Combining Image and Language to Predict and Understand the Usefulness of Yelp Reviews

David Z Liu (dzliu@stanford.edu)

Back Ground

- Predicting a reviews usefulness is important and challenging

- Knowing the usefulness of a review in advance, businesses can recommend high quality and fresh reviews to customers and gain business insight

Challenge

- For reviews with the same length, No obvious features that directly indicate the usefulness of a document

- Only a small amount of data have significant number of useful votes

Approach

- Cast the prediction problem to **Regression** and **Classification**

- **Make usefulness prediction by combing two neural networks**

  - **CNN** on an image associated with the review
  - Attention based **Convolutional RNN** on the text of the review

- **Analyze Attention weights** to gain insight into what make a review useful

Data Set

Using 10.2K Reviews
- Class 0: Reviews have 0 useful votes
- Class 1: Reviews have >9 useful votes
- A subset of images from the 200,000 business pictures provided by yelp

Link Image With Text

- Link a business photo to a review written about that business by match words in the image caption with words in the review text

- Pick the image with the highest Jaccard similarity

$$jac(x,y) = \frac{|x \cap y|}{|X| + |y| - |x \cap y|}$$
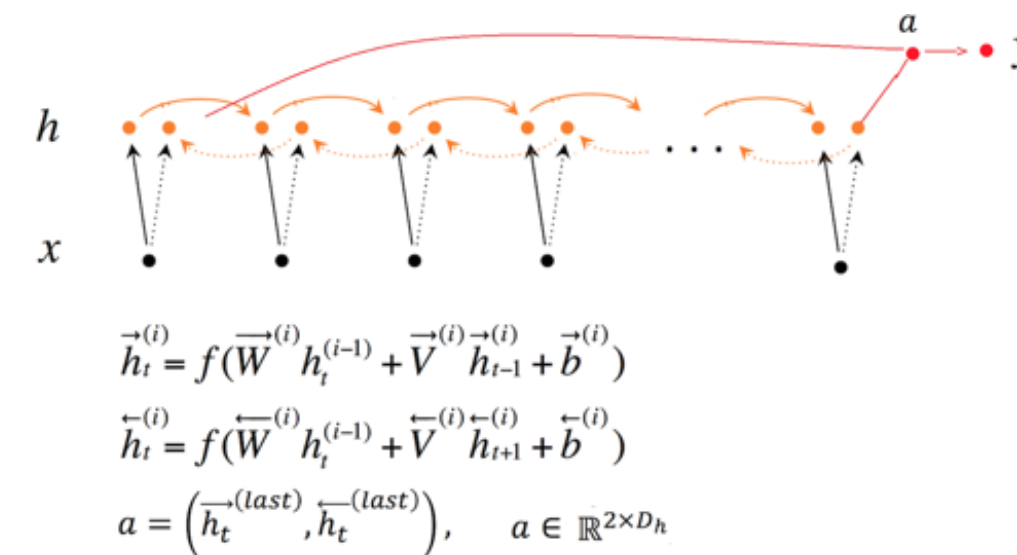
- **Input to the model:**
  - Written review to the RNN. Text are converted to 300 dimensional GloVe
  - Associated image to the CNN, converted to 64X64X3
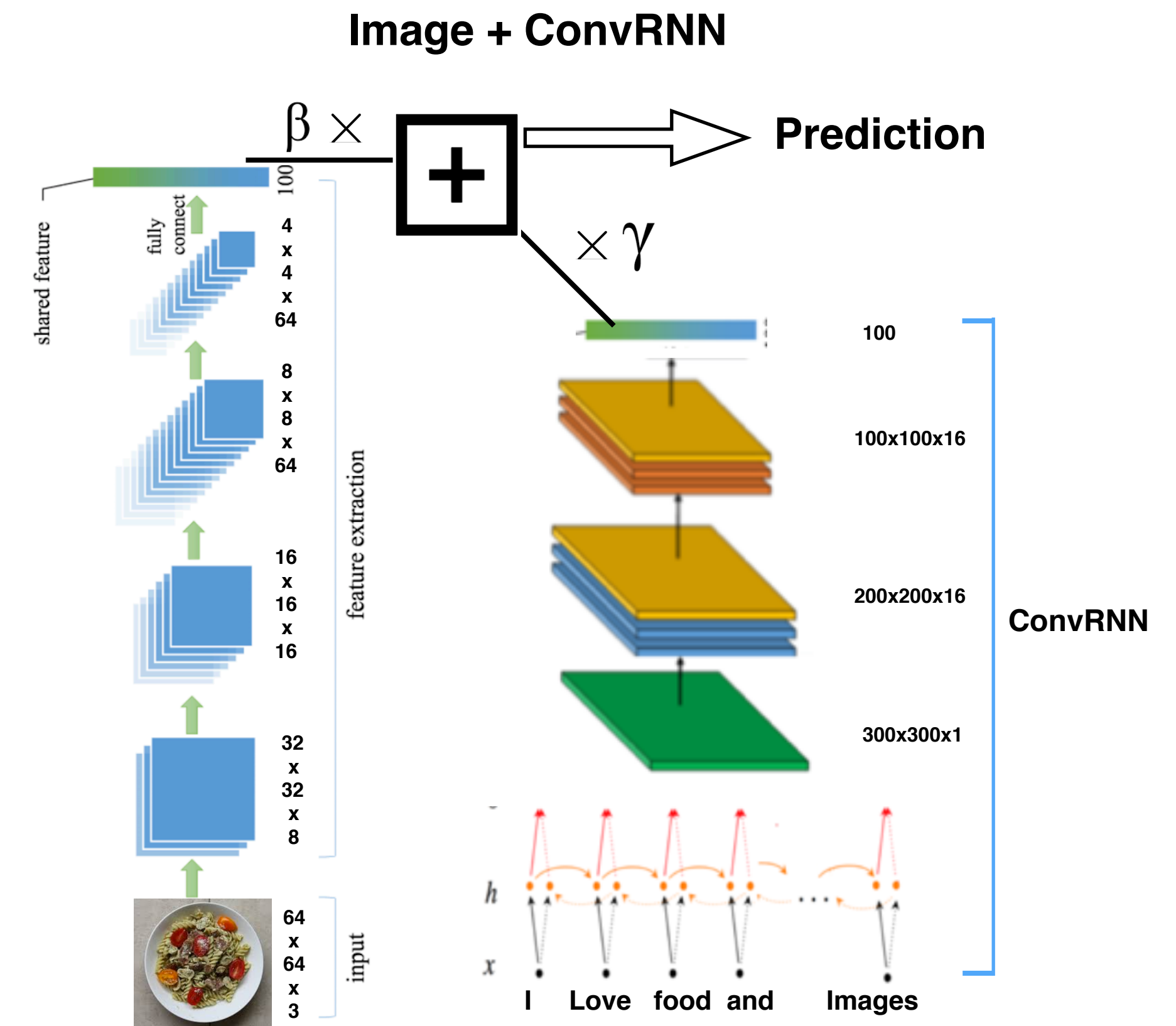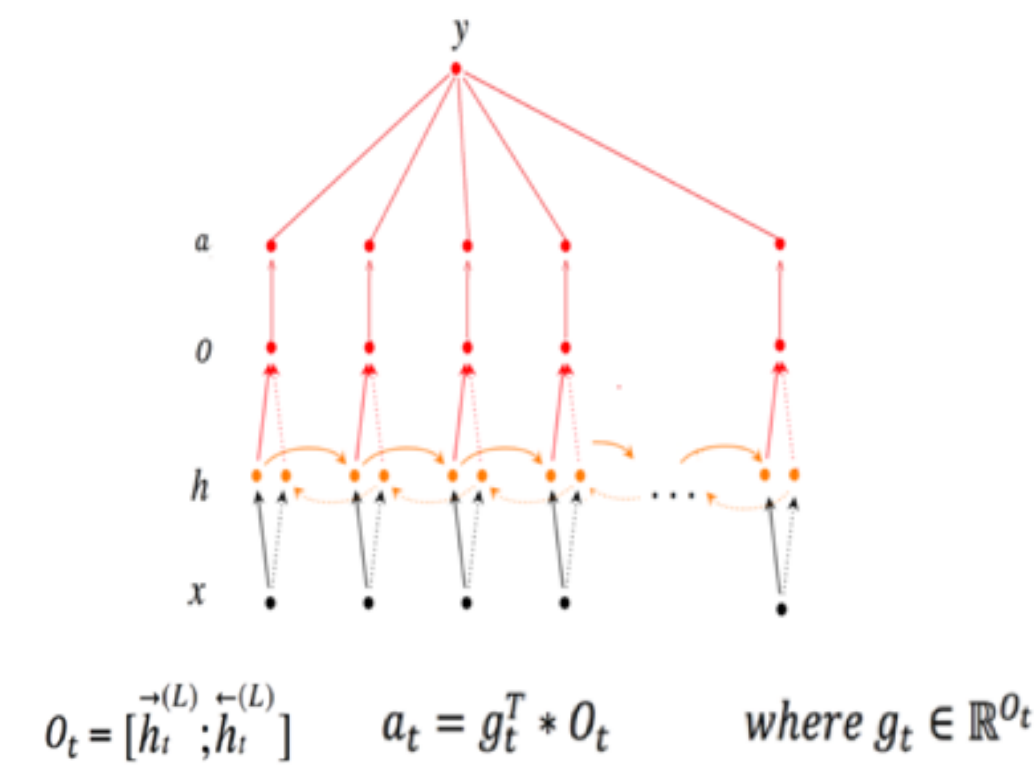
- **Output from the model:**
  - Classification or regression prediction
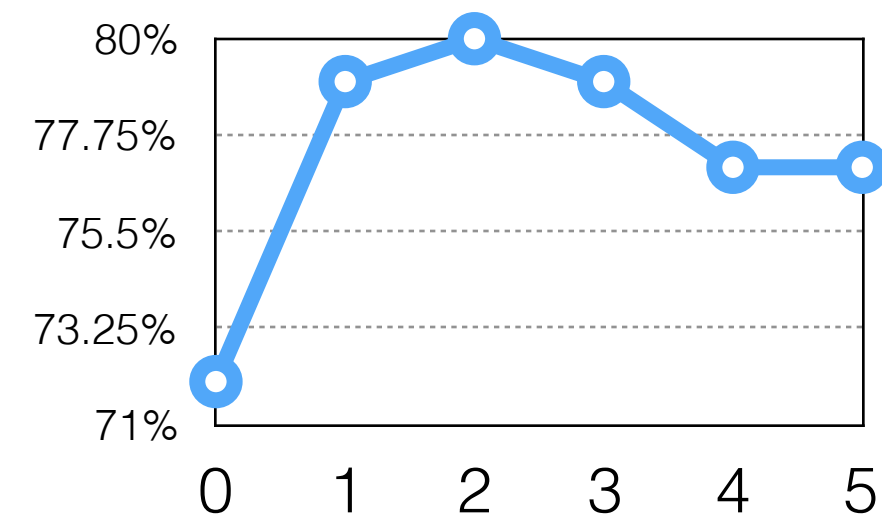
Model

Bidirectional RNN Without Attention

$$\vec{h}_t^{(i)} = f(\vec{W}^{(i)} h_t^{(i-1)} + \vec{V}^{(i)} \vec{h}_{t-1}^{(i)} + \vec{b}^{(i)})$$
$$\overleftarrow{h}_t^{(i)} = f(\overleftarrow{W}^{(i)} h_t^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{t+1}^{(i)} + \overleftarrow{b}^{(i)})$$
$$a = (\vec{h}_t^{(last)}, \overleftarrow{h}_t^{(last)}), \quad a \in \mathbb{R}^{2 \times D_h}$$

Bidirectional RNN With Attention

$$O_t = [\vec{h}_t^{(L)} ; \overleftarrow{h}_t^{(L)}] \qquad a_t = g_t^T * O_t \qquad where\ g_t \in \mathbb{R}^{O_t}$$

Image + ConvRNN
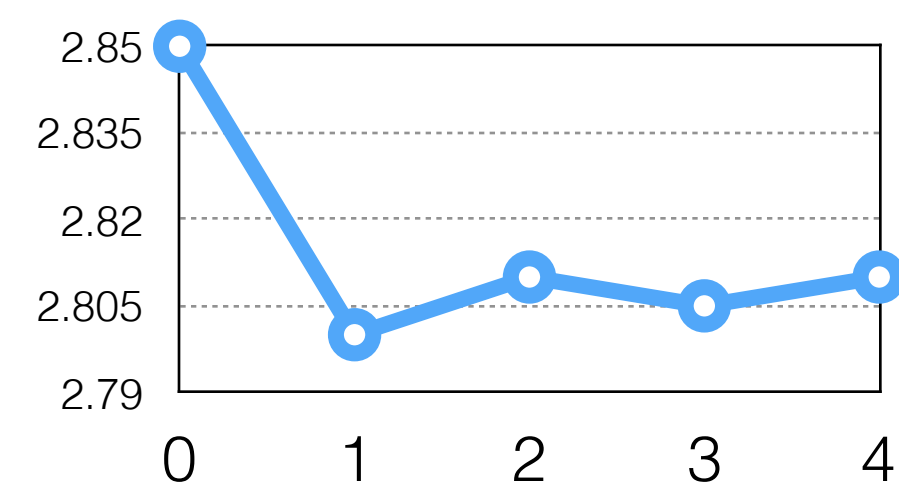

Results

# of ConvRNN Layer V.S. Classification Accuracy
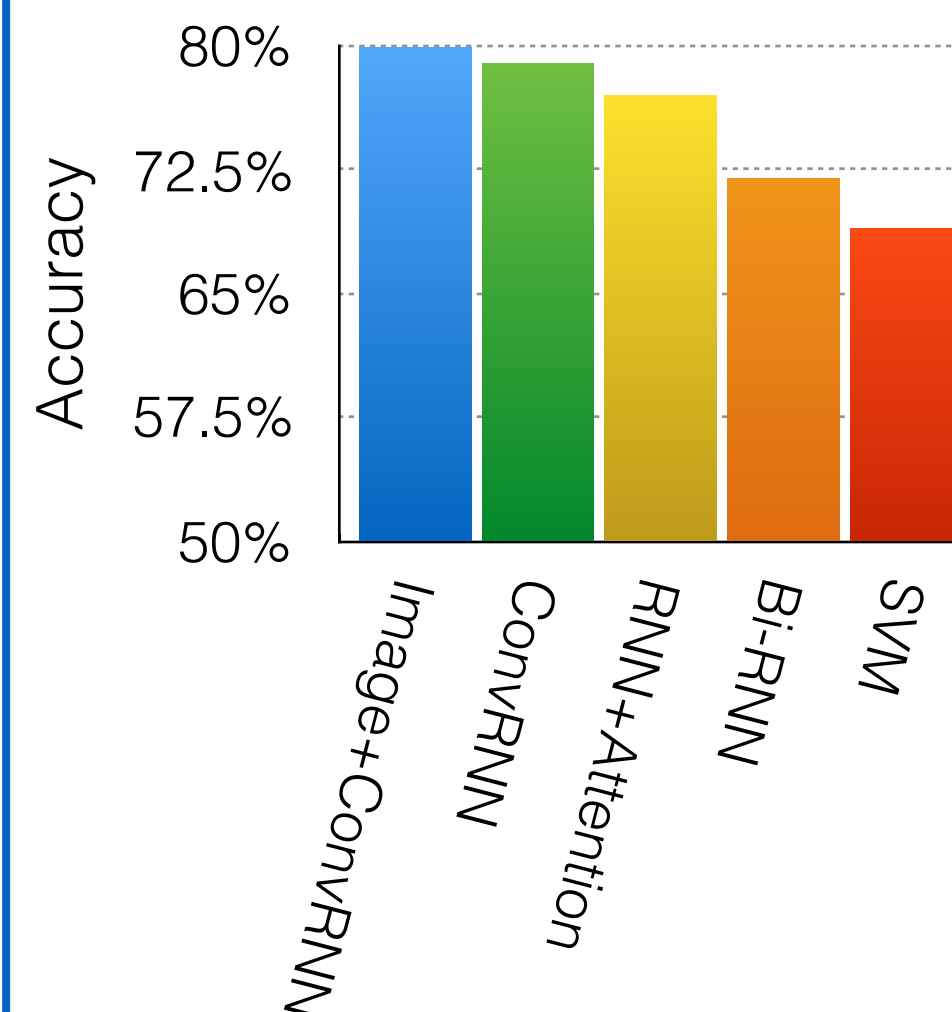

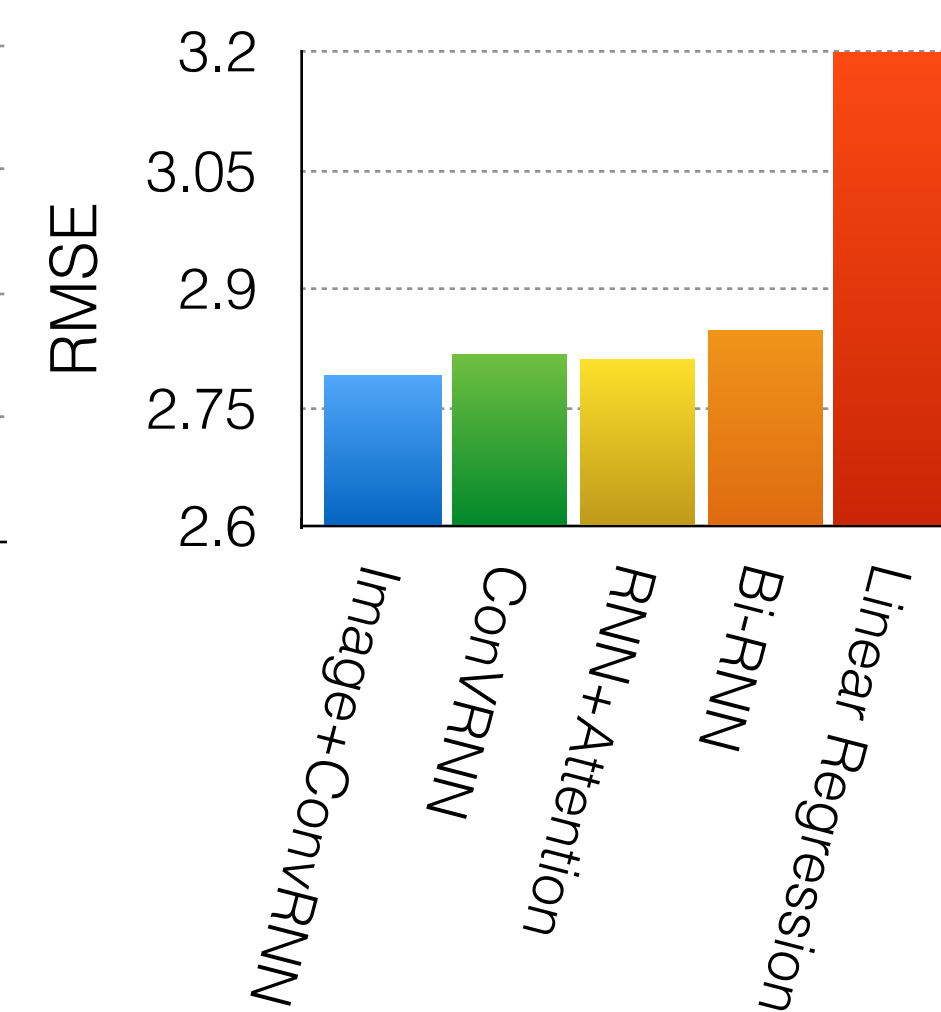**# of ConvRNN Layer V.S. Regression Accuracy (RMS)**

(0 = bidirectional RNN without attention)

Classification Results (All Model)

Regression Results (All Model)

Visualizing Attention Weight

A typical review with the top 25 words greened base on attention intensity.

If you want to splurge on a very tasty dinner, Del Frisco's is definitely one spot you should check out. This restaurant is a change of pace since it is located off the strip. The staff here provides impeccable service as well as an overall pleasant experience. The sesame bread that they give you is freakin addicting that I regret not saving room for my steak. I had their signature drink, the VIP (a pineapple infused vodka), but I stopped after one because it did taste like a bad hangover waiting to happen. I normally don't eat seafood, but I actually thought the crab cake tasted good. This is one of the very few places I've ever dined where I can say I was babied (in a good way)! They cut and served our lobster, served our lobster mac & cheese, served our potatoes au gratin, help split the humongous tomahawk wagyu into two portions, and even helped another Yelper take foodie pics muahaha. I should've visited the restroom to see if they wipe for me too! I had the prime ribeye which was awesomely delicious and cooked perfectly to medium rare as requested.

Size corresponds to the attention weight