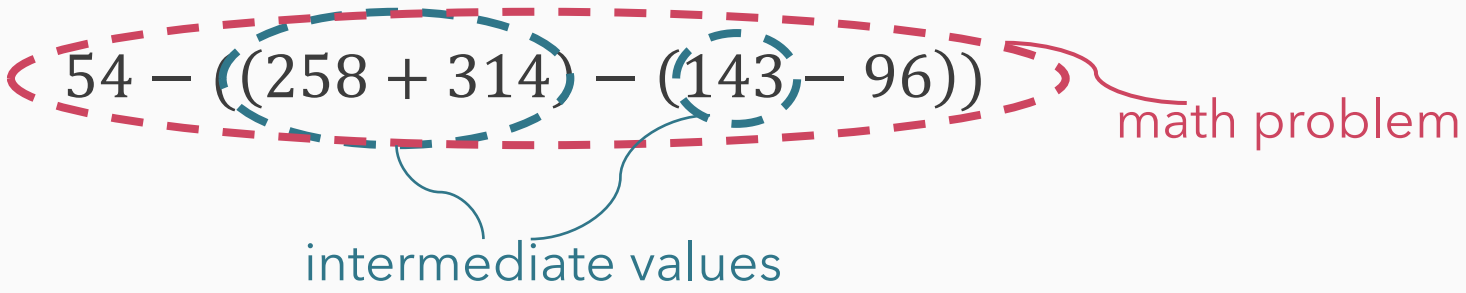
Tracing and Manipulating Intermediate Values in Neural Math Problem Solvers

Yuta Matsumoto[1], Benjamin Heinzerling[2], Masashi Yoshikawa[1], Kentaro Inui[1,2]

[1]·Tohoku Universiy [2]·RIKEN

Summary

- How do neural networks process complex inputs?
- We focus on the intermediate values of simple math problems



- We show that…
  - intermediate values are stored in particular directions of model representations
  - model representations can be manipulated as if inputs change

Neural math problem solvers perform well

- Solving math problems requires storing intermediate values

$$54 - \big((258 + 314) - (143 - 96)\big)$$
$$= 54 - (572 - 47)$$
$$= 54 - 525$$
$$= -471$$

  - We can guess that models capture those intermediate values

- Goal: Find out how intermediate values are stored

Contributions: Tracing and Manipulation

- Idea: focus on simple math problems and their intermediate values

0. Pre-training Transformer on simple math problems
   - e.g., (115-28)-(32-56), 25-(79+104),…

1. Tracing intermediate values by PCA
   - find components where information on intermediate values is encoded

2. Manipulating intermediate values
   - test these components influence model predictions

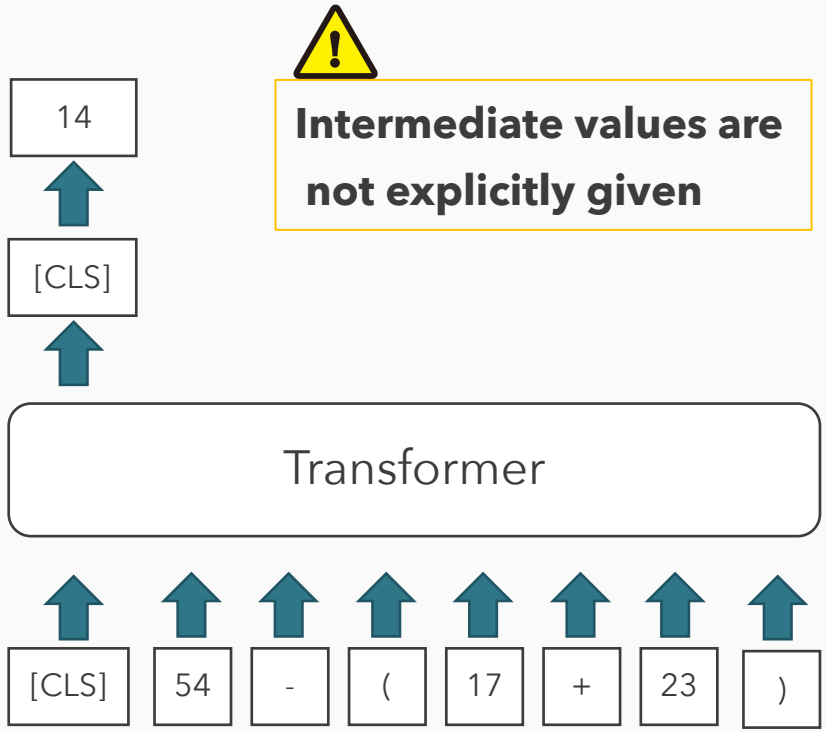
0. Pre-training on simple math problems

Method

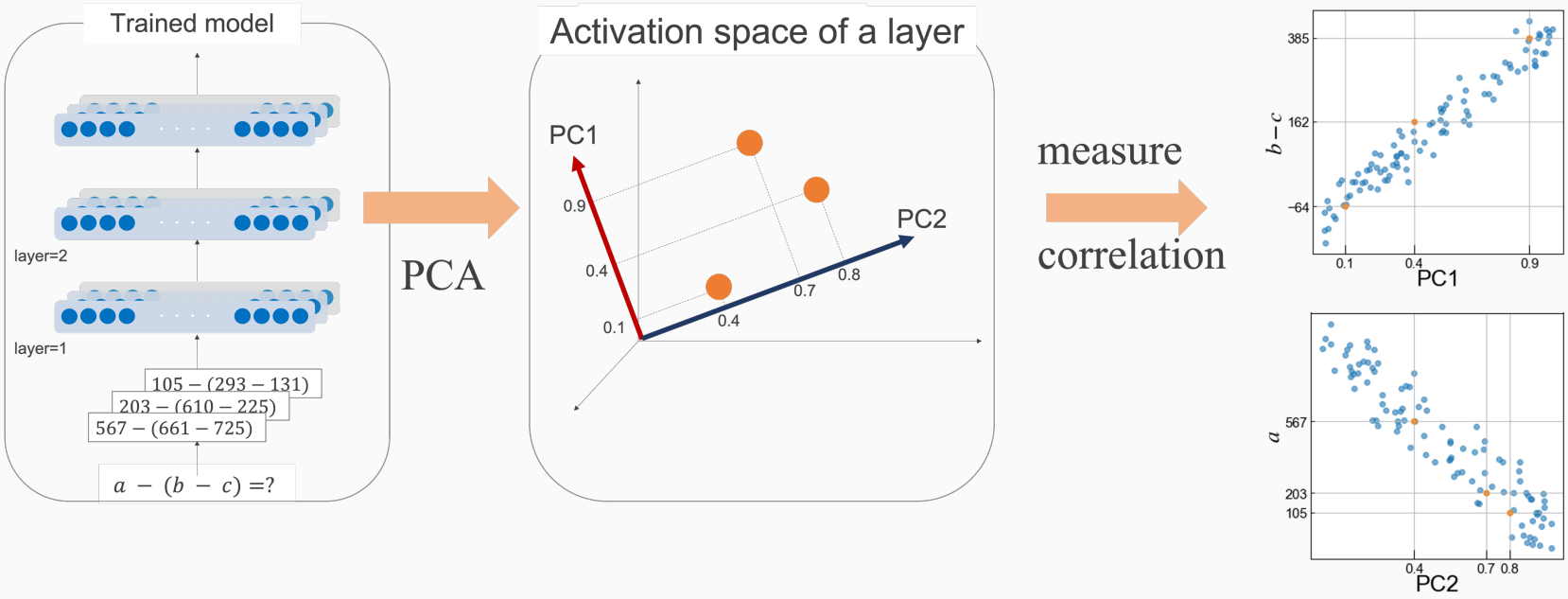
- Prepare multiple patterns of equations
- Train a Transformer model on the equations

Result

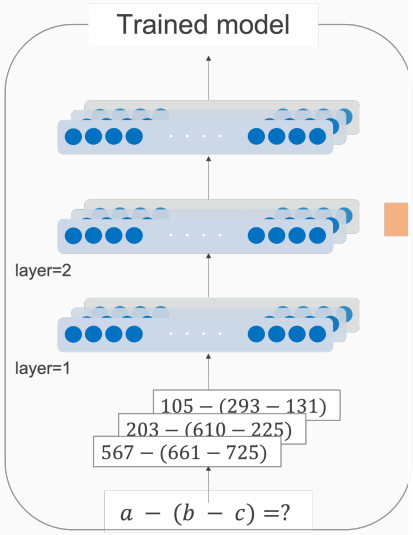
- High performance
  - R^2 score = 0.999



1. Tracing


Trained model
layer=2
layer=1
105 − (293 − 131)
203 − (610 − 225)
567 − (661 − 725)
a − (b − c) =?
PCA
Activation space of a layer
PC1
PC2
0.9
0.4
0.1
0.4
0.7
0.8
measure
correlation
385
162
−64
b − c
0.1
0.4
0.9
PC1
567
203
105
a
0.4
0.7 0.8
PC2

1. Tracing



PCA

Activation space of a layer

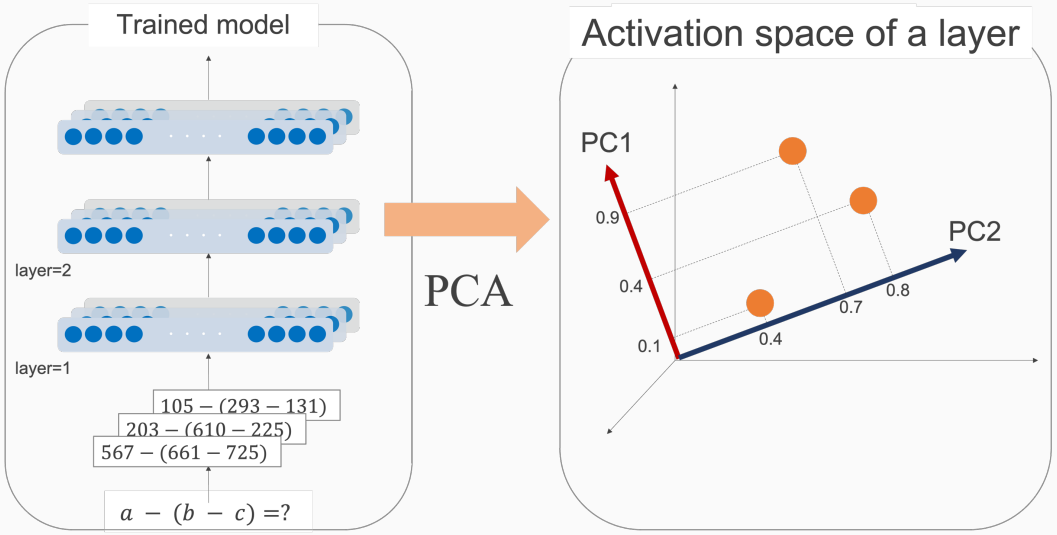
PC1

PC2

measure correlation

1. encode math problem

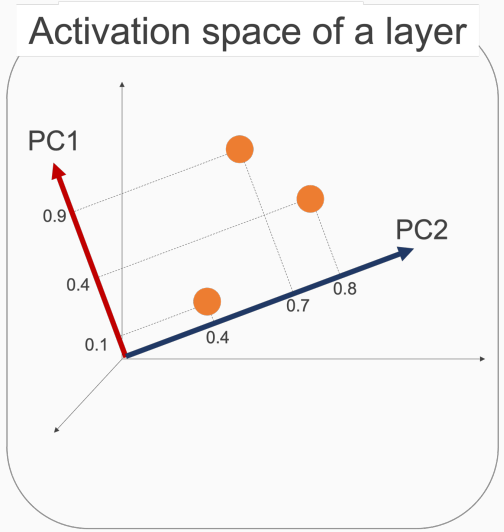
1. Tracing



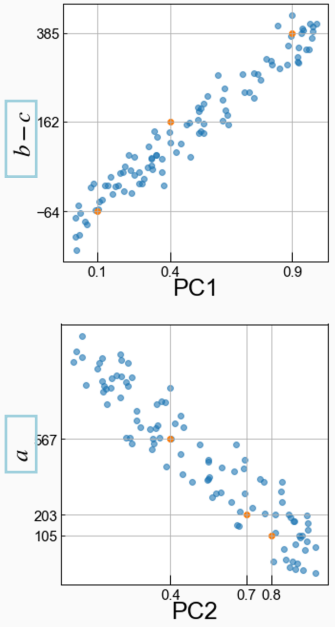
1. encode math problem
2. PCA on hidden representation

1. Tracing

1. encode math problem
2. PCA on hidden representation
3. measure correlation between principal components and intermediate values



measure correlation



$a-(b-c)$

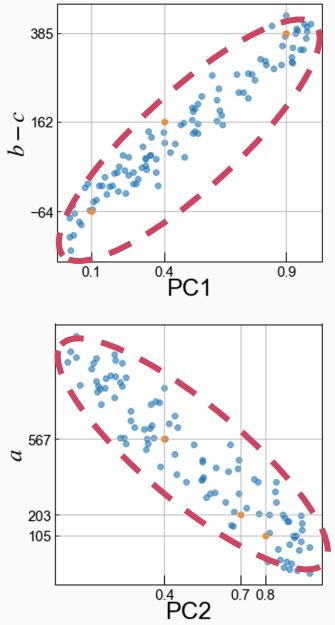
math problem

intermediate values

1. Tracing

If correlation is high…

A principal component encodes the intermediate values

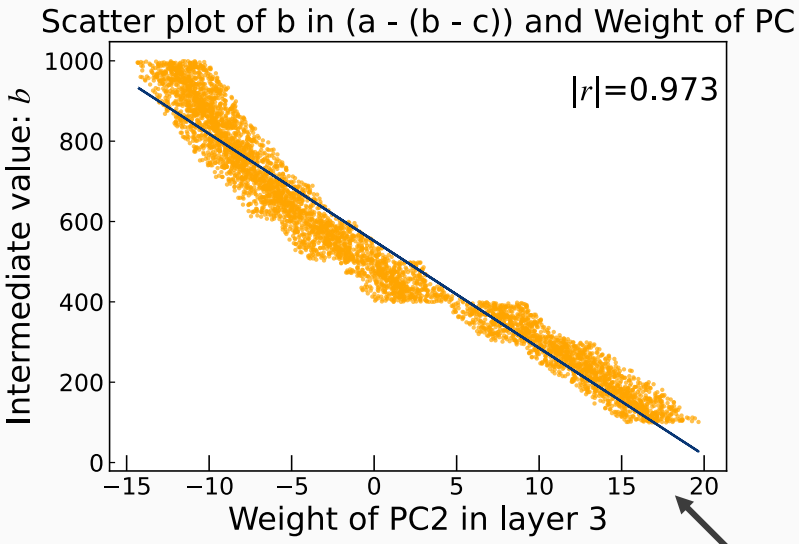


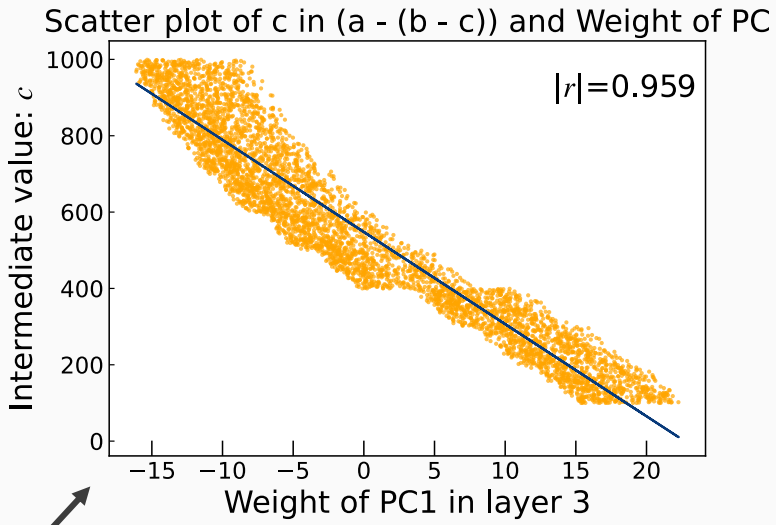
Correlation with intermediate values

Q. Are there principal components that are highly correlated with intermediate values?

A. Yes

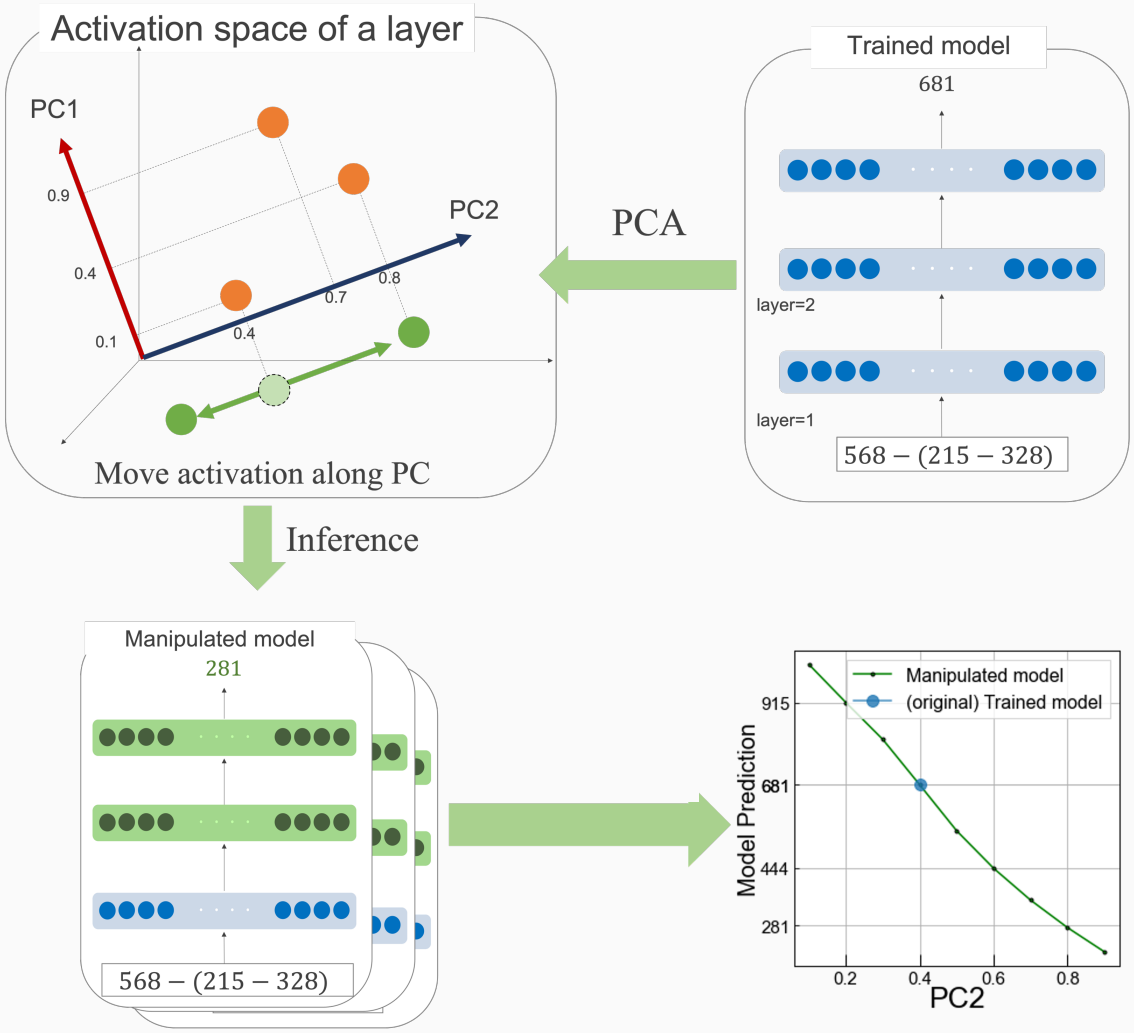
- Some components highly correlate with intermediate values.



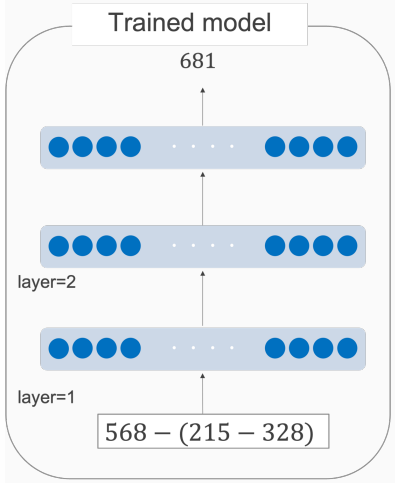


$$a - (b - c)$$

2. Manipulation

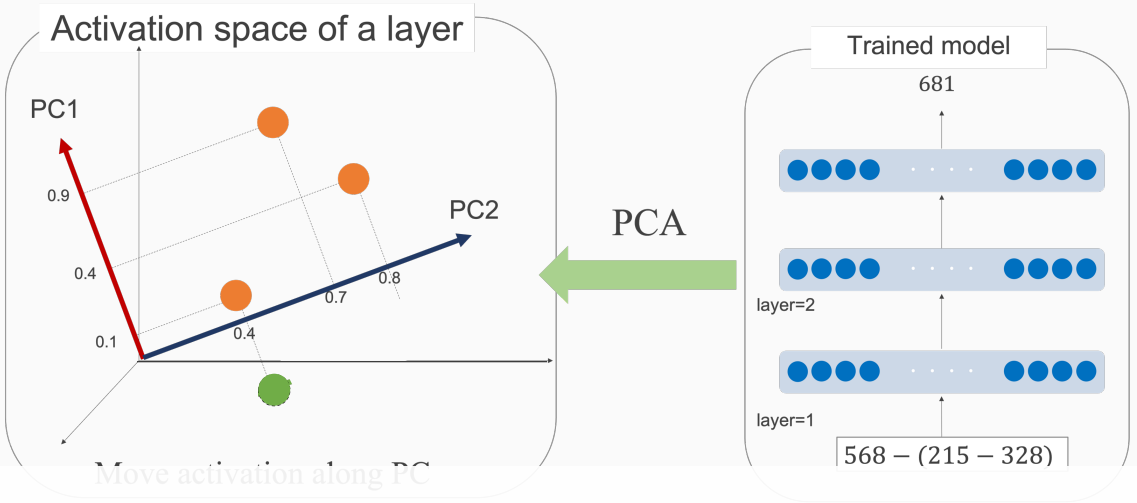


2. Manipulation



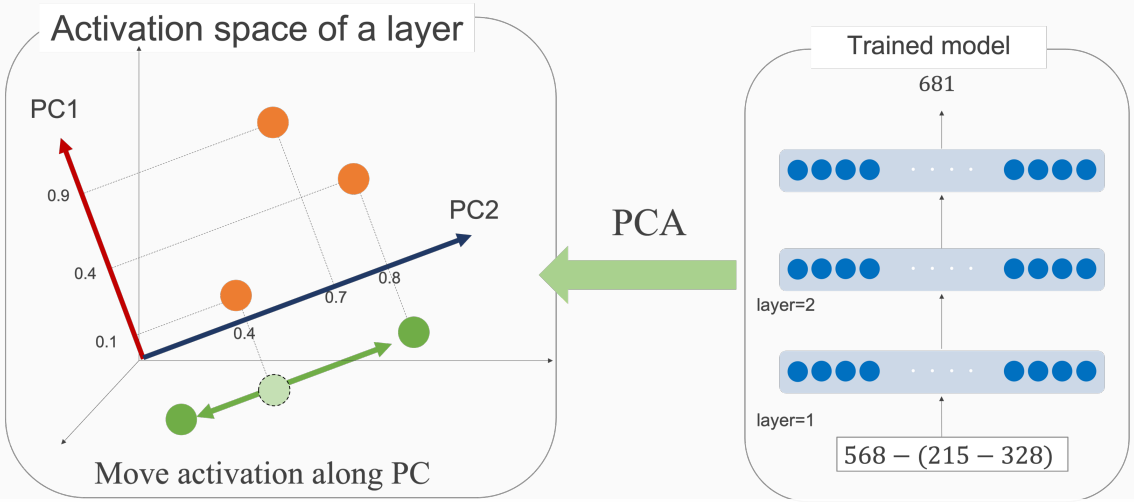
1. encode math problem

2. Manipulation



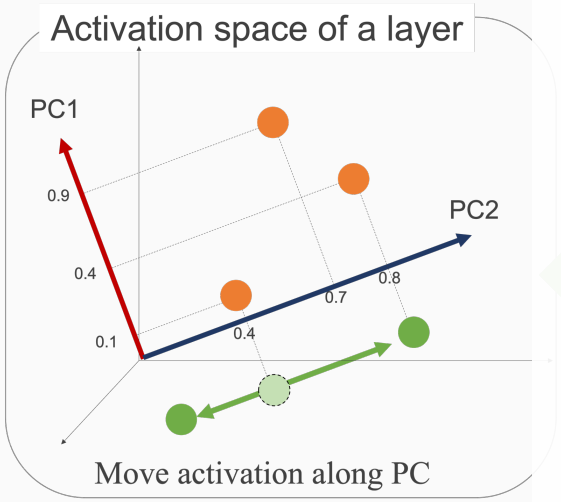
1. encode math problem
2. PCA on hidden representation (same as Tracing)

2. Manipulation



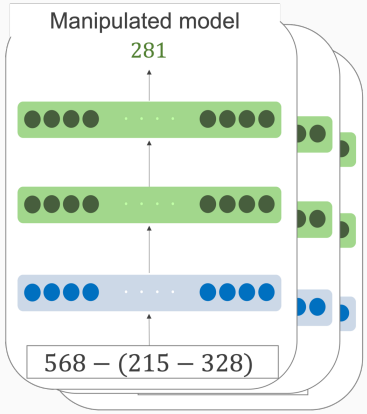
1. encode math problem
2. PCA for hidden representation
3. edit activations

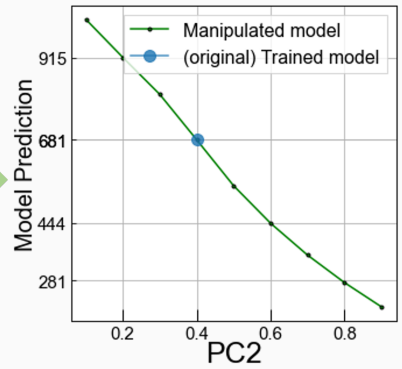
2. Manipulation



1. encode math problem
2. PCA for hidden representation
3. edit activations
4. observe the model prediction change

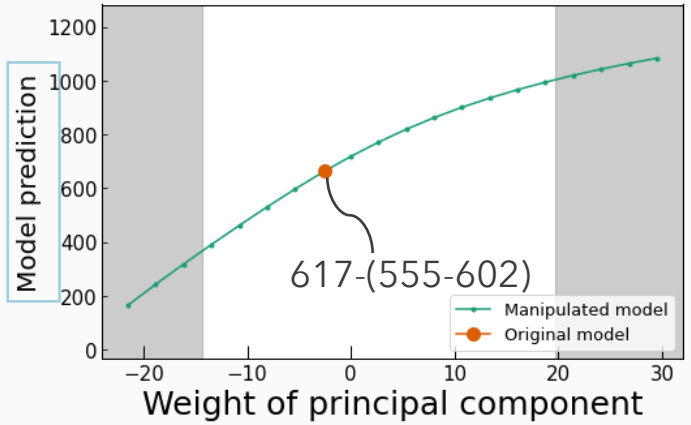
Inference



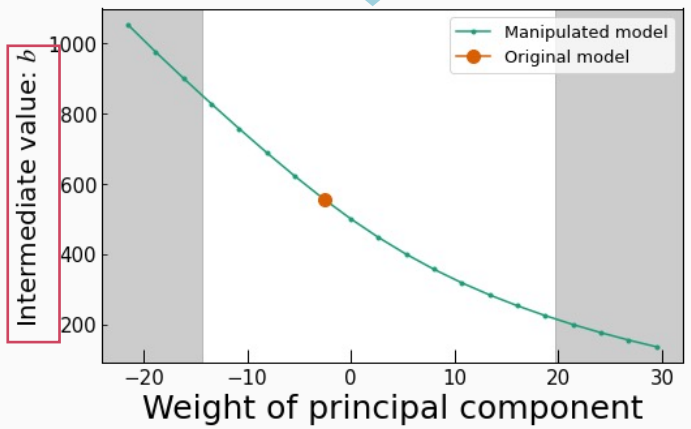


Manipulation results

- Move weights of PC2 in layer 3
  - highly correlated (r=0.97) with b in a-(b-c)
- Model predictions change consistently
  - the intermediate values may be manipulated

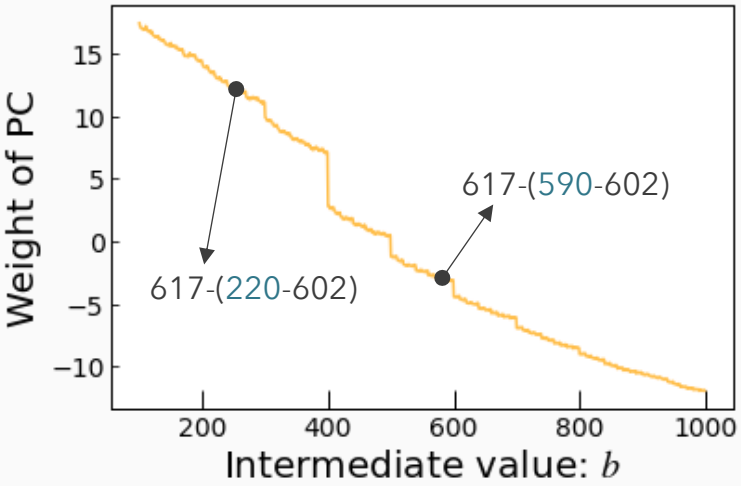


🤔 "We edit the representation of b"

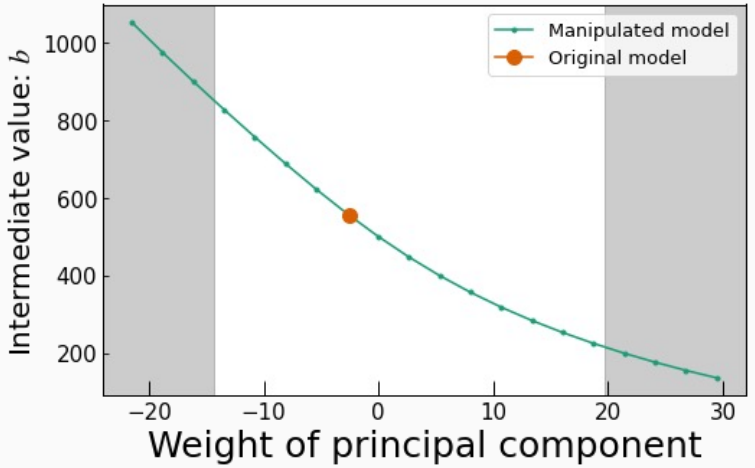


Are we moving intermediate values as if inputs change?

- Prepare inputs with different intermediate values
  - e.g., $617 - (100 - 602), 617 - (101 - 602), \ldots, 617 - (999 - 602)$
- Compare principal components after changing inputs and after manipulating activations



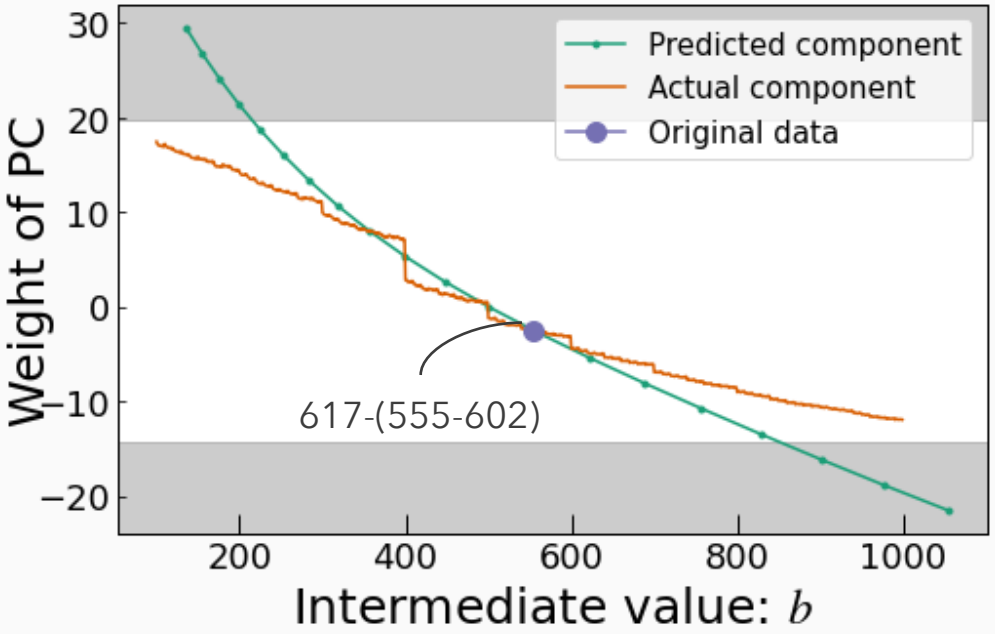
Change input



Manipulate activations

Are we moving intermediate values as if inputs change? →Yes

- High agreement
  - particularly around the original data
- We can change intermediate values by manipulating activations



Predicted: change activation

Actual: change the input

Summary

- We analyzed the representation of intermediate values in neural math problem solvers
- By Tracing, we found the directions which correlate with intermediate values
- By Manipulation, we showed that the directions are causally related with intermediate values