DUPLICATE QUESTION DETECTION

Guided by:
Dr. Sudeshna Sarkar

Project Mentor:
Aishik Chakraborty

Presented by:
- Nidhi Mulay (17CS60R75)
- Ami Ladani (17CS60R85)
- Archie Mittal(17CS60R82)
- Jagriti Jalal(17CS60R80)

Problem

- Q&A forums have many duplicate questions that tend to frustrate highly active users
- Detect semantic similarity in sentences having different wording and phrasing
- Cluster and join duplicate questions together
- Remove clutter and improve Quality of Service

Redundancy in questions

Same question, rephrased

- What would happen if you put milk in a coffee maker?
- What would happen if I put milk instead of water into my automatic drip coffee maker?

Related, but not asking the same question

- What got you into real estate investing?
- What is real estate investing

Dataset Description

1. The dataset is downloaded from :

http://qim.ec.quoracdn.net/quora_duplicate_questions.tsv

2. Number of question pairs: 404,278
3. Number of positive examples (duplicates): 161,711
4. Number of negative examples (non-duplicates): 242,567

id	qid1	qid2	question1	question2	is_duplicate
447	895	896	What are natural numbers?	What is a least natural number?	0
1518	3037	3038	Which pizzas are the most popularly ordered pizzas on Domino's menu?	How many calories does a Dominos pizza have?	0
3272	6542	6543	How do you start a bakery?	How can one start a bakery business?	1
3362	6722	6723	Should I learn python or Java first?	If I had to choose between learning Java and Python, what should I choose to learn first?	1

Fig. 1: Snippet of the data file

Baseline Model

- The baseline model uses **word2vec** to generate word vectors.
- Questions are converted to tf-idf embeddings to get 300-dimensional input vectors.
- 3-layered Feedforward network is applied with 128 hidden units in each layer to train the model.
- Objective function : Euclidean distance between two questions.
- Loss Function used : Contrastive loss
- Optimizer used : Adam Optimizer

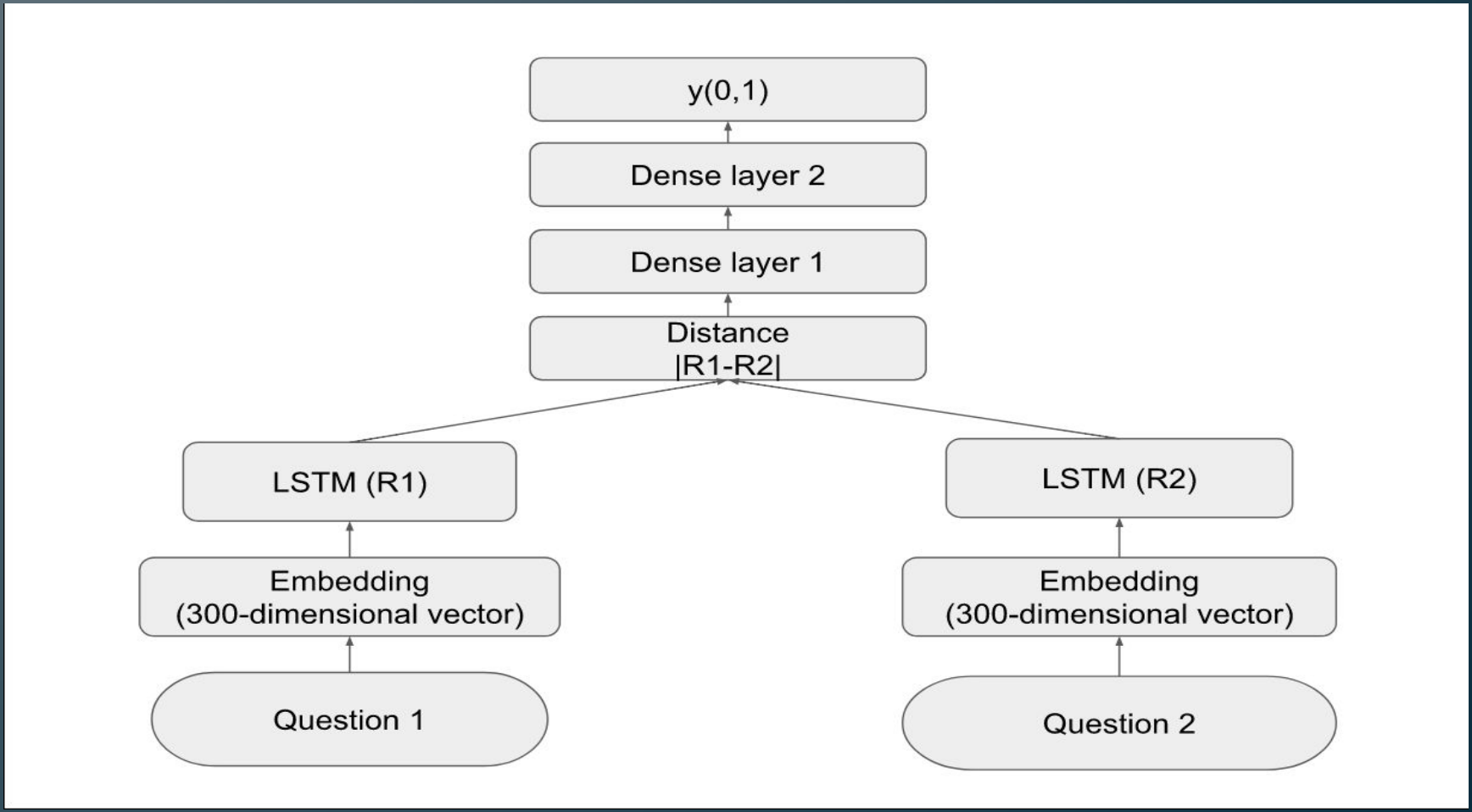
Contrastive Loss

$$L_{contrastive} = y\_true * (y\_pred)^2 + (1 - y\_true) * (\max(1 - y\_pred, 0))^2$$

Baseline Model Output

```
84997/84997 [==============================] - 12s 142us/step - loss: 0.1726 - val_loss: 0.2034
* Accuracy on test set: 60.14868%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 145us/step - loss: 0.1704 - val_loss: 0.2095
* Accuracy on test set: 59.70425%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 146us/step - loss: 0.1697 - val_loss: 0.2103
* Accuracy on test set: 56.27350%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 142us/step - loss: 0.1692 - val_loss: 0.2058
* Accuracy on test set: 58.38038%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 145us/step - loss: 0.1687 - val_loss: 0.2133
* Accuracy on test set: 55.50797%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 144us/step - loss: 0.1692 - val_loss: 0.2129
* Accuracy on test set: 54.91902%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 142us/step - loss: 0.1674 - val_loss: 0.2070
* Accuracy on test set: 61.32183%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 144us/step - loss: 0.1666 - val_loss: 0.2047
* Accuracy on test set: 57.45217%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 142us/step - loss: 0.1658 - val_loss: 0.2117
* Accuracy on test set: 55.65247%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 144us/step - loss: 0.1654 - val_loss: 0.2071
* Accuracy on test set: 57.82097%
Train on 84997 samples, validate on 15000 samples
Epoch 1/1
84997/84997 [==============================] - 12s 142us/step - loss: 0.1643 - val_loss: 0.2042
* Accuracy on test set: 59.16584%
jagriti@jagriti-HP-Pavilion-15-Notebook-PC:~/DL project/exp/QuoraDQBaseline-master$
```

LSTM Model Architecture



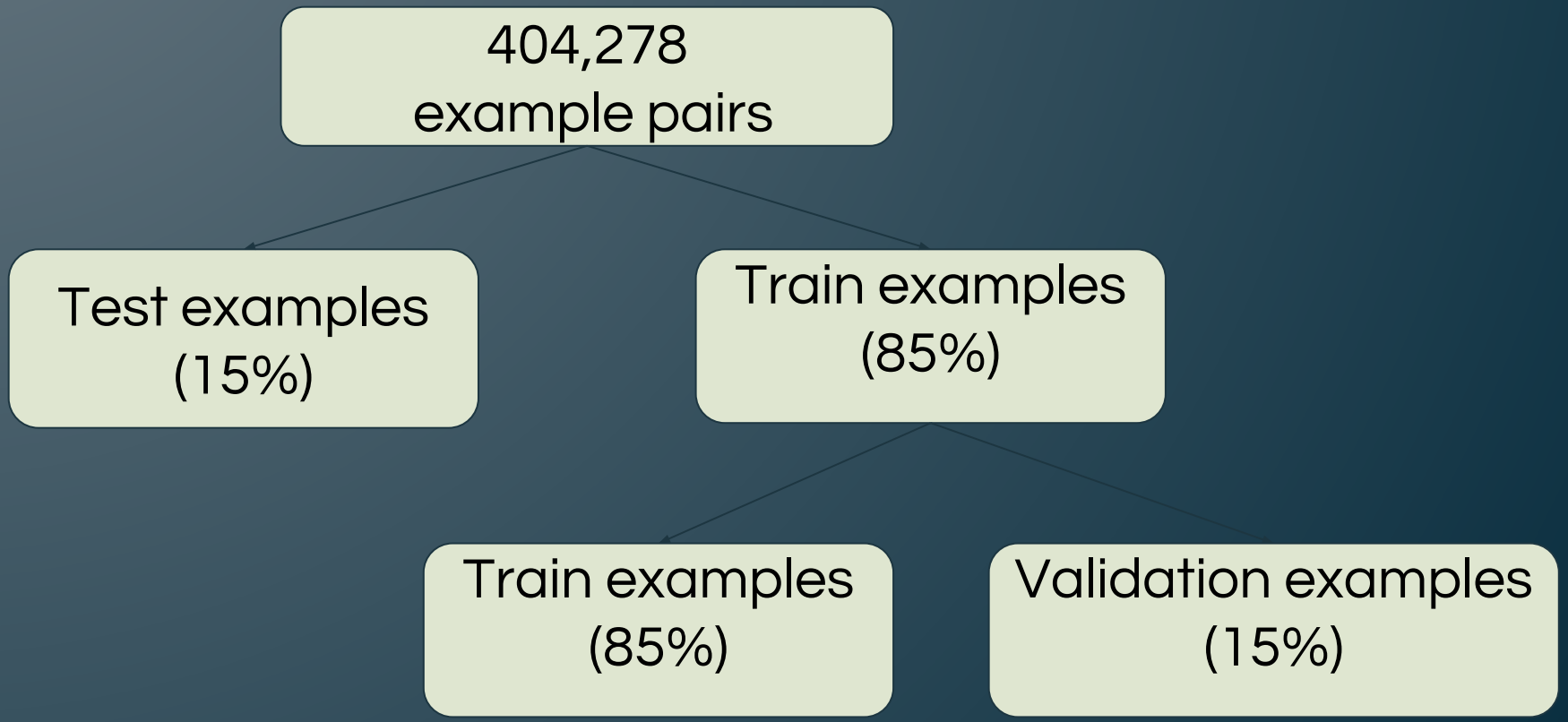
Model Description

- Word embeddings - Glove 300d
- Vocab size - 95602
- Max question length - 25
- LSTM Hidden layer units - 64, 128, 150
- Activation function - Sigmoid
- Dense Layer 1 - 16 units
- Dense Layer 2 - 1 unit
- Loss function - Cross entropy
- Optimizer - Adam Optimizer
- No of Epoch - 10
- Batch size - 256

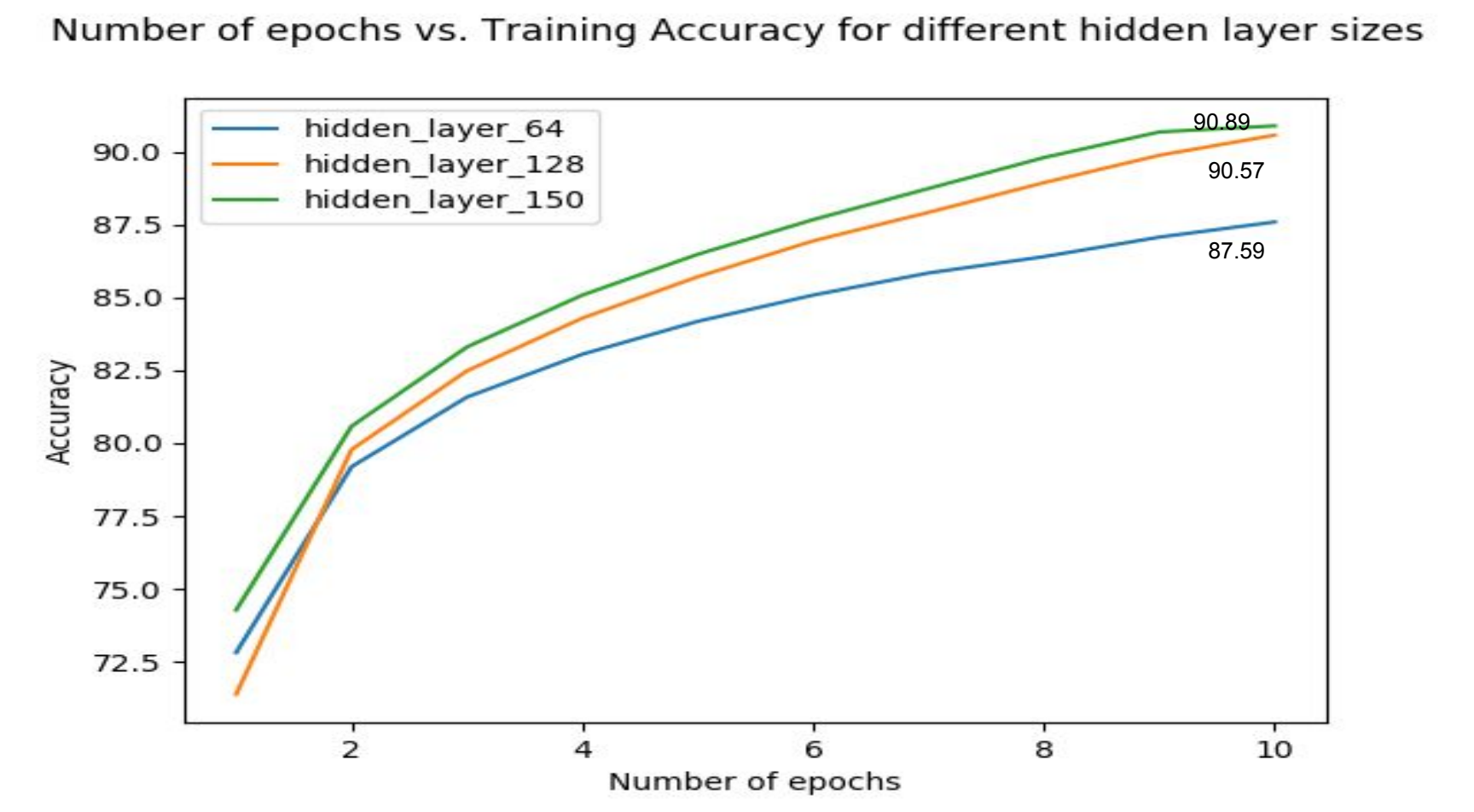
Similarity Results for Test Data

test_id	question1	question2	is_duplicate
343651	How do I learn to type with all 10 fingers?	Is it worth learning to type with the correct finger positions for a typing test in 10 days from now?	0.00111945
343653	Who is a physicist?	Who exactly is a physicist?	0.935248017
343656	How can I increase my intelligence as much as possible?	How can intelligence be increased?	0.740491509
343658	What is the best application to learn C and C++ from the basics?	How can I learn C and C++?	0.49376446

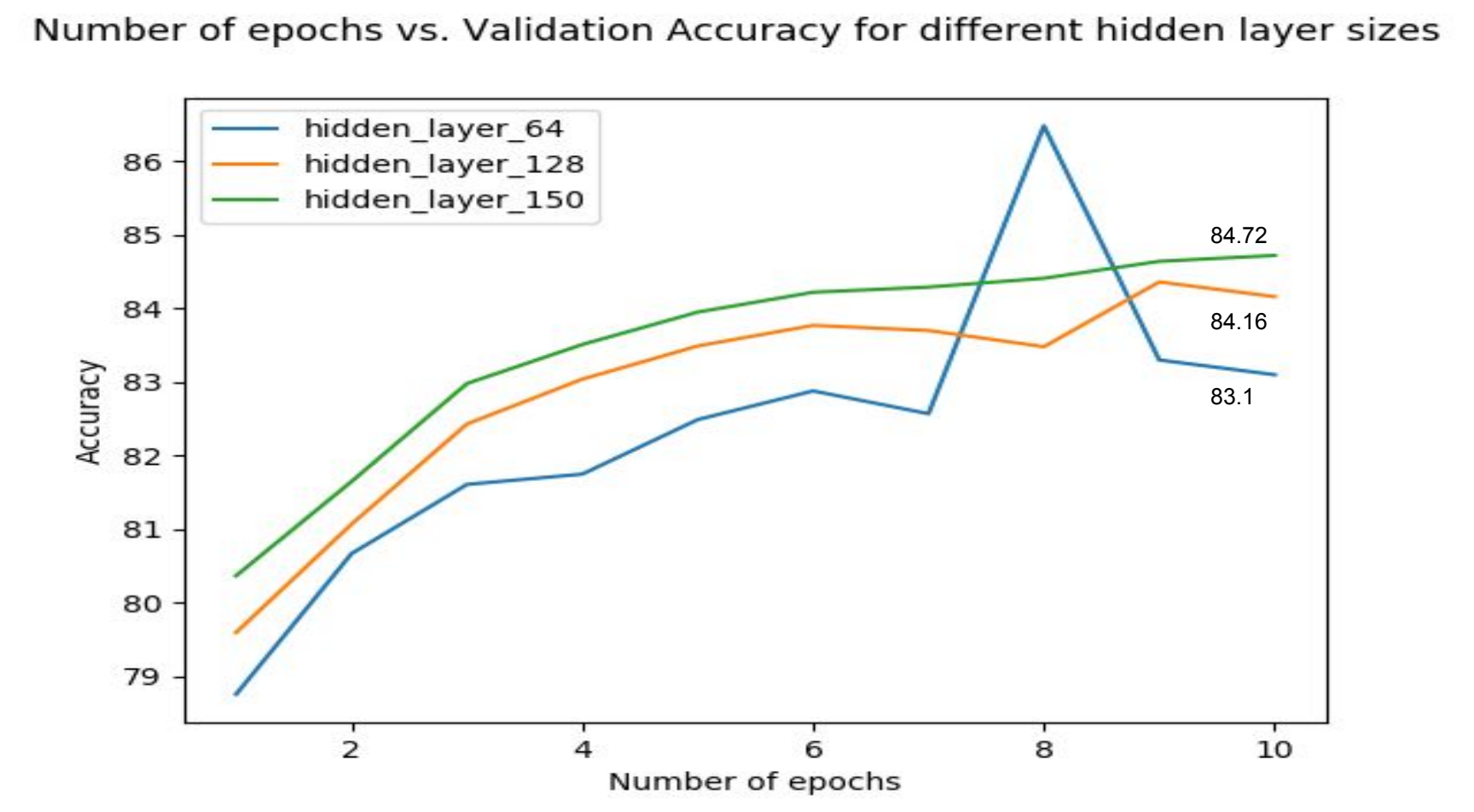
Dataset


404,278
example pairs
Test examples
(15%)
Train examples
(85%)
Train examples
(85%)
Validation examples
(15%)

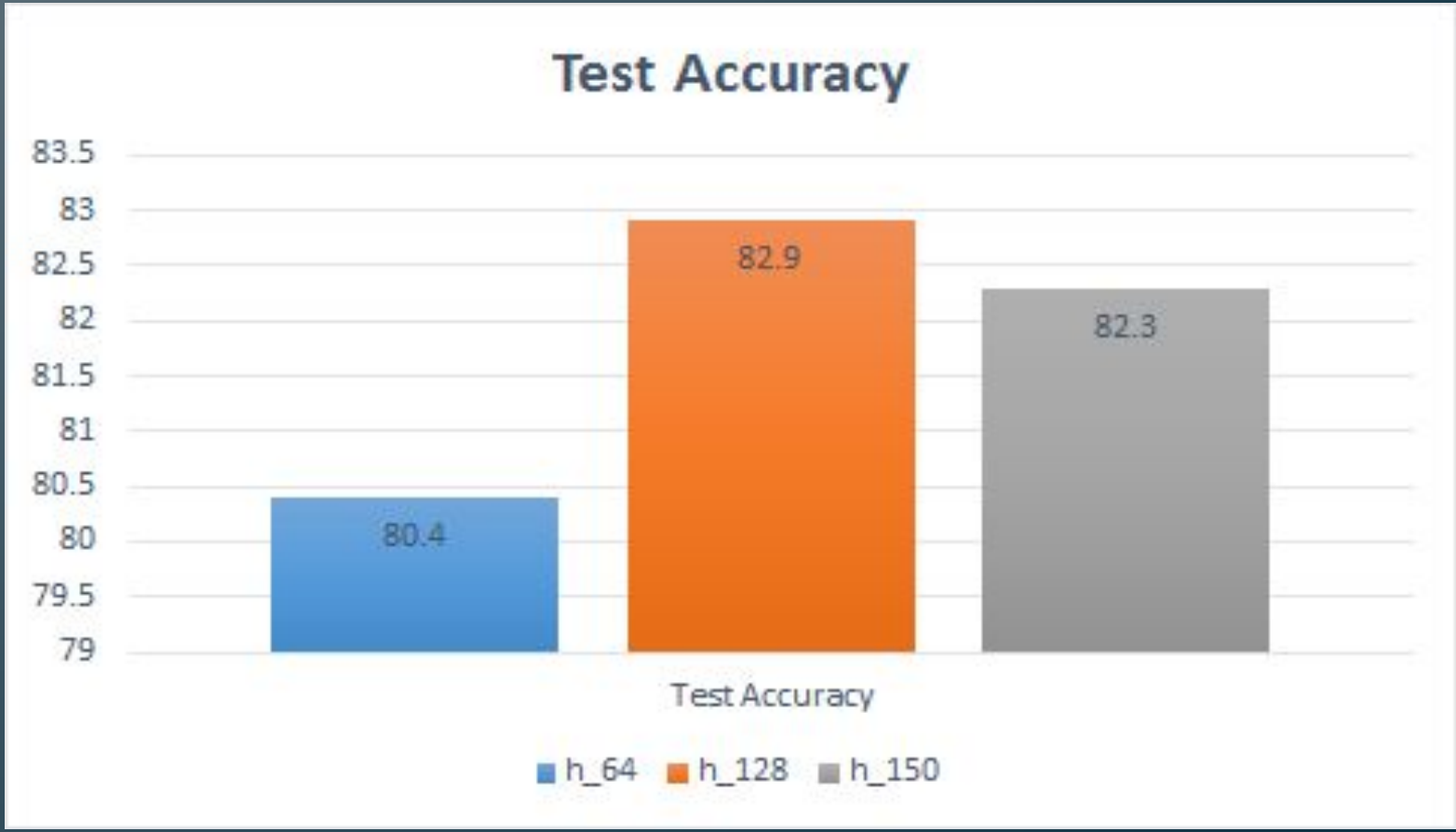
Training Accuracy



Validation Accuracy



Test Accuracy



Thank You!