CSCI 567 - Project Proposal
Exploring Complex Regularization Techniques on Image Classification and Sequence Labeling

Abhinav Gupta	Akshita Kapur	Hrishikesh Thakur	Shreyas Malewar
agupta67@usc.edu	kapuraks@usc.edu	hthakur@usc.edu	malewar@usc.edu

1 What research question are you trying to answer?

Amongst the topics listed, our team was very fascinated by the sub-topic summary for (2a). This sub-topic talks about the effects of variation in models, network structure, datasets, and most importantly, regularization on the output of a research experiment. We want to take the oldest and most reliable path of trial and error to understand the concept of regularization through this project.

2 Why is this question interesting to you?

In CSCI567 homework assignments, we had the opportunity to experiment with various *regularization methods* and noticed that even a tiny constant added to complex models during training can significantly improve performance and prevent overfitting. While it's often believed that copious amounts of data are generated in the modern age, a more analytical perspective reveals that not all available data is useful. For instance, in a tumor segmentation problem where data acquisition is costly, using a complex, high-performing model might lead to overfitting, ultimately reducing its effectiveness. This prompts us to explore different regularization techniques to prevent overfitting and effectively utilize all available data.

3 What kind of data are you collecting or what datasets will you use?

We have taken 2 vastly diverse datasets - Image Classification and Natural Language Processing. In addition to experimenting with these 2 different datasets, we also plan to scale down each dataset and re-run all experiments to observe and document the results obtained of using regularization on smaller size datasets. A brief description of the datasets:

1. **tiny-imagenet**
   (a) **Dataset Summary** - Tiny ImageNet contains 100,000+ images of 200 classes (500 for each class) downsized to 64×64 colored images. Each class has 500 training images, 50 validation images, and 50 test images.
   (b) **Data Feature Dimensions** -
      i. Image: A PIL.Image.Image object containing the image.
      ii. Label: an int classification label. -1 for the test set as the labels are missing. Check classes.py for the map of numbers and labels.
2. **nltk-brown + nltk-treebank + nltk-conll2000**
   (a) **Dataset Summary** - The combination of these 3 datasets gives us a large corpus of textual data that can be used for training a model that performs sequence labeling with a total size of 72,000+ tagged sentences. The nltk library takes the base dataset and performs tokenization to prepare it for the task of sequence labeling.
   (b) **Data Feature Dimensions** -

i. Input Sequence - A sentence in english.
ii. Output Sequence - POS tags of each word of the sentence.

4 What algorithms will you try?

From the tons of different regularization methods, we have selected 5 that will be applied to our 2 datasets in some combination. We will use a Classification model for the Image Classification task and a Sequence Labeling model for the NLP task. It is understood that in order to fit the model to these different datasets and regularization methods, cosmetic changes will be made to the models features and hyperparameters. A brief description of the different regularization methods:

1. **L2 Regularization** - modifies the loss function. Applied to both datasets.
2. **Data Augmentation** - modifies the data. For Dataset1 we plan to use **RandomErasing** - *RandomErasing* is concerned about removing and randomly adding information on the blank space, such as noise. For Dataset2 we plan to use **Random Synonym Replacement** - *Random Synonym Replacement* is concerned about removing and replacing with a synonym.
3. **MaxDropout** - modifies training approach. Applied to both datasets.
4. **Ensemble Regularization 1** - applying *RandomErasing* and *MaxDropout* together. Applied to Dataset1.
5. **Ensemble Regularization 2** - applying *RandomSynonymReplacement* and *MaxDropout* together. Applied to Dataset2.

5 What experiments and analysis will you run?

In our project, we aim to compare and analyze the use of different regularization methods on different types of datasets. We will apply the selected regularization techniques to the two datasets we've chosen. Additionally, we intend to experiment with combining different regularization methods to create ensemble regularization algorithms and check their impact on model performance. It is well known that regularization aids in mitigating overfitting, which often arises from insufficient data. Therefore, we intend to apply our regularization techniques to a subset of the two datasets and data sizes, to check if a certain regularization is good or the model is just performing well due to sufficient data.

6 What do you plan to finish by the midterm report

For the midterm report we plan to process our datasets and prepare the models that will be used for each of the tasks - *Image Classification* and *Sequence Labeling*. We also aim to start building the regularization methods for at least one of the two tasks by the midterm report.