CSED499I Research Project I

Neural Networks' Understanding of Question-Reply Relationship in Programming Education

Minkyung Kwon and Dongwoo Kim @ POSTECH
mkkwon@postech.ac.kr

Problem Overview

Research Question

To what extent can Neural Networks describe the knowledge acquisition process in programming education, using question (English text) – reply (code) pairs as input-output?

Knowledge Tracing

A technique or model used in education to infer a learner's knowledge state based on their interaction with learning materials (Piech et at, 2015).

Student A

Question 1	Question 2	Question 3	Question 4	Question 5	Question 6
O	X	X	X	O	O

Question 7	Question 8	Question 9
?	?	?

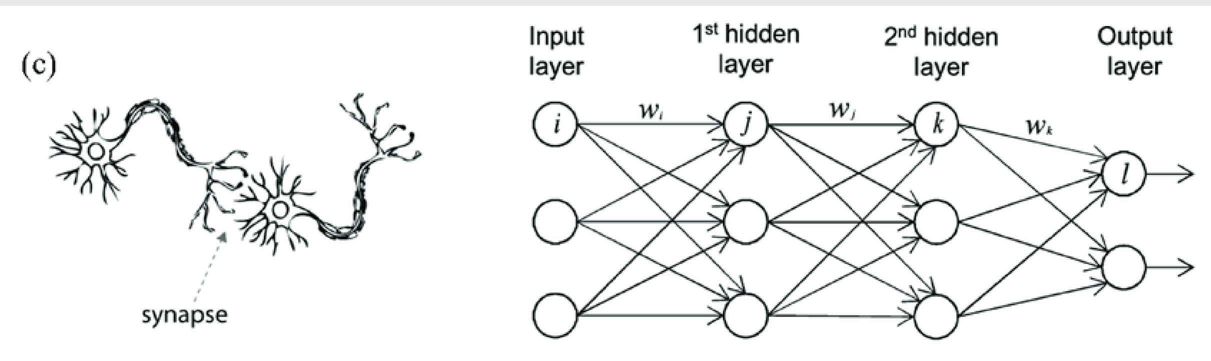
Artificial Neural Network vs. Biological Neurons

- **Similarities**

1) The incoming-outgoing connections in ANN – Neuron connections in the brain
2) The linear calculation & activation function – Neuron activation to nearby neurons

- **Difference**

1) The chemical-electrical mechanisms of neurons far surpass the mathematical equations composing artificial neurons.
2) Neurons are 'emergent' where the collective behaviors or functions arose from the interaction among neurons, rather than from individuals.



Dataset: CSEDM

CSEDM 2019

- Java code collected from a CS1 course in the Spring 2019 semesters at a public university in the U.S
- 46,835 sets of code submission from 264 students to 50 types of programming question
- e.g.

Question #1	Answer #1
Write a function in Java that implements the following logic: Given 2 ints, a and b, return their sum. However, sums in the range 10..19 inclusive, are forbidden, so in that case just return 20.	`public int sortaSum(int a, int b) { if (a + b >= 10) { if (a + b <= 19) { return 20; } return a + b; } else { return a + b; } }`

Preprocessing for Question

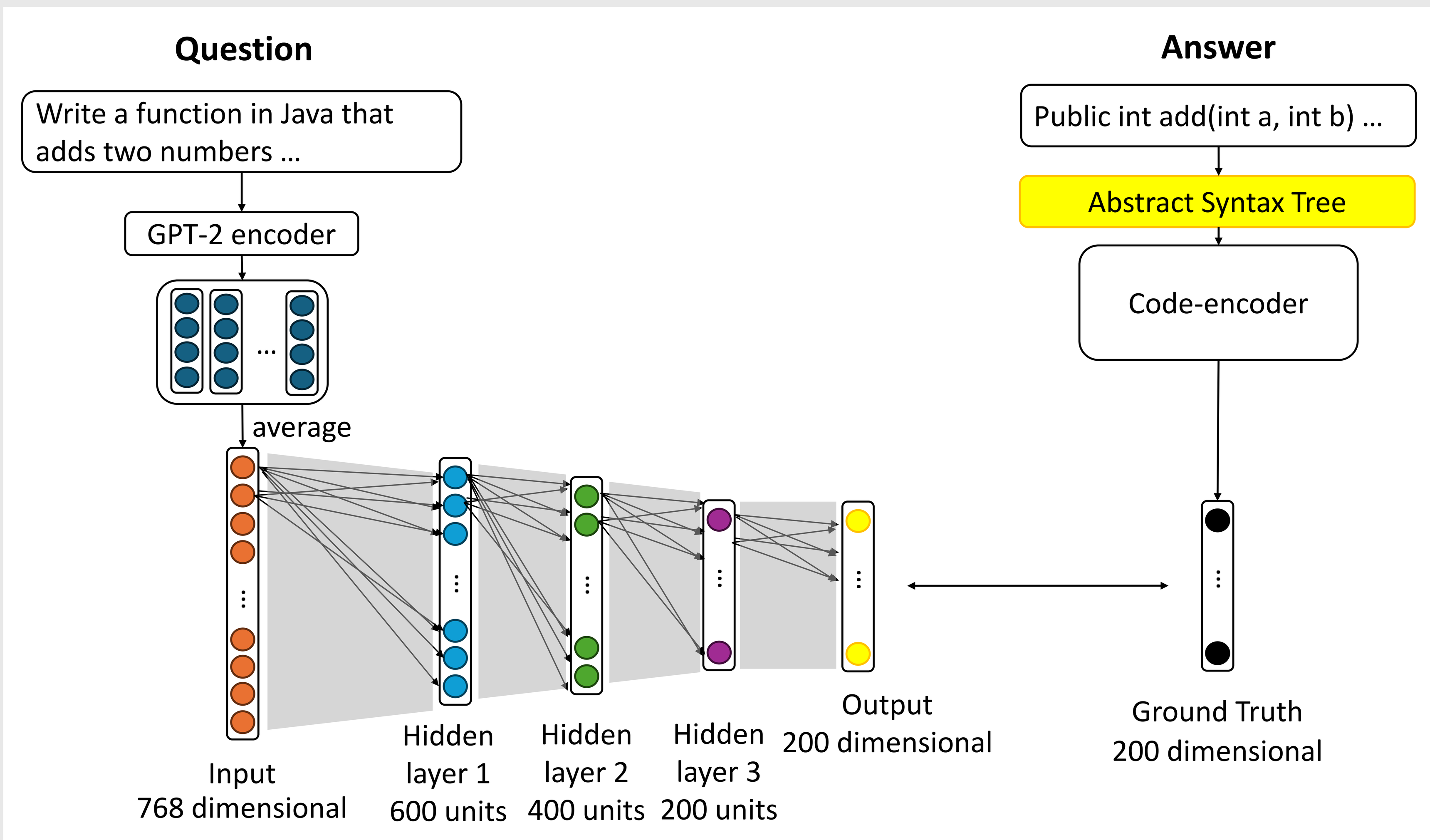
- GPT 2: transforms a natural text into a 768-dimensional vector

Preprocessing for Answer

- ASTNN: transforms a Java code into a 200-dimensional vector

Approach 1: MLP

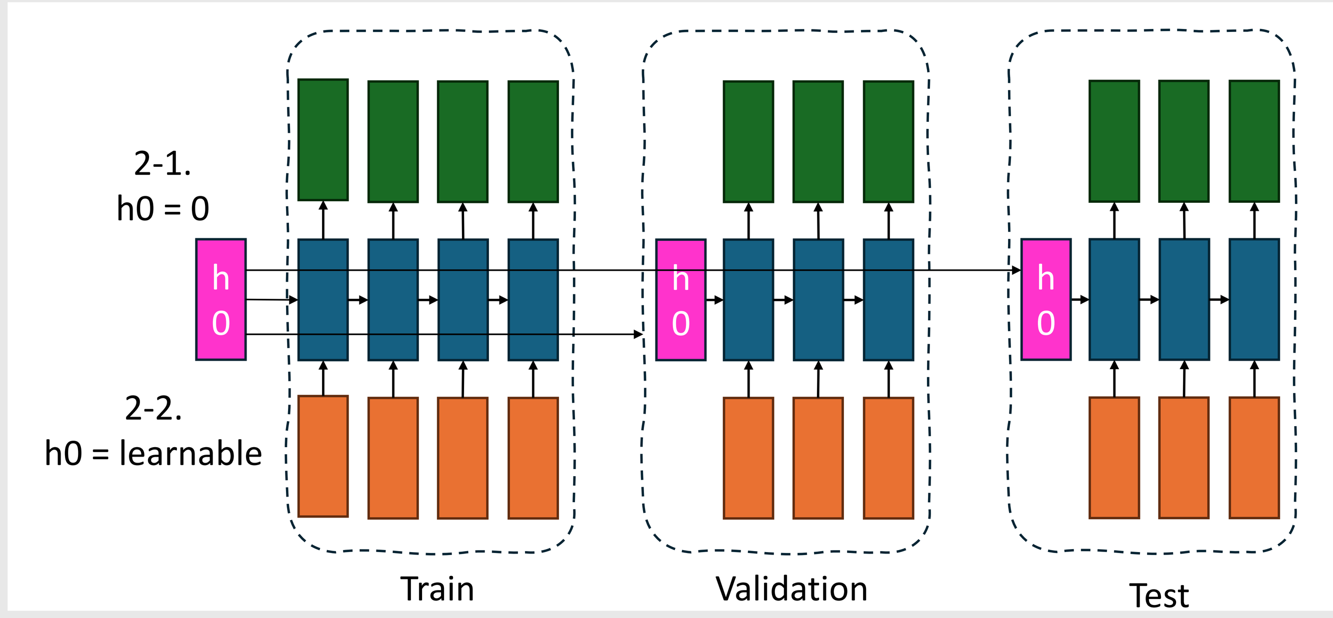
Experiment 1: Basic Multi Layer Perceptron



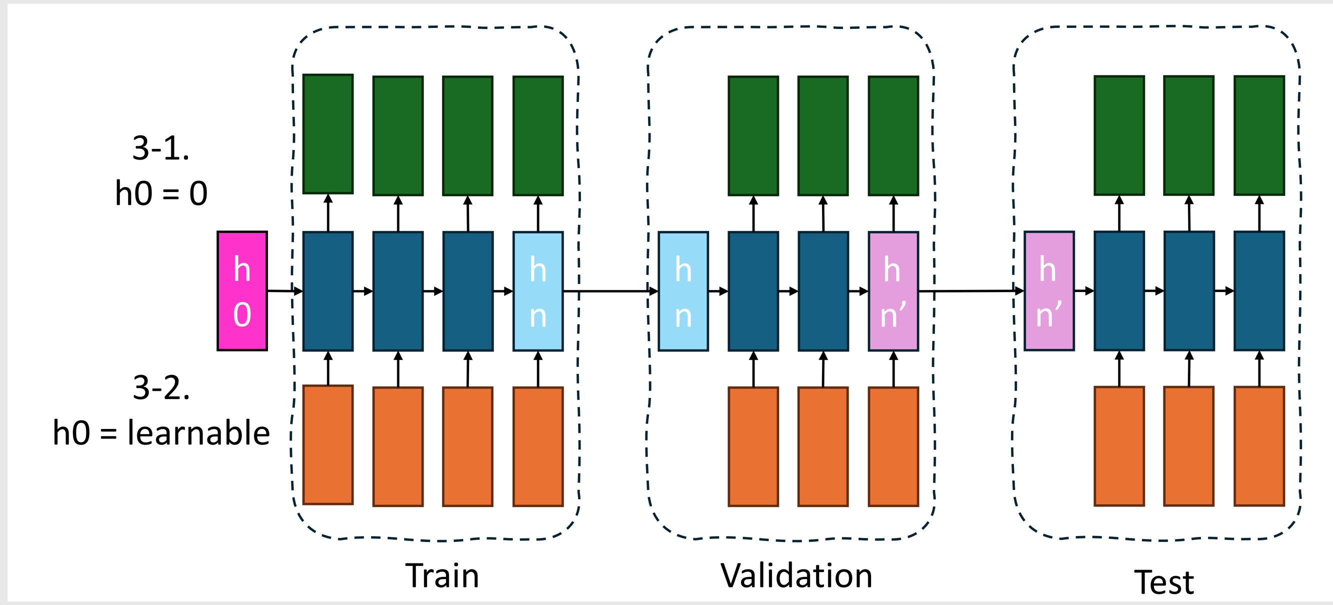
- Input: embedding of question sentence, Output: embedding of student's answer code
- MLP with 3 hidden layers, each 600, 400 and 200 units, MSE loss
- Train: Valid: Test = 64%:16%:20% per student

Approach 2: RNN

Experiment 2-1: h0 set to 0
Experiment 2-2: h0 being a learnable parameter



Experiment 3-1: h0 set to 0 + Passing final hidden state
Experiment 3-2: h0 being a learnable parameter + Passing final hidden state



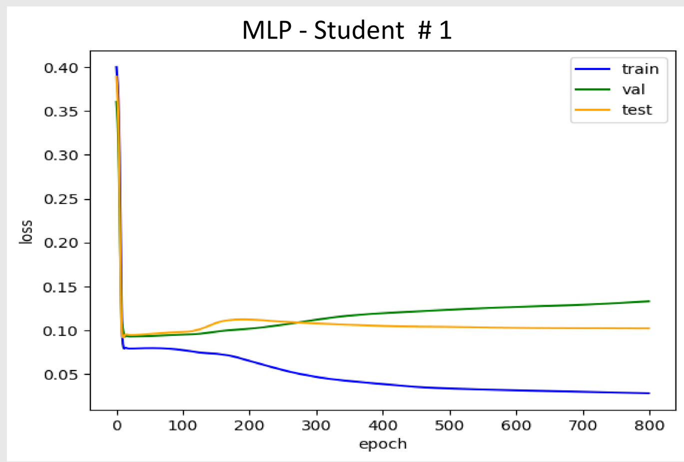
- Input: embedding of question sentence, Output: embedding of student's answer code
- Long Short-Term Memory (LSTM), MSE loss
- Train: Valid: Test = 64%:16%:20% per student

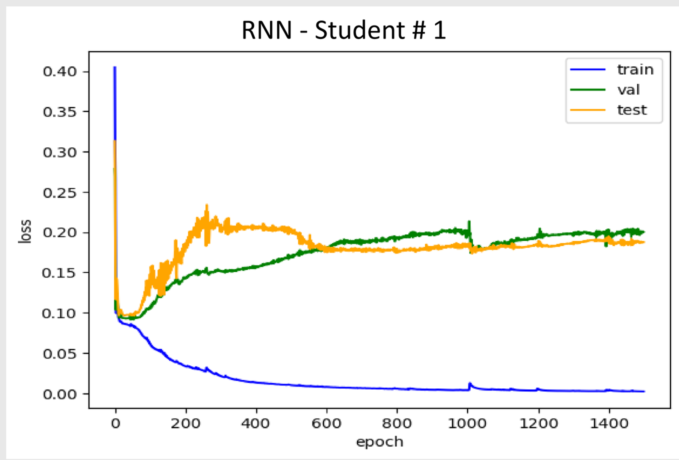
Experiment & Results

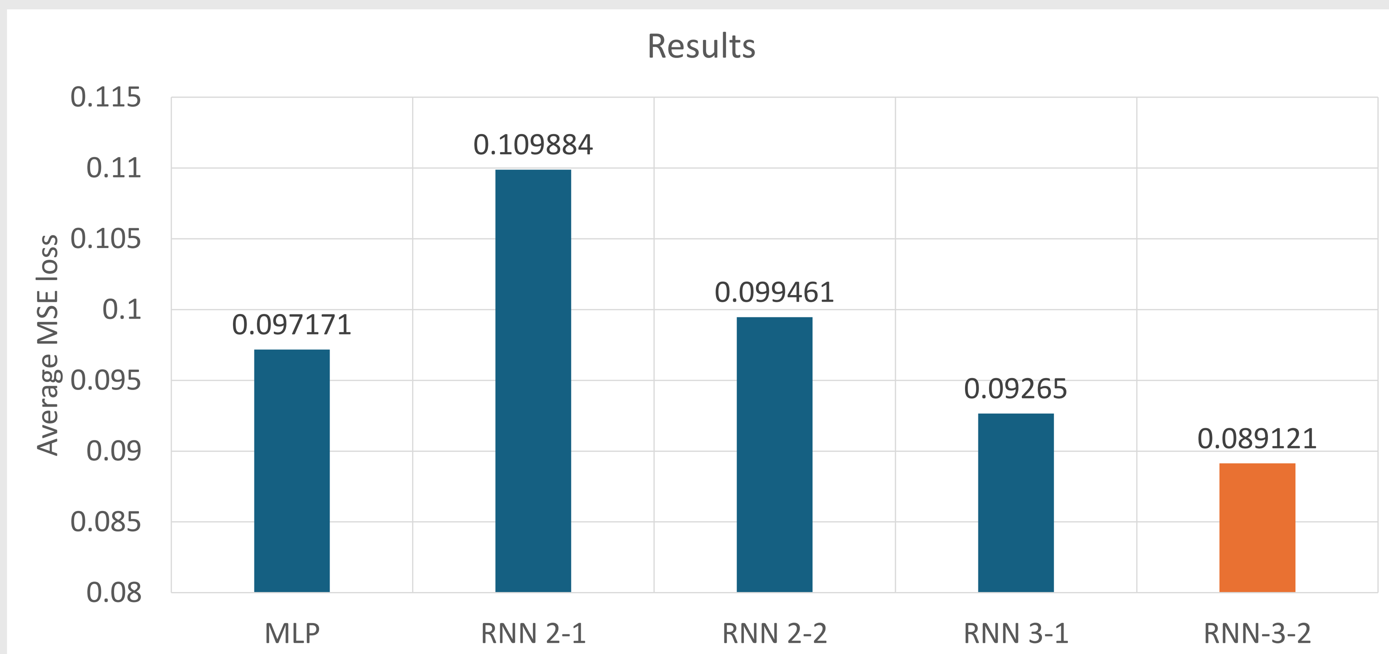
Procedure

- Model 264 students' knowledge state separately in five different experiments and the test loss was averaged out
- Early stopping at the lowest validation loss
- Code & data available at: https://github.com/mkxdxdxd/CSED499I.git

Results







Discussion & Findings

- **RNN** describes the learner's **knowledge acquisition process** better than MLP
- Preserving RNN's **final hidden state** is important in RNN's performance.
- One's **prior knowledge state** just before solving test set questions takes a crucial role in modeling.

Conclusion

- When modeling **a learner's knowledge state**, RNN outperforms MLP. This is because RNN reflects context through its **hidden states**.
- Implementing RNNs in educational tools can provide **more accurate assessments** of a learner's **knowledge state**.

1. Piech, Chris, et al. "Deep Knowledge Tracing." Advances in Neural Information Processing Systems, vol. 28, 2015, pp. 505-513. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf.
2. Zhang, Jian, et al. "A Novel Neural Source Code Representation Based on Abstract Syntax Tree." 2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE), 2019, pp. 783-794, https://doi.org/10.1109/ICSE.2019.00086.