Final Project: Dendrogram Loss Experiment

IL181.011 - Professor Patrick Lam

Guy Davidson

December 21, 2018

1 Introduction

As previously outlined in the grant proposal (assignment), I believe there is merit to investigate machine learning approaches that do not treat all mistakes as equally plausible, hopefully paving the way toward models that reason and learn more as humans do by *erring* more as people tend to Lake et al. (2017). While such an approach appears sensible across a variety of tasks, image classification is one that appears to conform to the Goldilocks principle, being neither sufficiently trivial as to be considered 'solved' (although the literature is trending in that direction), nor being impossibly difficult for state of the art methods. In image classification, the model is presented with an image at a time and asked to predict to which of several classes (or labels) it belongs. A picture of a cat lying on a couch, for instance, might be categorized as 'cat,' and one of a plane taking off as 'airplane.' Notably, there is no attempt to locate (and label) additional objects, or holistically describe the contents of the input; just to predict the correct label. Within human cognition, some types of mistakes are likelier than others. Let us consider conceptual mistakes, as opposed to visual illusions and objects with potentially confusing appearance, such as cow-shaped mugs (Bracci et al. 2018). A parent teaching a child basic vocabulary might not be surprised to see them confuse a cat for a dog, or fail to distinguish between a car and a truck; but mistaking a ship for a deer, or an airplane for a bird, would be far more worrisome[1]. This notion is lacking from current image classification approaches, which currently consider all mistakes to be equally harmful, and this paper will investigate a particular approach to operationalize it.

The current state of the art models for image classification are all deep learning models, and training them broadly follows the same pattern: first, present an image (or batch of images), and pass them forward through the network, outputting a prediction for each label (the maximal of whom is the label assigned to the image). Based on these per-label predictions, we use a loss function to assign a loss score to each image: the better the prediction, the lower the loss score is. The network's goal is to minimize this loss score. To do so, we take the gradient of this loss score with respect to the per-class predictions, and then continue the backward pass through the model, using the chain rule to compute a gradient with respect to each layer. An optimizer then updates the weights in the model according to the gradients, and we proceed to the next batch of examples.

This work will proceed as follows: first, we will describe the prevalent image classification loss function, the cross-entropy, noting how it considers all mistakes on equal footing. We will then suggest a particular way to conceptualize the distance between different image classes, and propose two other types of loss functions which can be adapted to weigh conceptually distant mistakes more heavily. Next, we will provide the results of experimenting with several such loss function variants on a relatively simple convolutional neural network model. We will then discuss the (mostly negative) results, providing several hypotheses attempting to explain the results and suggest further experiments to investigate each of these new hypotheses.

[1] All examples taken from the CIFAR-10 classes: airplane, auto, bird, cat, deer, dog, frog, horse, ship, truck.

2 Loss Functions and Implementation Details

The most common loss function chosen used to train classification models is the cross entropy between the prediction and the correct response, whose log-likelihood is $L(x, y) = -\sum_i y_i \log(x_i)$, where x is a vector of the predicted probabilities, y is a one-hot vector with $y[l] = 1$, and l is the correct label. In the case of image classification, since there is a single correct answer (for which all of the probability mass should be assigned, $y_i = 0 \forall i \neq l, y_l = 1$), this translates into the negative log of the probability assigned to the correct class. This minimizes at zero when the entire probability mass is assigned to the correct class, and otherwise returns a positive value. Notably, since the ground truth requires all mass to be placed in a single label, there is no contribution to the loss from masses assigned to incorrect labels (other than by not being assigned to the correct response). This formulation would not allow to differentially penalize mistakes to more conceptually distance classes and therefore makes for a poor candidate to modify.

Before we discuss the particulars of the loss functions, we must first consider how to relate between the different classes. While there are many approaches to this problem, in this paper, we focus on the notion of representing the different labels in a dendrogram, a diagram representing tree-based information. Starting from a root node, each node signifies a category, either a class in the dataset (such as 'cat' or 'ship') or a non-class category (for instance, 'mammal' or 'transport'). *Figure 1* represents the dendrogram used when running the experiments detailed below, and *Figure 2* an alternative, alphabetically-based dendrogram used to attempt to assess the effect of the choice of the dendrogram. Should these methods prove to have merit, there is substantial room for further investigation into the choice of dendrogram itself, and how to treat additional relationships, such as 'airplane' and 'bird' being related by both being airborne. For the methods described below, we need a distance matrix between every pair of categories. Using a dendrogram, this was computed using the Floyd-Warshall algorithm implemented in networkx. In principle, any suitable similarity matrix should suffice, such as a confusion matrix from human attempting to classify such images.[2].

2.1 MSE

The first option investigated involved the mean squared error (MSE) loss. It is not often used for classification, but it does penalize overall output units (label probabilities). This loss is coupled with a softmax activation, in order to map each output to a probability. The default formulation of the MSE is as follows: $L(x, y) = \frac{1}{n}\sum_i (x[i] - y[i])^p$, where x is a vector of the predicted probabilities, y is a one-hot vector with $y[l] = 1$, l is the correct label, and n is the total number of labels, the dimensionality of x and $y4$. To augment it with distances, we weigh each component of the loss by the distance between it and the correct class: $L(x, y) = \frac{1}{n}\sum_i d_{il}(x[i] - y[i])^p$, where d_{il} is the dendrogram distance between i and l. This means that additional mass placed further away on the dendrogram would be penalized further, theoretically producing the desired effect. Note that by default, the distance between each class and itself would be zero. To account for that, we add one to the entire distance matrix; setting the diagonal to one could have also worked.

2.2 Hinge/Multi-Margin

The hinge (or multi-margin) loss is traditionally used to fit support vector machines (SVMs) in multi-class settings and features the next property that it attempts to produce a margin between classes. Its default formulation is $L(x, y) = \frac{1}{n}\sum_i \max(0, m - x[y] + x[i])^p$ for input predictions x, correct label y, n different classes, a margin of m, and a power p. attempting

[2] **#cognitiveai:** Described a paradigm to translate conceptual knowledge to distance measures.

to create a maximal margin between each pair of classes. We consider two formulations of the hinge loss with dendrogram information. The first ('HingeDendrogramLoss'), weighted the mistakes similarly to the MSE loss described above: $L(x, y) = \frac{1}{n} \sum_i d_{iy} \max(0, m - x[y] + x[i])^p$, where $d_i l$ is the dendrogram distance between i and y. The same consideration with adding one to the entire distance matrix holds here as well. The second formulation ('HingeDendrogramMarginLoss') treats it differently, setting the margin to be the distance between the classes: $L(x, y) = \frac{1}{n} \sum_i \max(0, d_{iy} - x[y] + x[i])^p$. Note that in this case, we do not add one to the entire distance matrix, as enforcing a margin of 0 between the correct class and itself is perfectly sensible. This setup has the elegant conceptual interpretation as enforcing a wider margin between image categories that are further apart conceptually[3].

3 Experiment Details

3.1 Technical Details

All loss functions were evaluated on the same convolutional neural network model. There is nothing particularly unique or meaningful about this choice of model, other than it being a fairly standard and simple convolutional architecture I have implemented for another project. *Figure 3* illustrates the overall model architecture. The model included two main components: a convolutional image-processing segment, followed by a fully-connected (multilayer perceptron, MLP) segment. The convolutional part of the model was comprised of four layers, each of which used a different number of 3x3 (pixel) filters, from 16 in the first layer to 64 in the last one. All layers were executed with padding of 1, a stride of 1, ReLu nonlinearities, batch normalization, and 2x2 max pooling, effectively halving the input size and doubling the receptive field between layers. These layers were also executed with a spatial dropout of 0.2 in order to aid generalization. Each of the four fully connected layers included 512 units, and used ReLu nonlinearities with a dropout of p = 0.5, other than the final, output layer, which had ten units (one for each class), and used a different activation based on the demands of the loss function.

All models were trained for 200 epochs of the 50,000 32x32 pixel images in the CIFAR-10 (Krizhevsky, 2009) training set, belonging to ten classes (footnoted above) and evaluated on the 10,000 image test set. The training set images were augmented using random cropping, horizontally flipping with a probability of 0.5, and normalizing each channel (red, green, and blue) independently to zero mean and a standard deviation of one. The test set is only normalized, as is standard. The Adam (Kingma & Ba, 2015) optimizer was used for all experiments, with an initial learning rate of 1e-3 and wight decay of 1e-4, and a scheduler allowing the model to halve the learning rate if the test set loss stagnates for ten consecutive epochs.

3.2 Model descriptions

- **Cross Entropy:** standard cross entropy loss using softmax activations.
- **MSE:** standard mean squared error loss using softmax activations.
- **MSE Dendrogram:** mean squared error with individual error terms weighted by their dendrogram distance from the correct label.
- **MSE Dendrogram (Alphabetical):** same as above, but using the alphabetical dendrogram rather than the categorical one.

[3] **#ailearning:** Discussed and modified several different loss functions to attempt to allow image classfication models to make mmore human-like errors.

- **Hinge (L1 SVM):** a hinge (multi-margin) loss with the default parameter settings, $p = 1, m = 1$, using log-softmax activations.
- **Hinge Squared (L2 SVM):** as above, but using the squared error, $p = 2$.
- **Hinge Dendrogram:** a hinge loss with the default parameter settings, weighting each error term by its distance from the correct class.
- **Hinge Squared (L2 SVM), Large Margin:** a hinge loss with $p = 2, m = 4.5$, selected as a fairly arbitrary example of attempting to enforce a larger margin.
- **Hinge Squared Dendrogram, Large Margin:** as above, weighting each error term by its distance from the correct class.
- **Hinge Squared Dendrogram (Alphabetical):** as above, but using the alphabetical dendrogram rather than the categorical one.
- **Hinge Squared Dendrogram Margin:** a hinge loss with $p = 2$, and using the distance from the correct label as the margin to enforce.

4 Results

Table 1 provides a summary of all model performance after 20 epochs, and *Figures 4, 5, and 6* compare the performance of all models on the test set. *Figure 4* plots classification accuracy, *Figure 5* the region under the receiver operating characteristics curve (ROC AUC, a more holistic measure than raw accuracy), and *Figure 6* the mean rank order assigned to the correct category (a metric we developed, providing one measure of how much (or little) mass is placed on the correct class). The results provide evidence for why cross-entropy might be the standard loss function, as it clearly outpaces the pack. Equally clearly, the margin-based dendrogram approach appears least effective, and broadly, the MSE-based approaches are outperformed by the ones drawing on the hinge loss. *Figures 7, 8, and 9* show top-2, top-3, and top-4 accuracy (is the correct class in the top-k of responses?), in which the dendrogram margin performs much more in line with the rest of the dendrogram approaches, but the cross-entropy continuing to outperform the rest. *Figure 10* plots another measure I devised, in order to investigate the effectiveness of the dendrogram based approach. That plot computes the mean dendrogram distance of mistakes from the correct class -- in other words, whenever an incorrect category is assigned the most probability mass, how far away this prediction is on the (conceptual) dendrogram from the correct label[4]. In this plot, where lower is better (as it implies close to the correct class), the dendrogram-margin approach performs *better* than all other approaches, and most loss functions cognizant of the true dendrogram produce better scores than the ones that are not. The alphabetical dendrogram loss also performs poorly on this metric, suggesting it captures well the original notion of 'making better mistakes.'

By most measures, the MSE-based approaches perform the worst, and cross-entropy outperforms all others; therefore we will omit those from further discussion. Comparing all hinge-based approaches, *Figures 11-17* reproduce the previous visualizations, plotting only results from loss functions deriving from the multi-margin loss. Within those, it is clear that the dendrogram margin approach performs substantially worse than the others. This result makes us wonder if there is any literature suggesting using different margins between different classes, irrespective of what these margins are derived from. Furthermore, it appears that the approaches without the conceptual dendrogram (both the default multi-margin loss, and the one employing the

[4]One could also examine the mean distance from the correct class, rather than the distance of the prediction from it, but I reckon I have enough plots already.

alphabetical dendrogram) perform better than the ones with it. This curiosity points at another potential future direction, to perform an ablation study over different dendrograms (or other distance matrices), evaluating how much of the effectiveness of such a loss function is derived from the particular matrix specified (which I would have performed had these approaches outperformed traditional methods, such as the cross-entropy).

5 Discussion

The results are not as promising as we had hoped. While on the metric we devised, which explicitly evaluates the quality of mistakes made, the dendrogram-based approaches outperform the standard ones, in all traditional metrics (namely, accuracy and ROC AUC), none of my approaches cognizant of the conceptual dendrogram reach better performance than cross-entropy does. While slightly disheartening, we do not believe this spells doom for the idea. Until proven otherwise, the null hypothesis must remain that this idea does not work. However, we can hypothesize several other reasons that might have contributed to the lack of results:

- **Underexpressive model:** the model I chose to evaluate these approaches with is fairly simplistic, and while its architecture derives from the AlexNet (Krizhevsky et al., 2013) model that won the 2013 ImageNet (Russakovsky et al., 2015) classification challenge, it is substantially smaller. This makes it easier and faster to train, allowing to run such a high number of experiments within the last two weeks[5]; but it also means its capacity to learn such relationships are limited. Note that the cross entropy loss model peaked around 75% accuracy, whereas state of the art on CIFAR-10 is over 90%.

  We suspect the limited expressive capacity of this model might have forced optimization to 'choose,' in some sense, between learning to classify at peak accuracy and between learning to make better mistakes. Perhaps a larger, more powerful model could learn both. To investigate these, we can repeat these experiments with a newer model, such as one of the ResNet variants which achieve over 90% accuracy[6] on the same dataset[7].

- **Small data and few classes:** The CIFAR-10 dataset used in these experiments is comprised of fairly small images (32x32 pixels each), belonging to a mere ten categories. The small images provide little information to learn complicated visual relationships over -- on the one hand, allowing to execute the model faster and using less memory, but on the other, limiting the richness of the learned representations. I find it plausible that the additional information regarding the quality of mistakes would be helpful in regimes where it is possible to learn richer representations (by virtue of being provided more data), and in settings where the similarities between classes might be stronger (for instance, the ImageNet competition has different dog breeds as separate classes[8]). Investigating this hypothesis further is simple: try a larger dataset with more classes.

- **Overly simplistic task choice:** Image classification is among the simplest visual tasks currently investigated in the field, and there has been significant progress made in state of the art performance over the past five years. While computer vision researchers use many tasks to investigate different properties, one that strikes as particularly relevant for

[5] A single epoch (training the model on all 50,000 training set images and testing it on all 10,000 test set ones) took my model about 30 seconds to run; it takes the smallest ResNet model, ResNet18, about five minutes, or approximately 10x as long.

[6] https://github.com/kuangliu/pytorch-cifar

[7] **#ailearning:** Described why an underexpressive moedl might fail to learn sufficiently powerful representations. The "overly simplistic task choice" hypothesis also represents an understanding of why a simple task might not urdge toward learning powerful representations.

[8] http://image-net.org/synset?wnid=n02084071

making better mistakes is visual question answering (VQA; Agrawal et al., 2015). In this paradigm, a model is presented with an image (say, a horse galloping through a pasture), and is asked a question ("Which animal is running through a field?"). In this setting, one could strongly argue that different mistakes carry substantially more weight: when asked about animals, it makes little sense to answer anything that is not one, and in the context of a field, aquatic species make little sense either. By virtue of its difficulty, this task requires developing richer representations than image classification does, potentially allowing for additional to be derived from the increased information provided by one of the loss functions proposed[9].

- **Dendrogram and loss function misspecification:** Another potential culprit is the choice of dendrogram and specification of the loss function. As mentioned above, we intended to do some ablation studies (beyond the alphabetical dendrogram) if this method proved promising, but the negative results meant my time was better spent investigating alternative formulations of the loss. It plausible that either an alternative distance matrix or a completely different formulation of the loss function would prove more effective. To further this hypothesis, we would have to design and benchmark additional formulations of the loss functions and consider alternative distance matrices or dendrograms. One novel way to devise an alternative distance specification is to run an experiment on human subjects, asking them to classify such images, in a manner that makes the task actually hard on humans (either classifying over many labels, beyond the ten in CIFAR-10, or perhaps allowing only little time to view each image). We could then derive the distances to be used from the confusion matrix generated by human subjects' classifications. This approach would genuinely allow the model to make human-like errors, avoiding the decision-making required to derive the dendrogram.

6 Conclusion

This work investigated an idea previously proposed, implementing it using several different formulations, running several experiments investigating these variants, reporting the outcomes, and discussing potential hypothesis for the mostly negative results obtained. On a personal level, I found this investigation quite rewarding, even if not as fruitful as I could have hoped. It was interesting to explore different implementations of the same concept I had in mind, run several experiments with them, and try to reason about the results. I remain excited to continue investigating this idea in the future, as I have not lost all hope in it, and the experience of persevering to work on a problem even when initial approaches prove suboptimal is one I cherish. I am also excited to dive into alternative explorations of this idea, such as explicitly examining the representations learned by models trained with these loss functions, and seeing if they behave substantially differently than representations learned with standard approaches to the loss function.

7 Code

The main Colaboratory notebook with my work is available here: https://colab.research.google.com/drive/1QUPsY5B_vET0-WYiCLppRq4YVU7CcKh3.

[9] **#cognitiveai:** Discussed why a choice of task is meaningful to the representations a model learn. Humans learn such flexible and robust representations for knowledge because we wield our knowledge over many different goals and objectives, which perhaps should nudge us toward multi-task learning if we wish to develop machine which represent more like we do. The next hypothesis, regarding dendrogram specification, suggests a method to capture more human-like mistakes into the distance matrix used by the loss functions devised.

8 References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2015). VQA: Visual Question Answering www.visualqa.org. In *International conference on computer vision (iccv)*. arXiv: 1505.00468v6. Retrieved from www.visualqa.org

Bracci, S., Kalfas, I., & de Beeck, H. O. (2018). The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *bioRxiv*, 228932. doi:10.1101/228932

Kingma, D. P. & Lei Ba, J. (2014). *ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION*. arXiv: 1412.6980v9. Retrieved from https://arxiv.org/pdf/1412.6980.pdf

Krizhevsky, A. (2009). *Learning Multiple Layers of Features from Tiny Images*. Retrieved from https://www.cs.toronto.edu/%7B~%7Dkriz/learning-features-2009-TR.pdf

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2013). *ImageNet Classification with Deep Convolutional Neural Networks*. Retrieved from http://code.google.com/p/cuda-convnet/

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. doi:10.1017/S0140525X16001837

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. In *Ijcv*. arXiv: 1409.0575v3. Retrieved from http://image-net.org/challenges/LSVRC/

9 Figures

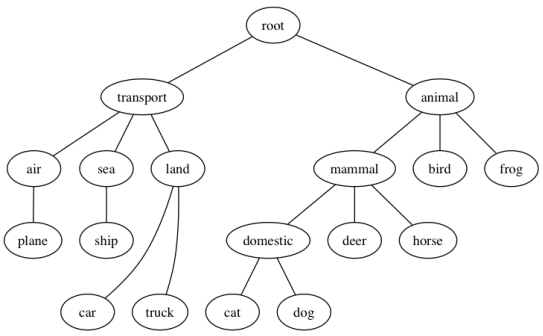

Figure 1: Categorical dendrogram used in experiments.

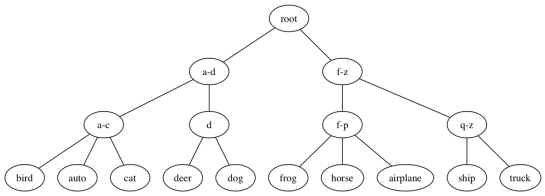

Figure 2: Alphabetical dendrogram used in experiments.

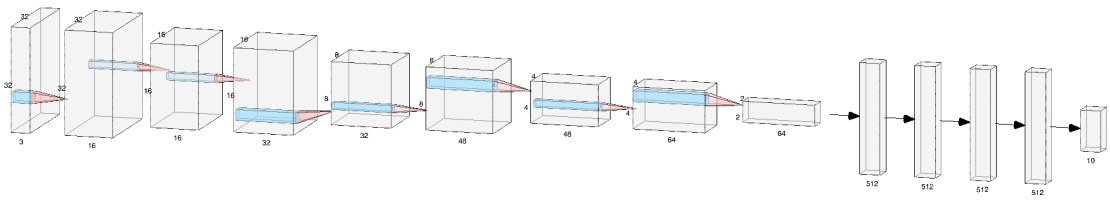

Figure 3: Schematic of model used in experiments.

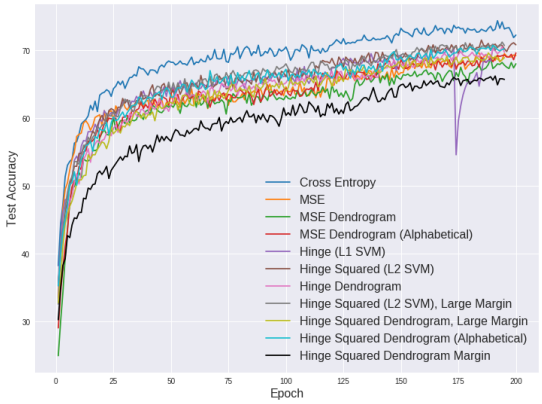

Figure 4: Test set classification accuracy over all models.

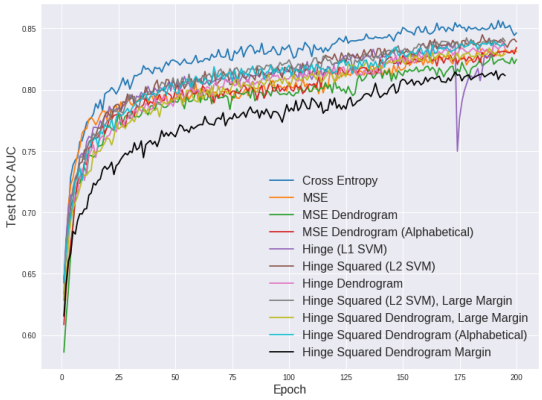

Figure 5: Test set ROC AUC over all models.

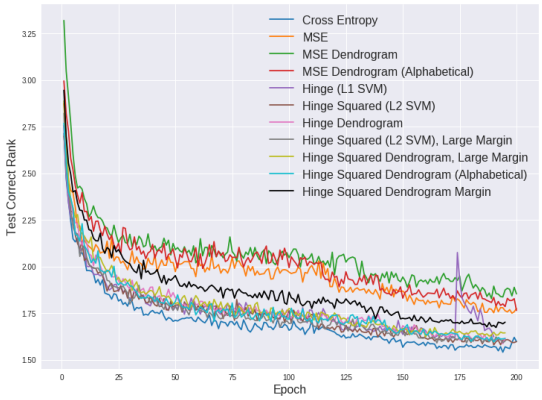

Figure 6: Test set correct answer rank over all models.

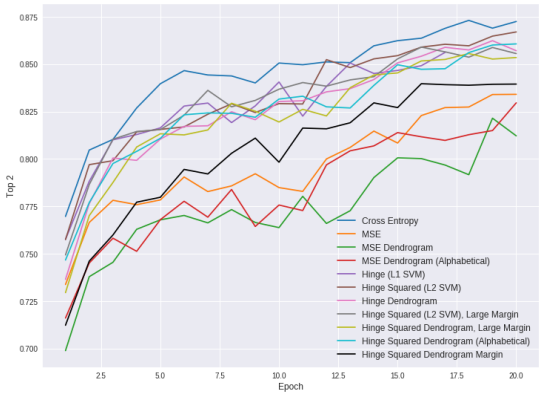

Figure 7: Test set top-2 classification accuracy over all models.

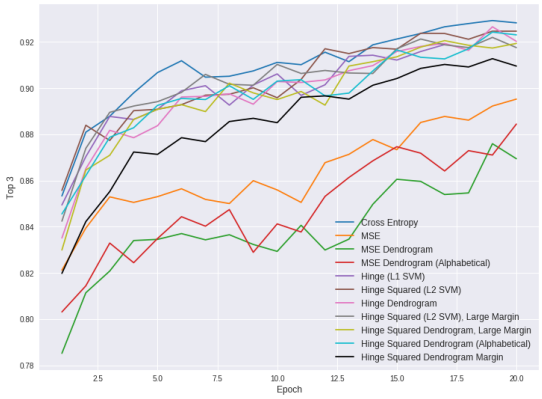

Figure 8: Test set top-3 classification accuracy over all models.

Figure 9: Test set top-4 classification accuracy over all models.

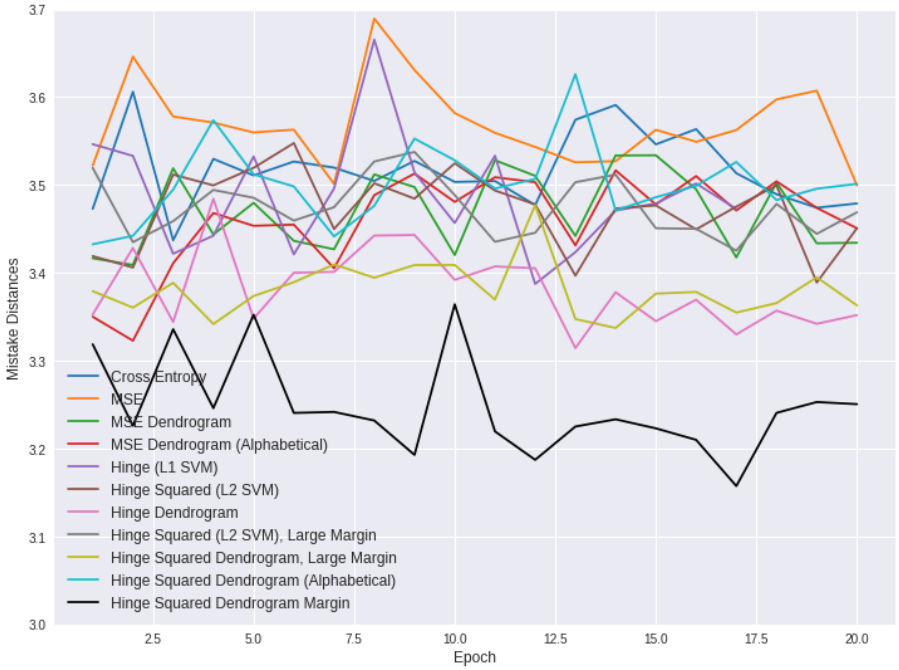

Figure 10: Test set mistake dendrogram distance over all models.

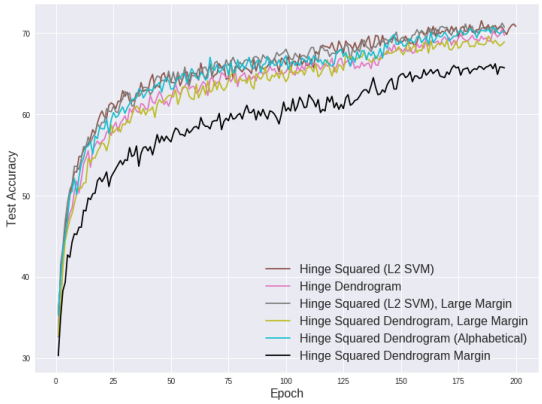

Figure 11: Test set classification accuracy over hinge-loss models only.

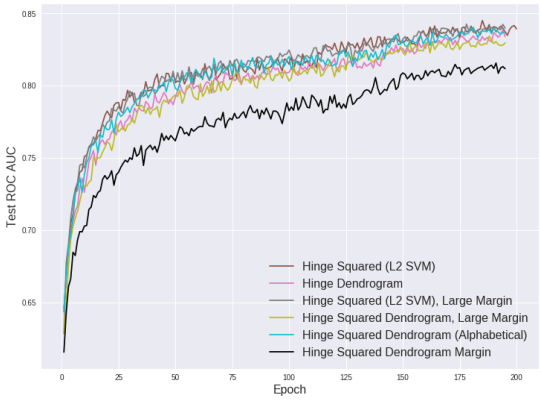

Figure 12: Test set ROC AUC over hinge-loss models only.

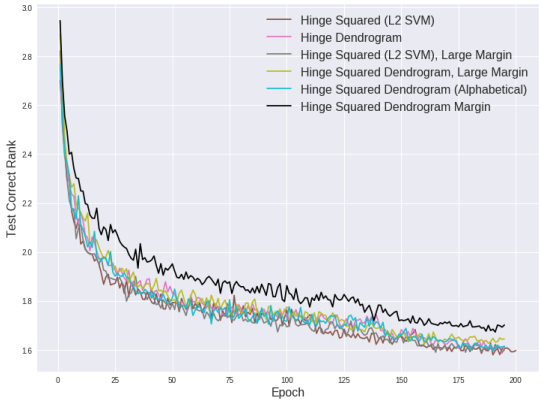

Figure 13: Test set correct answer rank over hinge-loss models only.

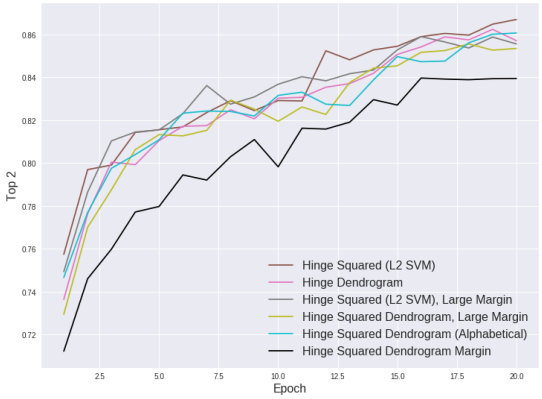

Figure 14: Test set top-2 classification accuracy over hinge-loss models only.

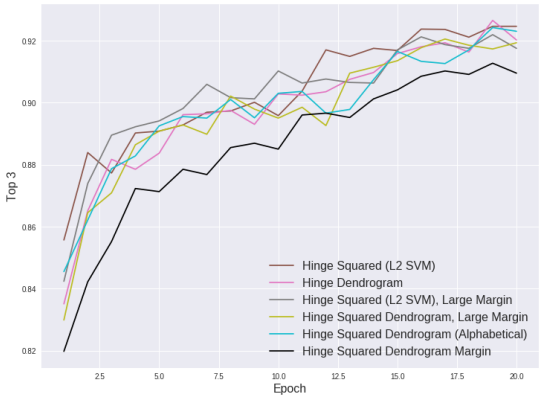

Figure 15: Test set top-3 classification accuracy over hinge-loss models only.

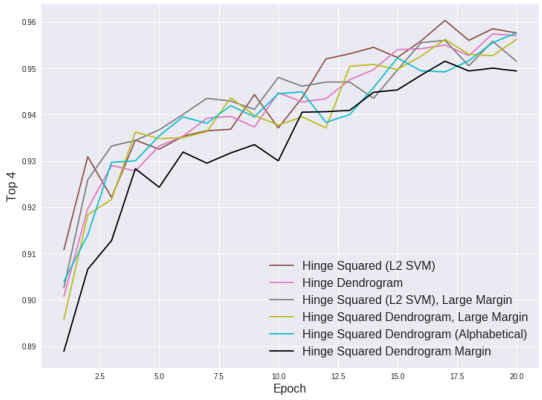

Figure 16: Test set top-4 classification accuracy over hinge-loss models only.

Figure 17: Test set mistake dendrogram distance over hinge-loss models only.

10 Tables

Table 1: Experiment Results Summary

Model Name	Accuracy	ROC AUC	Correct Result	Top 2 Acc.	Top 3 Acc.	Top 4 Acc.	Mistake Dend. Distance
Cross Entropy	72.43	0.847021	1.5911	0.8725	0.9282	0.9595	3.47904
MSE	69.3	0.83267	1.7522	0.8341	0.8952	0.931	3.4995
MSE Dendrogram	67.9	0.823794	1.8526	0.8122	0.8694	0.907	3.43431
MSE Dendrogram (Alphabetical)	68.5	0.827536	1.8197	0.8296	0.8844	0.9182	3.45082
Hinge (L1 SVM)	70.3	0.837807	1.6229	0.8562	0.9188	0.9519	3.47278
Hinge Squared (L2 SVM)	70.4	0.838646	1.6082	0.867	0.9246	0.9576	3.45086
Hinge Dendrogram	69.7	0.834273	1.6131	0.8571	0.9202	0.957	3.35198
Hinge Squared (L2 SVM), Large Margin	70.3	0.838681	1.611	0.8556	0.9175	0.9515	3.46899
Hinge Squared Dendrogram, Large Margin	68.7	0.828875	1.6501	0.8535	0.9193	0.9562	3.36305
Hinge Squared Dendrogram (Alphabetical)	70.3	0.837823	1.6157	0.8607	0.923	0.9576	3.50136
Hinge Squared Dendrogram Margin	65.6	0.811673	1.6987	0.8395	0.9095	0.9494	3.25068