Perceptual Magnet Effect in Handwritten Digit Categorization

Palak Bansal
NYU CDS (2024)
pb2766@nyu.edu

Siri Desiraju
NYU CDS (2024)
scd4156@nyu.edu

Rodrigo Kreis de Paula
NYU CDS (2023)
rk4197@nyu.edu

Jin Ishizuka
NYU CDS (2023)
ji721@nyu.edu

Abstract

The perceptual magnet effect is a well-studied phenomenon wherein knowledge of predefined categories influences human perception. Attempts to model this phenomenon with Bayesian statistics have largely focused on its effects on perceived audio stimuli. In our project, we aim to extend work in this area into the domain of visual perception. Specifically, we develop a Bayesian model for the task of handwritten digit classification. We then demonstrate a perceptual magnet effect under our Bayesian framework by comparing the probabilistic outputs from our Bayesian model to the probabilistic outputs from a traditional convolutional neural net. Our results further demonstrate the utility of a Bayesian approach when attempting to model the biases present in human perception.

1 Introduction

The influence of predefined categories on human perception has been well documented. One well-studied consequence of this influence is the perceptual magnet effect, a phenomenon that is characterized by the warping of perceived stimuli toward categorical prototypes. Typically, this involves a shrinking of the perceptual space near category centers and an expansion of the perceptual space near category boundaries.

Studies of this phenomenon have primarily focused on its effect on perceived audio stimuli. Notable among these studies is one by Feldman and Griffiths (2009), who introduced a Bayesian model that formulated the task of speech perception as a statistical problem. In their study, they assumed that listeners attempting to reconstruct distorted speech sounds were leveraging their prior knowledge of phonetic categories. They then formalized this problem using Bayesian statistics, and, with this framework, they were able to closely replicate human judgments through simulated speech perception tasks.

Our project aims to extend prior work in this area by applying the Bayesian framework to a visual perception task. Specifically, we develop a Bayesian model for the task of categorization of handwritten digits. We then compare the assessments of our Bayesian model to the outputs from a traditional convolutional neural network to demonstrate how a Bayesian approach may more closely replicate biases in human perception. Notably, we see that the perceptual space around ambiguous images is pulled toward categorical prototypes in a manner that is characteristic of the perceptual magnet effect.

2 Methods

Our project can be broken down into two primary components. First is the creation of a convolutional neural network (CNN) to extract learned latent features from the image data and produce baseline class probability estimates. Second, is the development of our Bayesian model. This component of the project largely consisted of estimating the various probability distributions needed for our Bayesian framework. After completing these two tasks, we then work to demonstrate a perceptual magnet effect when using a Bayesian framework by comparing the outputs from our CNN and Bayesian model on a visual categorization task.

We hypothesize that the class probability estimates from our CNN will contain more uncertainty than the estimates from our Bayesian model. Furthermore, we predict that ambiguous images will be pulled toward areas of higher probability under our Bayesian framework, resulting in more definitive predictions than those from our CNN.

2.1 Dataset

The dataset used for this project was the MNIST dataset (link here), which is a large database of handwritten digits widely used in computer vision research. It consists of a set of 60,000 grayscale images of size 28x28 pixels, representing the digits from 0 to 9. For the purposes of our project, we filtered the dataset to only include handwritten 5s and 8s. These digits were selected due to their relatively similar shape and their ability to potentially confuse classification algorithms, as demonstrated in Figure 1

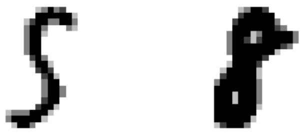

Figure 1: A pair of similar-looking 5 and 8, extracted from the MNIST dataset

Thus, the primary task for our classification models in this study was to correctly distinguish between handwritten 5s and 8s. We further split the dataset into a training set of 9,370 images, a validation set of 1,902 images, and a test set of 1,866 images. All datasets were comprised of an approximately 50-50 split between the two classes.

2.2 Convolutional Neural Network

The first task of this project was to construct a CNN model. This model would be used to both extract latent features in the image data for use in our Bayesian model as well as generate class probability estimates for comparison with our Bayesian class probability outputs. The network consists of two convolutional layers with ReLU activation functions and max pooling along with two fully connected layers. The output from the final fully connected layer is a vector of size 2 which is then passed through a log softmax module to generate class log probabilities. In addition to the class log probabilities, the model also outputs a set of latent features. The array of features is extracted prior to the final fully connected layer and is of size 10. This architecture is shown in Figure 2.

In order to train the model, we used an Adam optimizer with a learning rate of 0.001, and a negative log-likelihood loss (NLLLoss). The CNN was trained for five epochs, and the model with the lowest validation loss was saved.

After training our CNN model, we iterated over all images in the training set and saved the extracted latent features for each image. These hidden features were then used to determine the feature value probability distributions used in our Bayesian classifier.

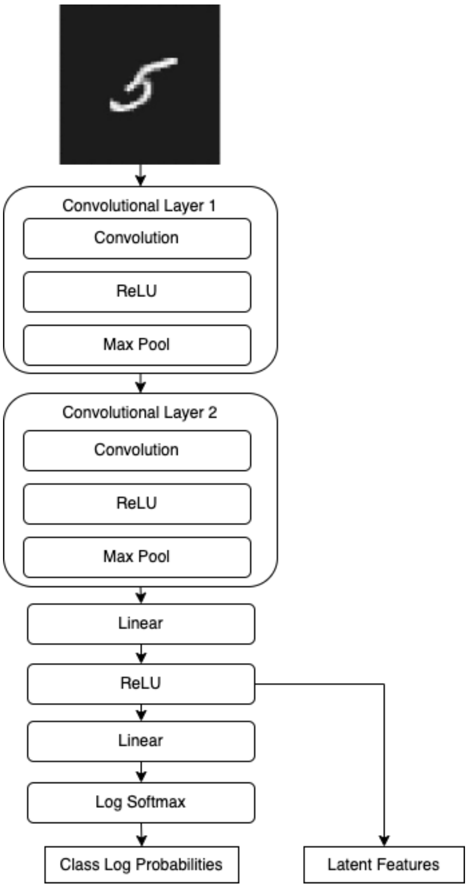

Figure 2: CNN architecture

It is important to mention that out of the 10 resulting hidden features, two of them (namely “Feature 1" and “Feature 3") were composed of only zero values. As a result, they could not generate any value for our model and, therefore, were excluded from our resulting set of image features.

2.3 Bayesian Model

Our Bayesian model calculates the probability of class membership given a set of image features. The data used in this model are the latent features obtained through the CNN described in subsection 2.2. We leverage Bayesian statistics to calculate the probability of each category given a set of features, $P(C|F)$. In the context of our project, Bayes Theorem is defined as $P(C|F) = \frac{P(F|C)P(C)}{P(F)}$. Therefore, we first need to compute each of the terms $P(F|C)$, $P(C)$, and $P(F)$. The priors are simply assumed to be equal for both categories given the approximate 50-50 split between the two classes and are therefore assigned a value of 0.5. The majority of work on this approach was therefore focused on computing $P(F|C)$ and $P(F)$.

In order to compute $P(F|C)$, we first computed the likelihood of each feature given the category, $P(f_i|C)$, for all features $f_i \in F$. In order to estimate these conditional probability distributions, $P(f_i|C)$, we attempted to fit a variety of popular continuous probability distributions to our data. Among the distributions we tested were normal, exponential, beta, gamma, t, and chi-squared distributions. For each distribution, we found the set of parameters that best fit our data and then selected the distribution with the best overall fit, as determined by p-value. This process was then repeated for each feature-category combination. The best-fitting distributions for each feature-category pair are shown in Figure 3.

Once we obtained our estimated $P(f_i|C)$ distributions, we were able to compute the overall likelihood of an image given a category $P(F|C)$. To do this we made a common, simplifying assumption and assumed conditional independence between all of our features. This allowed us to calculate the overall probability of an image given a category as the product of the probabilities of each feature given the category: $P(F|C) = \prod_i P(f_i|C)$.

Next, we worked to compute $P(F)$. We achieve this by again using our assumption of independence between features and first attempting to estimate each distribution for $P(f_i)$. However, one issue we encountered was the irregular distribution of these feature values which made it difficult to adequately fit some of the common continuous distributions we used previously. To remedy this, we instead generated binned histograms of our feature values and computed $P(f_i)$ as the proportion of data points in a given bin to the total number of data points. We then computed $P(F) = \prod_i P(f_i)$

After computing all of the necessary probabilistic components, we were able to define our Bayesian model. The model takes as input an array of latent feature values for a given image and returns the probabilities of category membership for the image as defined by Bayes Theorem.

3 Results

In order to demonstrate our Bayesian framework's ability to replicate a perceptual magnet effect, we compared the class probability outputs produced by our CNN and Bayesian models when evaluated on our test set. Initially, a magnet effect was difficult to visualize as both models recorded high confidence in their predictions and generated high classification accuracies of approximately 96-97%. This strong performance was likely due to the relatively simple nature of the digit classification task. In order to increase the difficulty of the task we introduced noise in the form of random visual distortions to our test set. Specifically, we leveraged Pytorch's Elastic Transform augmentation. We experimented with five different levels of distortion. The amount of distortion is quantified by the alpha parameter with a higher alpha value corresponding to a higher level of noise. Examples of these transformations are shown in Figure 4.

After generating our test sets with varying levels of distortion, we compared the outputs of both our CNN and Bayesian models. In order to visualize the perceptual magnet effect, we generated histograms displaying the distribution of confidence scores for the correct labels in our test set. These plots are shown in Figure 5. We can see that at lower levels of distortion, the magnet effect is not clear as both models generate high confidence in the correct class. However, as the amount of noise increases, we can see the influence of a perceptual magnet effect. Specifically, we see an expansion of the perceptual space around ambiguous data points (i.e. data points with a CNN class probability of around 20%-80%). Under our Bayesian framework, these ambiguous data points are pulled toward areas of greater certainty (note: a confidence score of 0% for the correct class still corresponds to a confidence score of 100% for the incorrect class). In this way, we can see that our Bayesian model is able to distort the perceptual space around the class boundary in a manner that is consistent with a perceptual magnet effect.

In addition to comparing confidence scores, we also evaluated the accuracies of our two models. We see that our CNN generated consistently stronger performance than our Bayesian model across all levels of image distortion. Unsurprisingly, we also see that accuracy for both models declines as the amount of noise increases. These accuracy scores are displayed in Table 1.

4 Discussion and Limitations

In this project, we investigated the perceptual magnet effect by developing a Bayesian model that closely replicates human-like biases when attempting to classify handwritten digits. We first trained a CNN to extract latent image features and generate baseline class probability estimates. We were then

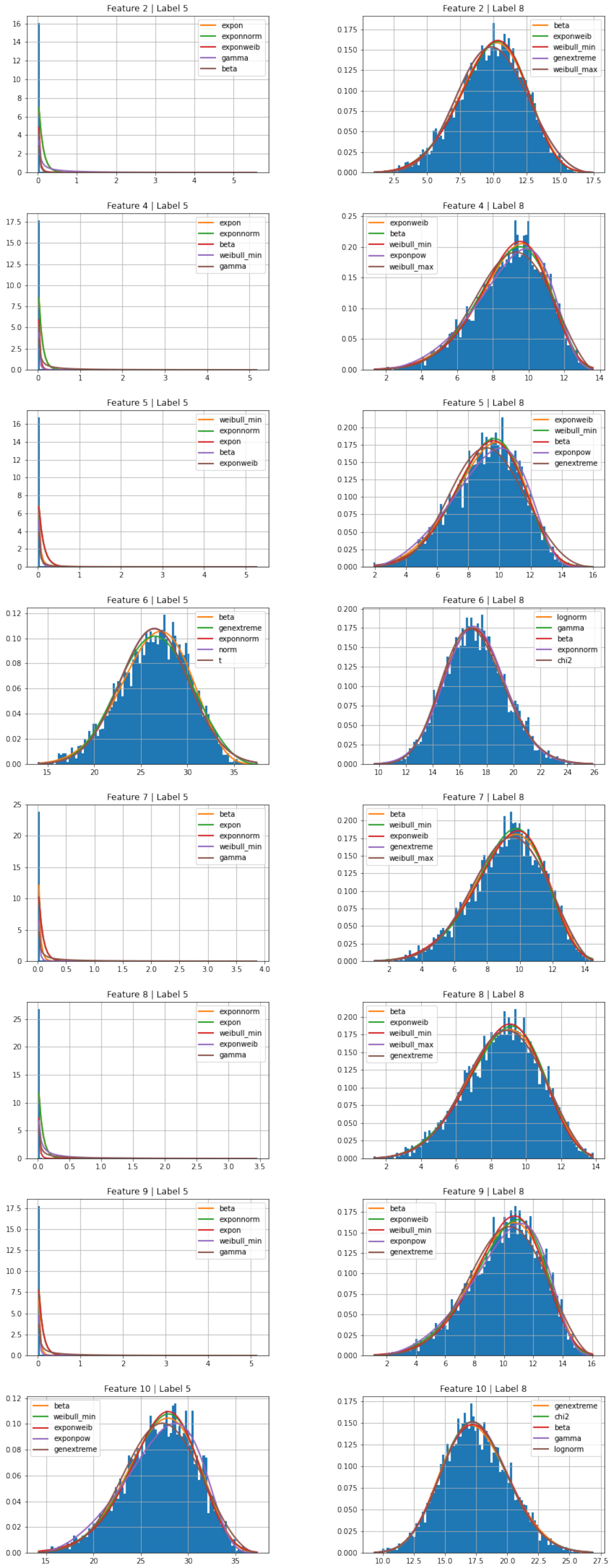

Figure 3: Distributions for each Feature and Label

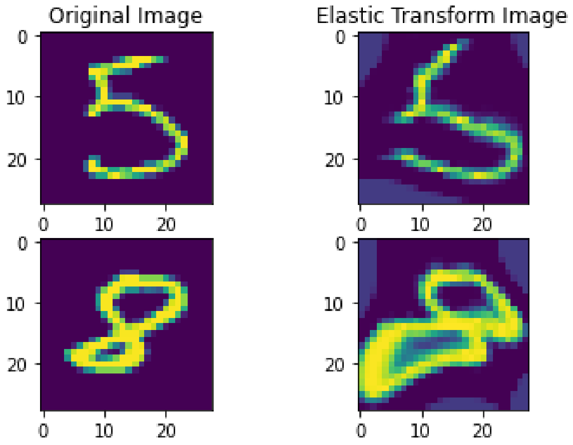

Figure 4: Elastic Transformations

Alpha	CNN Accuracy	Bayesian Accuracy
50	0.987	0.951
75	0.966	0.917
100	0.934	0.861
125	0.884	0.807
150	0.833	0.767

Table 1: Test set accuracy across varying levels of distortion

able to use the extracted latent features from our training set to generate the probability distributions needed for our Bayesian model. After generating several test sets with varying levels of image distortion, we compared the outputs from our CNN and Bayesian models.

Our results support our original hypothesis. We were able to show that our Bayesian model was able to closely simulate human-like biases on a task of image classification. Specifically, our Bayesian model's estimates were able to capture the warping of perceived stimuli toward categorical prototypes, which is characteristic of the perceptual magnet effect, as demonstrated in Figure 5. This was particularly prevalent when using datasets with higher levels of noise. Although our Bayesian model produced more confidence in its predictions, these predictions were sometimes incorrect, particularly at higher levels of image distortion. However, the manner it which it was incorrect closely mirrors the manner in which humans also make errors in classification. Just as humans often over-rely on knowledge of predefined categories in the presence of noise, our Bayesian model also generated less nuanced predictions when evaluating images with high degrees of distortion. In this way, we demonstrate the utility of a Baysian approach when attempting to model the biases present in human decision making.

Our study is not without limitations, however. First, our study focused on the relatively simple task of classification between two pairs of handwritten digits. In order to demonstrate the robustness of our findings, future studies could explore the perceptual magnet effect on classification tasks involving more than two classes and eventually on tasks involving more complex image datasets. A second limitation would be our Bayesian model's reliance on oftentimes rough estimates of certain feature value distributions. This is particularly true when attempting to estimate the overall distributions for our features $P(F)$. As shown in Figure 3, many of our features have a roughly exponential distribution for class 5 and roughly normal distributions for class 8. Attempting to estimate the overall distribution for each of these features was difficult, and our solution of using binned histograms produced only moderately accurate estimates of the true underlying distribution. We feel that this could have limited the performance of our Bayesian model. Lastly, a final way to extend and add robustness to this study could be to compare our Bayesian model's evaluations to evaluations from a panel of human evaluators. Due to the relatively simple nature of MNIST dataset, we felt as though the task of classification would be trivial for humans. However, future studies working with more complex datasets could compare the evaluations produced by a Bayesian framework to human evaluations on tasks of visual classification.

Overall, our study contributes to the growing body of research on the perceptual magnet effect and extends it to the realm of visual perception tasks. Our findings suggest that a Bayesian approach may be a useful tool for investigating perceptual biases in visual perception tasks, and future research could further explore the potential benefits of this approach in other classification tasks.

Group Contribution

All members of our group contributed equally to the research and development of this project's coding, results, and final paper.

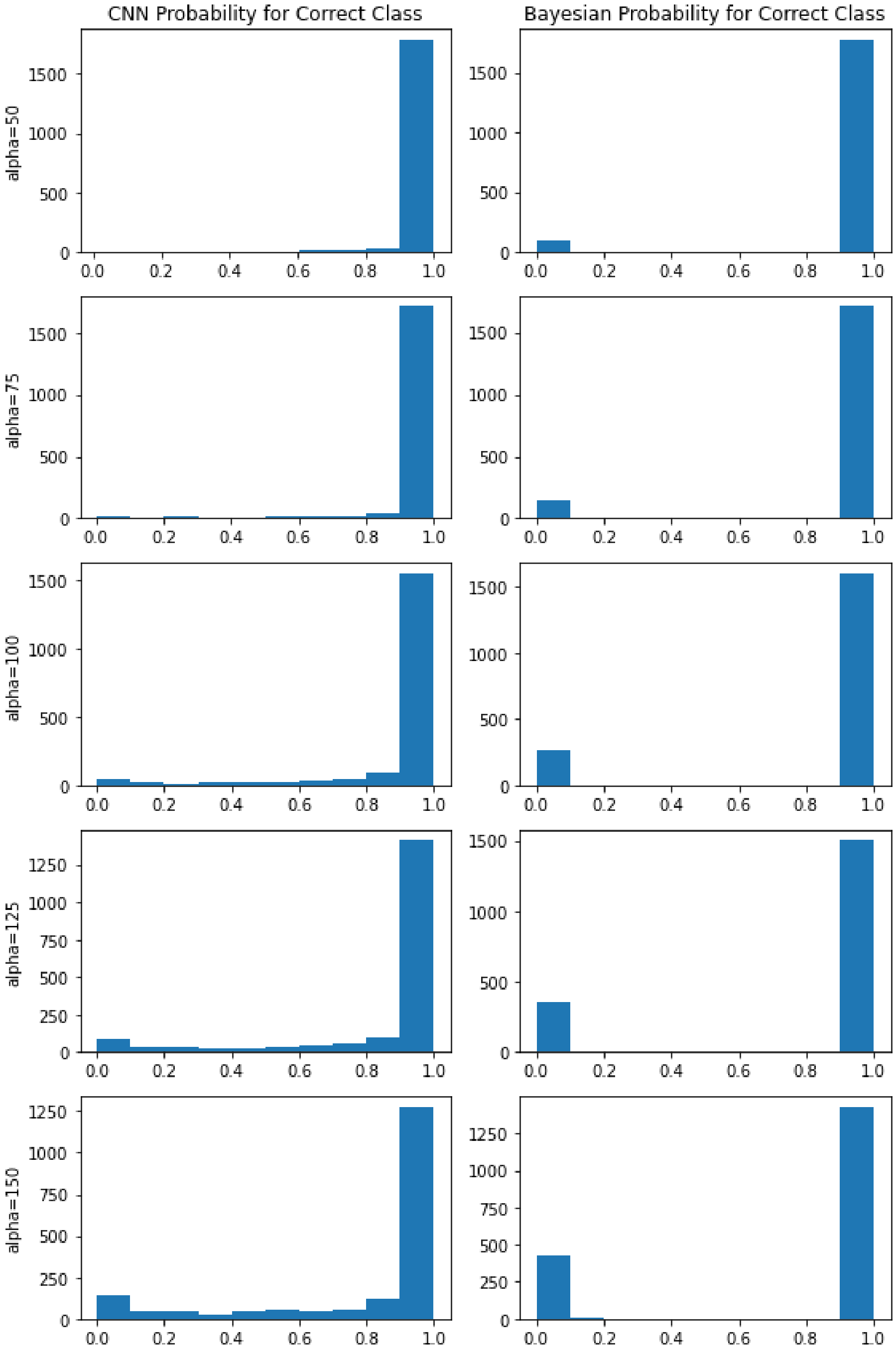

Figure 5: Histograms of probability estimates for correct class

Acknowledgements

We would like to express our gratitude to Professors Brenden Lake and Todd Gureckis for teaching us DS-GA 1016 - Computational Cognitive Modelling for their expertise, dedication, and efforts in making the course an invaluable learning experience. We also extend our thanks to Teaching Assistants Pat Little and Francesco Mantegna for their support and feedback on our homework assignments and course project.

A Appendix - Code

All codes developed and used in this Project are publicly available on the following GitHub Repository: https://github.com/jnshzk/CCM_project.

References

Mark Feldman and Thomas Griffiths. 2009. A bayesian framework for word segmentation: exploring the effects of context. *Cognitive science*, 33(4):569–599.

Yann LeCun, Corinna Cortes, and CJ Burges. 2010. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

(Feldman and Griffiths, 2009) (LeCun et al., 2010)