. We would expect larger time gaps to produce larger noise, or a heavier reliance on the label rather than the specific image. It would also be interesting to compare the Markov Chain model to a control case that was not given a label; without a label, people may simply vary around the original image with some noise, or it may prove to be an unstable equilibrium, where extremes are eventually reached with some random probability. As the mean of the original Bayesian analysis did not appear to produce the prototypical "cat"/"dog" we expected, we may be able to use the Markov chain process as a sort of human gradient descent towards finding a number of peaks in a multi-peaked distribution. Another extension of this research would be to change the variables chosen to draw the images, such as including fur patterns or fur tufts. This would add complexity Alternatively, allowing for a broader labelling process (such as specific breeds or a wider set of animals) may provide more realistic identification of how people simplify animal features into cartoons.

We believe this work is particularly interesting for the levels of abstractions that are encoded within the design of this experiment. The abstraction of the cartoon from a realistic image of a cat or dog is representative of a visual language that has evolved and is culturally dependent. The decisions made by people in deciding "cat" or "dog" rely on years of experience of not only seeing cats and dogs, but also in seeing *cartoon* images of cats and dogs. There is then abstraction in the memory of the person, then in the recalled image, and finally in the limitations of the drawing tool given to them to represent their recalled image. This work attempts to use Bayesian modeling to analyze how all these levels of abstraction interact. However, it also notes the limitations of Gaussian modeling on data that we do not know to be normal.

Individual Contributions

All of us collaborated on the overall design of the experiment, as well as the general analysis steps.

Srilaya: Srilaya collected classification data from five individuals after generating random face images. She primarily worked on modeling the features for each category as multivariate Gaussian distributions. She fit multivariate Gaussian distributions for the features classified as cat and dog separately, sampled from this distribution, and obtained a posterior mean of feature values for each category using `python`, `numpy`, `scipy`, and `pymc3`. She analyzed the results, including the differences in the means of both distributions and the covariances between the feature variables. Additionally, she wrote the Abstract, Introduction, Related Work, and Modeling Features as Multivariate Gaussian Distributions sections.

Priya: Priya designed and wrote the code to generate the cat and dog images. She also designed and wrote the UI to allow people to generate their own cat/dog images. These were done in `Python` using `drawSvg`, `pandas`, and `ipywidgets`. She wrote the Dataset Creation and Experimental Setup and Conclusions and Future Work sections

Meena: Meena did the noise and Markov chain experiments and writeup. She helped collect classification data from five individuals after generating random face images. She conducted the experiments about recall (from 10 subjects) and analyzed the data to find the noise variance. She did the Bayesian inference analysis of the noise and looked into the results of the different feature variances and how they related to the Gaussian model. She also ran the Markov Chain experiment for both the cat and dog chain. She wrote the Bayesian Modeling of Noise in Recall and Markov Chain sections.

References

[1] Feldman, Naomi H., Griffiths, Thomas L, Morgan, James L. (2009). The Influence of Categories on Perception: Explaining the Perceptual Magnet Effect as Optimal Statistical Inference. *Psychological Review, 116*. 752-782.

[2] Liberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*. 358-368.

[3] Davidoff J., Davies, I., K.S., Roberson, D (1999). Colour categories in a stone-age tribe. *Nature, 398*. 203-204.

[4] Etcoff, N. L., Magee, J. J.(1992). Categorical perception of facial expressions. *Cognition, 44*. 227-240.

[5] Beale, J. M., Keil, F. C.(1995). Categorical perception of facial expressions. *Cognition, 57*. 217-239.

[6] Goldstone, R. L.(1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*. 178-200.

[7] Guenther, F.H., Husain, F.T., Cohen, M.A., Shinn-Cunningham, B.G. (1999). Effects of categorization and discrimination training on auditory perceptual space. *Journal of the Acoustical Society of America, 106*. 2900-2912.

[8] Livingston, K.R., Andrews, J.K., Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*. 732-753.

[9] Goldstone, R.L., Lippa, Y., Shiffrin, R.M. (2001). Altering object representations through category learning. *Cognition, 78*. 27-43.

[10] Levin, D.T., Beale, J.M.(2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception Psychophysics, 62*. 386-401.

[11] Huttenlocher, J., Hedges, L.V., Vevea, J.L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General, 129*. 220-241.