# Named Entity Recognition using LSTM, GloVE

Harsh Karia
University of Southern California

## Task:

The task at hand is to perform named entity recognition using the CoNLL-2003 dataset using LSTM

## Data Preparation:

To prepare the data, we first form a vocabulary from the dev and train sets. We then tokenize the states as well as the joint vocabulary.

We use PyTorch's "Dataset" class to create a dataset that returns the tokenized sentence and states for each entry in the corpus and use the data loader in PyTorch to create a collate function that creates a dataset of a desired batch and pads the sentences less than the longest sentence in the batch with 0, as well as the tags.

We also account for class imbalance by computing label weights, inversely proportional to the number of times they appear within the dataset.

## Bidirectional LSTM Model

As a first task, we build a baseline bidirectional LSTM model. The architecture is as follows:

embedding dim 100
number of LSTM layers 1
LSTM hidden dim 256
LSTM Dropout 0.33
Linear output dim 128

We employ cross entropy loss as our loss function, use SGD, with weights computed above, as well as a cosine annealing rate of 1e-1 annealing every 5 epochs for a total of 240 epochs. Our batch size is 64.

Dev Data Results:

- Precision: 79.49%
- recall: 74.47%;
- F1: 76.90%

## GloVe Word Embeddings Model

Instead of learning the embeddings as above, we initialize the embedding layer using GloVe (100-d) embeddings. We compute the embeddings for every word in our vocabulary and create a frozen embedding matrix.

We account for uppercase words by appending a boolean mask to the embedding of the word, making it 101-d, and account for unknown words with a seeded random vector initialized with a mean of 0 and std. deviation of 1.

We train this using a weighted cross-entropy loss function with a batch size of 1 and a learning rate of 1e-3 (cosine annealing) every 20 epochs using stochastic gradient descent.

Dev Data Results:

- Precision: 88.50%
- recall:  90.02%;
- F1:  89.25%