ECE324 Final Report
FearNet

Avi Kraft: 1004341535
Victor Yip: 1003795333

December 3rd, 2019

Word Count: 1975

1 Introduction

The objective of FearNet is to design an image classification model capable of identifying images likely to trigger phobias. Online web browsing is an environment where it is often difficult to provide a safe, personalized experience due to the volume of data on image-sharing platforms. As such, FearNet aims to serve as the necessary framework for an image-filtering system able to selectively handle images based on each user's individual phobias.

A neural network architecture excels in this problem due to the possibly abstract and highly varying visual descriptions of individual classes. Using claustrophobia as an example, it is not immediately obvious what in an image constitutes a claustrophobic trigger; it has numerous contrasting descriptions, such as an image of a single person stuck in an elevator, as compared to an image of a crowded space.

2 Background and Related Work

Based on a brief study, no public machine learning-based algorithm or paper was found addressing the challenge of detecting common visual phobia triggers. There are however existing ML models with the goal of filtering disturbing or inappropriate content. Facebook, among other companies, is currently attempting to develop deep-learning based content moderation algorithms which can be trained to recognize and filter out violent content (1); however, the company's algorithms have been controversial due to repeated failures to flag clear violent content (2). Similar controversies have attached themselves to the filtering algorithm used by Youtube Kids (3). From these industry examples, it is clear that not only is there room for improvement, but the testing accuracy that should be

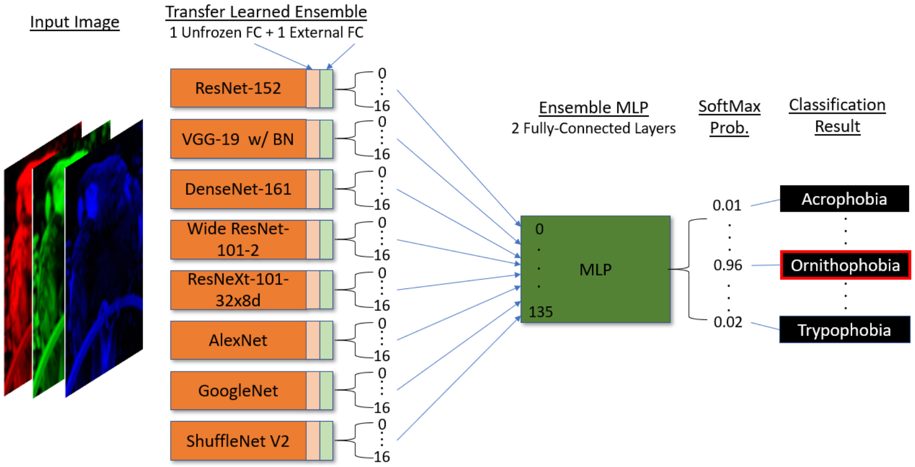


Figure 1: Illustration of model architecture

achieved must also be relatively high (regardless of the number of classification labels).

3 Data and Data Processing

For this project, 16 phobias were chosen as the target classes. This makes 17 image datasets, including the general non-phobia containing class.//

3.1 Image Sources

Data was collected by going through image collection sites, finding collections associated with phobias. The main sites used for this were pexels.com, pixabay.com, gettyimages.ca, 123rf.com, and istockphoto.com. With the exception of carcinophobia and acrophobia (where relevant collections did not exist or had too much irrelevant content on these five main sites), two or three sites were selected for each phobia - resulting in approximately 1000 images per phobia. Other sites used for specific phobias include:

- artinnaturephotography.com/gallery/openspaces (Agrophobia)
- digital-photography-school.com/15-spectacular-lightning-images/ (Astraphobia)
- freeimages.com/search/blood (Hemophobia)
- reddit.com

  - r/spiders (Arachnaphobia)
  - r/lightning (Astraphobia)
  - r/dogs (Cynophobia)
  - r/vomit (Emetophobia)
  - r/snakes (Ophidiophobia)
  - r/BACTERIA (Mysophobia)
  - r/trypophobia (Trypophobia)

- (Trypophobia)
  - distractify.com/trending/2018/09/06/19Ieyf/trypophobia-triggers
  - ranker.com/list/find-out-if-you-have-trypophobia-/brian-gilmore
- (Carcinophobia)
  - oralcancerfoundation.org/dental/oral-cancer-images/
  - cancer.org/cancer/skin-cancer/skin-cancer-image-gallery.html
  - webmd.com/melanoma-skin-cancer/ss/skin-cancer-and-skin-lesions-overview
  - skinvision.com/skin-cancer/pictures
  - dermnetnz.org/image-catalogue/lesion-tumour-and-cancer-images/
- goodhousekeeping.com/life/pets (Cynophobia)
- (Acrophobia)
  - tripjaunt.com/5-bridges-will-trigger-acrophobia-fear/
  - pinterest.ca/melodyannmyers/a-fear-of-heights/
  - ranker.com/list/afraid-of-heights-photos/ashley-reign
- buzzfeed.com/karstenschmehl/claustrophobia-woes (Claustrophobia)

For these other sites, collections were taken in their entirety.

3.2 Data Scraping Methods

All collections were bulk-downloaded using Image Downloader in Google Chrome.

3.3 Data Cleaning

The data was noisy initially for a handful of reasons:

- Images tangentially related to the phobia
- Arguably unrelated images (in abstract phobia datasets in particular)
- Unrelated website features such as logos or user profile photos

Each set had to be cleaned manually. After cleaning, the number of images per phobia ranged from approximately 200-1300.

All images in the dataset are processed and resized to a 128x128x3 resolution with normalized pixel values of a mean of 0 and variance of 1 across all 3 RGB channels.

3.4 Data Statistics

The category with the largest number of photos is emetophobia (fear of vomit) at 1305. The fewest photos were found for trypophobia (fear of holes) at 177. On average, there are 712 images per phobia.

3.5 Example Images

Figure 2: Snake

Figure 3: Group of people hanging off a cliff, potentially triggering for people with fear of heights

4 Architecture

The final model makes use of transfer learning and ensemble learning techniques on pre-trained convolutional models.

8 pre-trained convolutional models are used, namely:

- ResNet-152
- VGG19BN
- AlexNet
- DenseNet-161
- ResNeXt-101 (32x8d)
- GoogLeNet
- Wide ResNet-101-2
- ShuffleNet V2-x1-0

In order to tune each model to the task, each pre-trained model has its output layer unfrozen, and a separate linear output layer added to model. The number of neurons in the original unfrozen output layer corresponds to half the number of neurons in the previous layer in the model, while the added linear output layer has 17 neurons.

The 17x1 outputs from each pre-trained model are concatenated together to form the 136x1 input vector for a multi-layered perceptron. The MLP consists of two fully-connected layers, with 68 and 17 neurons respectively. Batch normalization and a ReLU activation function are used on the first, as well as on both unfrozen and added linear layers in each pre-trained model.

The final output of the MLP is a 17x1 tensor, which is then softmaxed so that each element represents the probability of the presence of that phobia in the image.

It should be noted that the the individual pre-trained models were trained separately from the MLP (using the same train/validation/test data split). Although the potential benefits of doing so are not exercised here, this gives more freedom in hyperparameter tuning between the transfer models and the MLP.

5 Baseline Model

The baseline model is a convolutional neural network with two convolutional layers - each with 50 kernels (first layer is 3x3, second is 5x5), batch-normalized output, a ReLU activation function, and is followed by a pooling layer of size 2x2 with stride of 2. Following the convolutional layers is a batch-normalized fully-connected output layer with 17 neurons and a softmax activation function.

6 Quantitative Results

The final model achieves the following quantitative results:

- Training Accuracy: 99.5%
- Validation Accuracy: 86.2%
- Testing Recall Score: 85.8%
- Testing Accuracy: 85.7%

With the following hyperparameters:

- Learning rate = 0.001
- Loss function = Cross-Entropy Loss
- Optimizer = Adam

- 20 epochs
- Random seed = 1

Recall is used as the primary testing metric due to the emphasis on false negatives in classification problems with potential health impacts. If the model is deployed in the real-world, a false positive is unlikely to have serious consequences, while a false negative may result in a phobia being triggered.

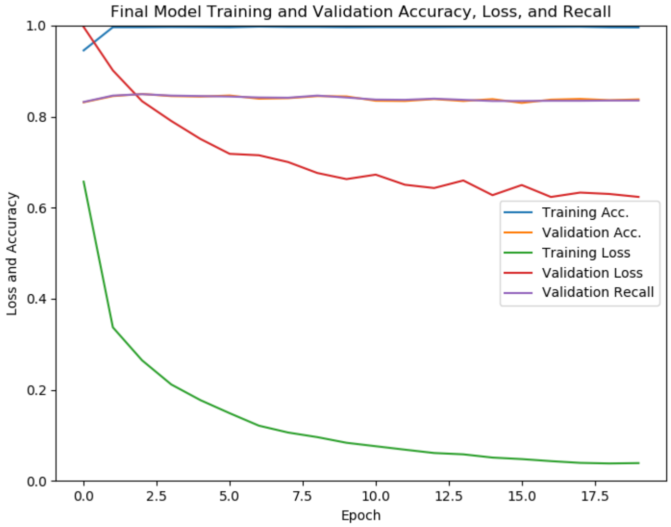


Figure 4: Training and Validation Loss and Accuracy

The ensembled MLP (Figure 4) begins training at a high accuracy/recall due to the transfer-learned networks, however this quickly plateaus over a few epochs. Some overfitting is observed in the architecture as the training and validation accuracy plateau at different scores, however, the affect is not severe as the validation loss continues to decrease.

The final transfer-learning ensemble architecture achieves a higher accuracy and recall score than each individual transfer-learned pre-trained network, all of which far exceed the performance of the baseline (Figure 5,6).

It should be noted that the results shown during final presentation were inaccurate due to a mix-up in the training and validation/testing sets during the separate training of the pre-trained networks and MLP.

The normalized confusion matrix (Figure. 7) was chosen to represent the per-class classification accuracy due to the varying number of data samples per phobia.

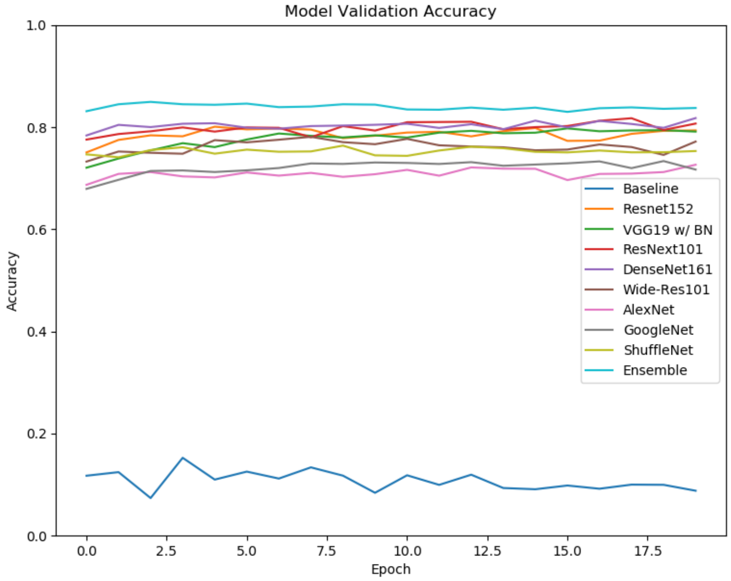


Figure 5: Validation Accuracy of Transfer Models, Baseline, and Architecture

7 Qualitative Results

An example correct I/O of our final model is shown in the form of a GUI (Figure 8). The GUI allows the user to specify their phobia(s), select the image to be inspected, and blocks the image if triggers are present. When an indicated phobia is detected, a prompt is displayed displaying the confidence (softmax'd probability) of prediction.

A second, incorrect example I/O is shown in Figure 9. This image shows an irregular pattern of holes which may be triggering to people with Trypophobia, yet the image misclassifies it as a dog. Misclassifying images as dogs is a common error for our model due to the Cynophobia (fear of dogs) dataset being much larger than most of the others, whereas Trypophobia has a low accuracy due to having the smallest dataset.

Another incorrect example I/O is shown in Figure 10; here, an image of a group of people carrying a coffin – a potential thanatophobia trigger – is misclassified as containing birds. Thanatophobia is one of the phobias on which our model achieves lowest accuracy, probably due to death being a very abstract concept as well as its dataset being of below-average size. In the case of this particular image, the incorrect result appears to be due to the shape of the flowers on the coffin, which somewhat resemble the wings of a bird taking flight.

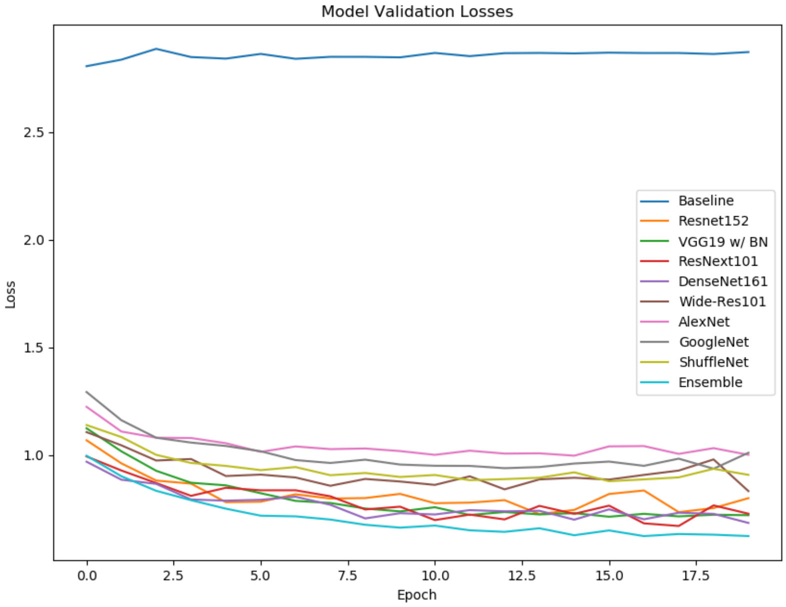


Figure 6: Validation Loss of Transfer Models, Baseline, and Architecture

8 Discussion

8.1 Model Performance

8.1.1 Abstract Classification

One of the hypotheses that is tested is the neural network's strength in extracting abstract patterns from data. Although our model performs decently, the result is inconclusive. This is because a majority of the accuracy in the classification problem relies on the pre-trained model's prior learning on large general image datasets such as ImageNet. As such, non-ambiguous categories containing distinct objects or features such as animals, clowns, blood, holes, and needles perform extremely well (¿ 90% recall), whereas more abstract triggers such as death and tight spaces perform worse (60%).

8.1.2 Ensemble Hypothesis

Although computationally expensive, ensembling multiple classifiers - of transfer-learned models - led to a clear and significant improvement over any individual model. This suggests that some (if not all) of the transfer-models extract slightly different features and representations from the same image.

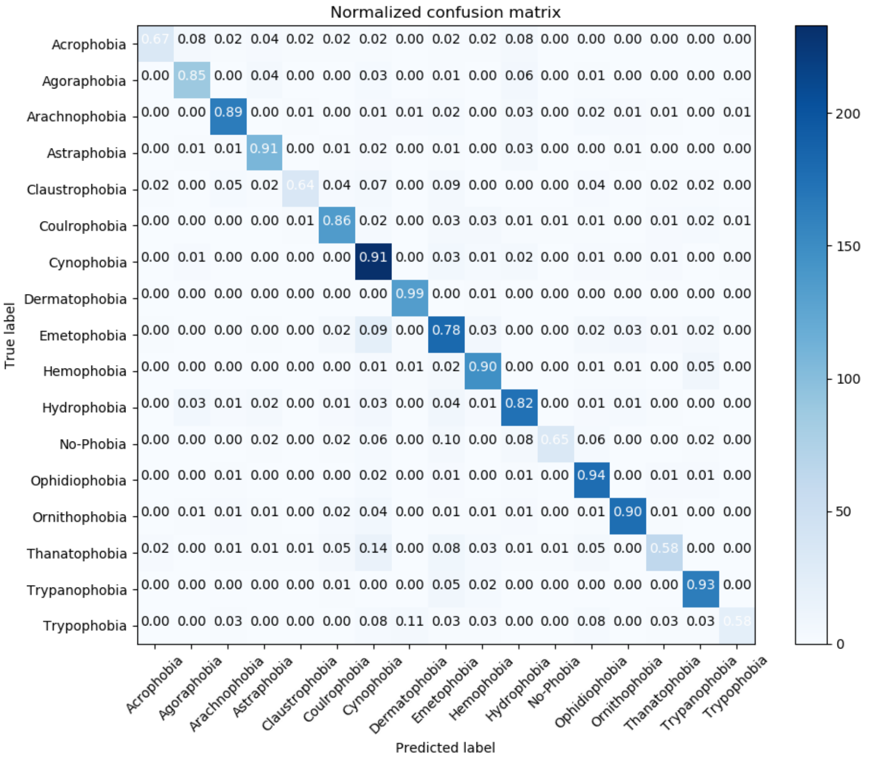


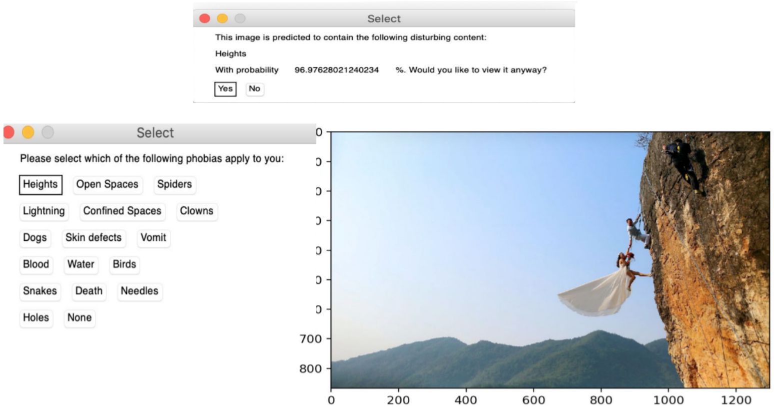
Figure 7: Normalized Confusion Matrix

8.1.3 Dataset Size Impact

The per-class accuracy of the model is highly correlated with the size of the phobia's dataset. Upon inspection of Figure 7, there most likely exists a causal relationship between small dataset size and poor performance, as the 5 lowest per-class accuracy classes also have the 5 smallest dataset sizes. Future improvements should consider gathering more data, and using data augmentation techniques (discussed later) to increase small dataset sizes.

8.2 Light Model Experimentation

Computational time was not considered during the design of the model architecture; however, it becomes extremely important when considering practical usage of an image-filtering system. By comparing the results of the final model to the individual transfer-learned models, there is only a few percentage im-



Select

This image is predicted to contain the following disturbing content:

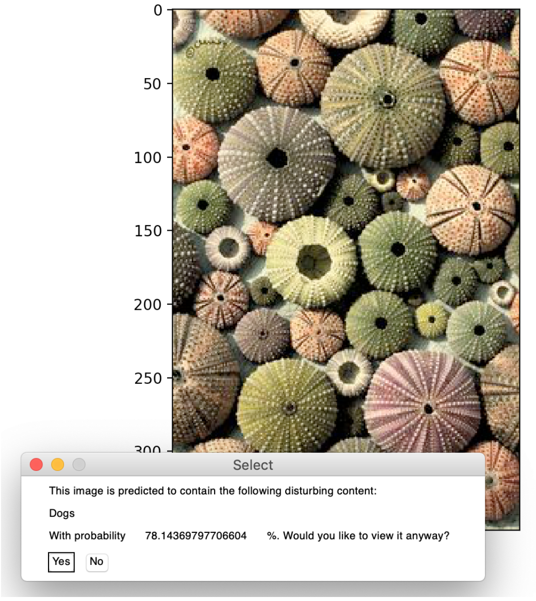
Heights

With probability 96.97628021240234 %. Would you like to view it anyway?

Select

Please select which of the following phobias apply to you:

Figure 8: GUI and Example Model Prediction



Select

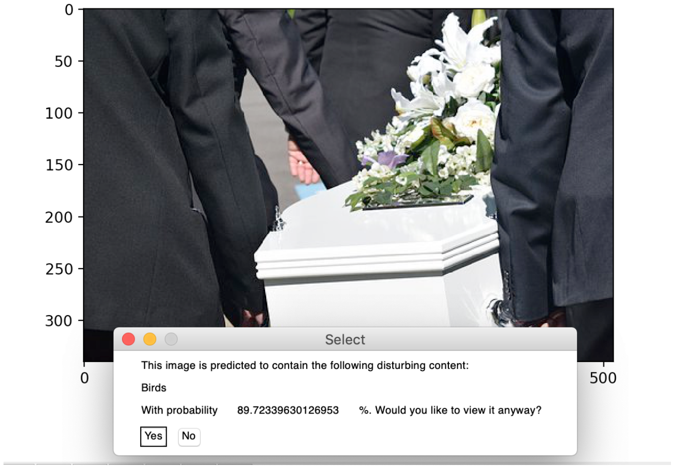
This image is predicted to contain the following disturbing content:

Dogs

With probability 78.14369797706604 %. Would you like to view it anyway?

Figure 9: Trypophobia trigger misclassified as a dog

provement in testing accuracy and recall despite increasing the complexity and number of parameters by nearly an order of magnitude. As such, it may be more desirable to look at more sophisticated techniques such as boosting (to target poorly performing classes), or drop-out (to reduce over-fitting) instead of ensembling more networks.



This image is predicted to contain the following disturbing content:

Birds

With probability 89.72339630126953 %. Would you like to view it anyway?

Figure 10: Thanatophobia trigger misclassified as a bird

8.3 Future Work

8.3.1 Data Augmentation

Due to the large variance in the representations of visually abstract phobias, conventional data augmentation may not be as useful as simply collecting more data. Regardless, data augmentation can still be used in future iterations of this project to reduce disparities between the sizes of different subsets. Techniques which may be useful in this case include:

- Image Rotation
- Cropping
- Explicit Colorization of Grey-scale Images

9 Ethical Framework

From a media platform user's perspective, or from their parent's perspective in the case of a child, this classification system clearly benefits autonomy and beneficence, by allowing users/ parents to have greater control over what content they view while protecting them from phobia triggers. It does contain, however, risk from the standpoint of non-maleficence; if the model mistakenly flags an image as a trigger, the user may miss out on something they would have actually wanted to view. There is also a bigger ethical risk in terms of non-maleficence from the perspective of content creators, as if enough people avoid content flagged as triggering, then it may become harder for people creating certain types of content to get exposure.

For the platforms that use this algorithm, this model being effective would be a positive in terms of non-maleficence and justice. Media platforms using such an image filtering system would protect themselves from public backlash regarding phobia-triggering photos being shared to all users.

References

[1] Facebook twitter violence ai. Undark. [Online]. Available: https://undark.org/2017/06/26/facebook-twitter-violence-ai/

[2] Facebook's failed ai showcases the dangers of technologists running the world. Forbes. [Online]. Available: https://www.forbes.com/sites/kalevleetaru/2019/03/22/facebooks-failed-ai-showcases-the-dangers-of-technologists-running-the-world

[3] Algorithms fail to filter graphic content from youtube kids. EdScoop. [Online]. Available: https://edscoop.com/algorithms-fail-to-filter-graphic-content-from-youtube-kids/