**Data preprocessing**

Extract all information from jason files. (image ID, question, answers, annotations )

1. Image data:
   1. Resize image to (299, 299, 3)
   2. Loading inceptionV3 model with ‘imagNet’ weights and extracting the image features.
2. Text data (question and answer):
3. Transform the question sequence and answer using label-encoding
4. Add start, end, and padding token to question sequence
5. Given multiple answer choices by 10 human experts based on the question, extract only binary answers with yes/no and use the most frequent answers among all answers given by human experts as the label. Also, transform it using label-encoding
6. Feed these data into embedding layers to learn text features.

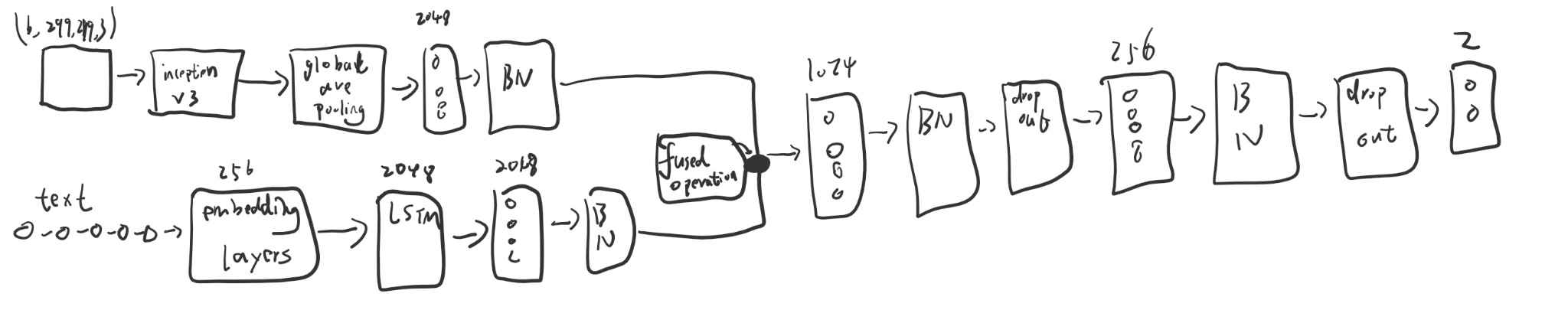
**Hyper-parameters**

* **Batch size: 128**
* **Epochs: 50**
* **Samples: 5000**
* Buffer size: 5000
* **Train Test Split Rate: 8:2**
* Top-k-words for vocab: 1000
* Image resize: 299,299,3
* **Learning Rate:**
  + ExponentialDecay(initial\_learning\_rate=1e-3, decay\_steps=50,decay\_rate=0.9)
* **Adam Optimizer:**
  + Beta1: 0.99
  + Beta2: 0.999
* Embedding Layer Units: 256
* LSTM Units: 2048
* **Dropout Rate: 0.3**
* Dense Units: 2048, 1024, 256, …

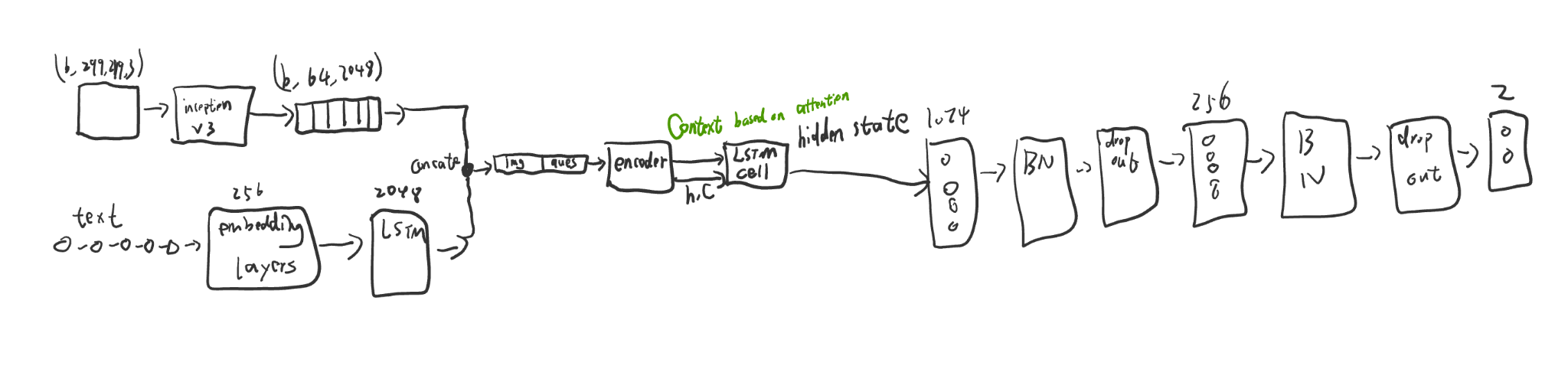
…

**Model architecture (Onelayer, Shortcut, Deep)**

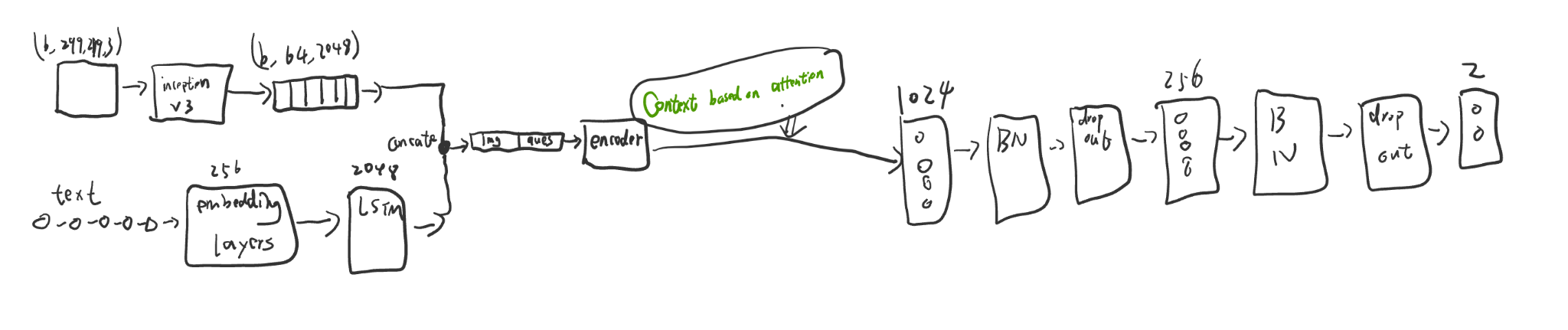
* **Middle Fusion: One Layer Architecture with Concatenation or Addition or Dot-Product**



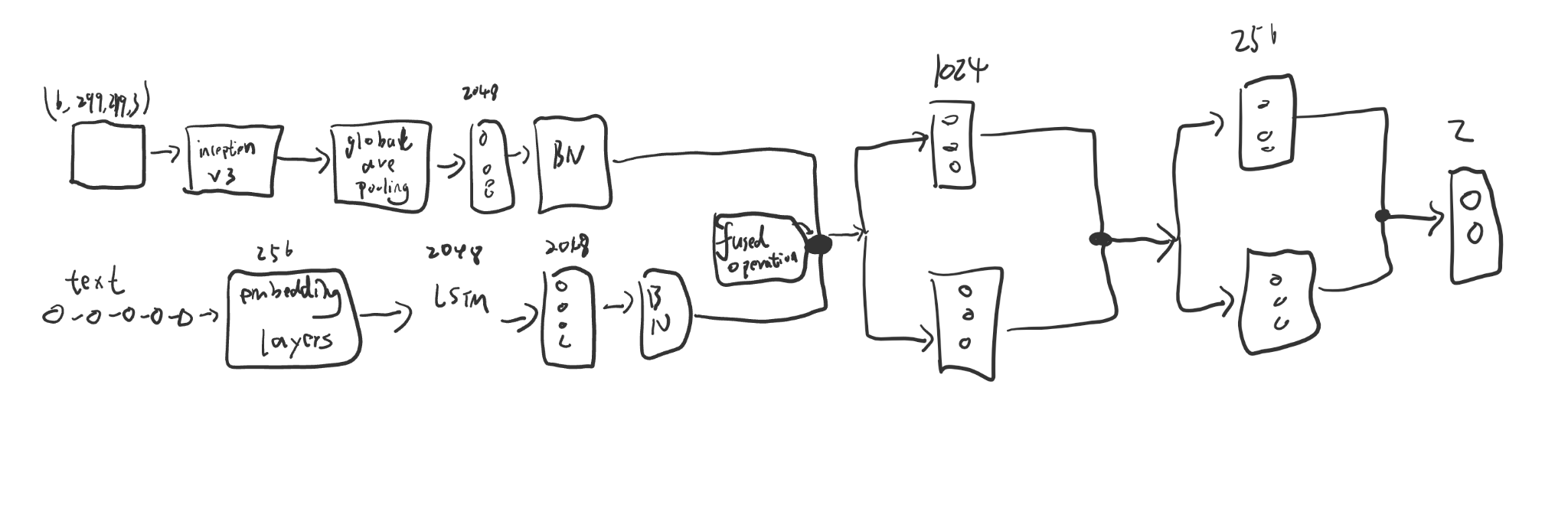
* **Middle Fusion: One Layer Architecture with attention including decoder(LSTM Cell)**

****

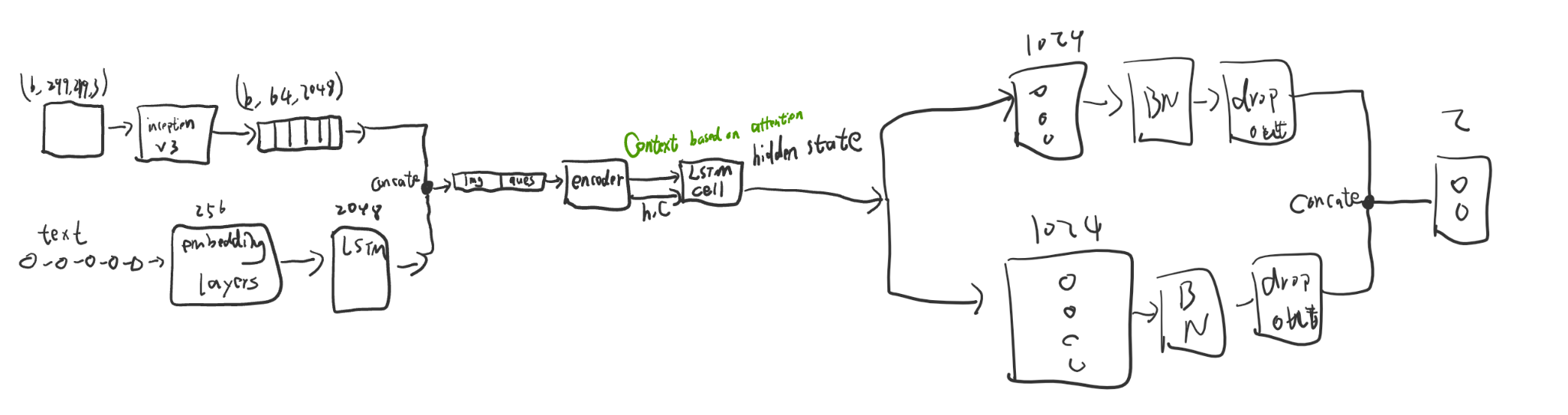
* **Middle Fusion: One Layer Architecture with attention excluding decoder(LSTM Cell)**

****

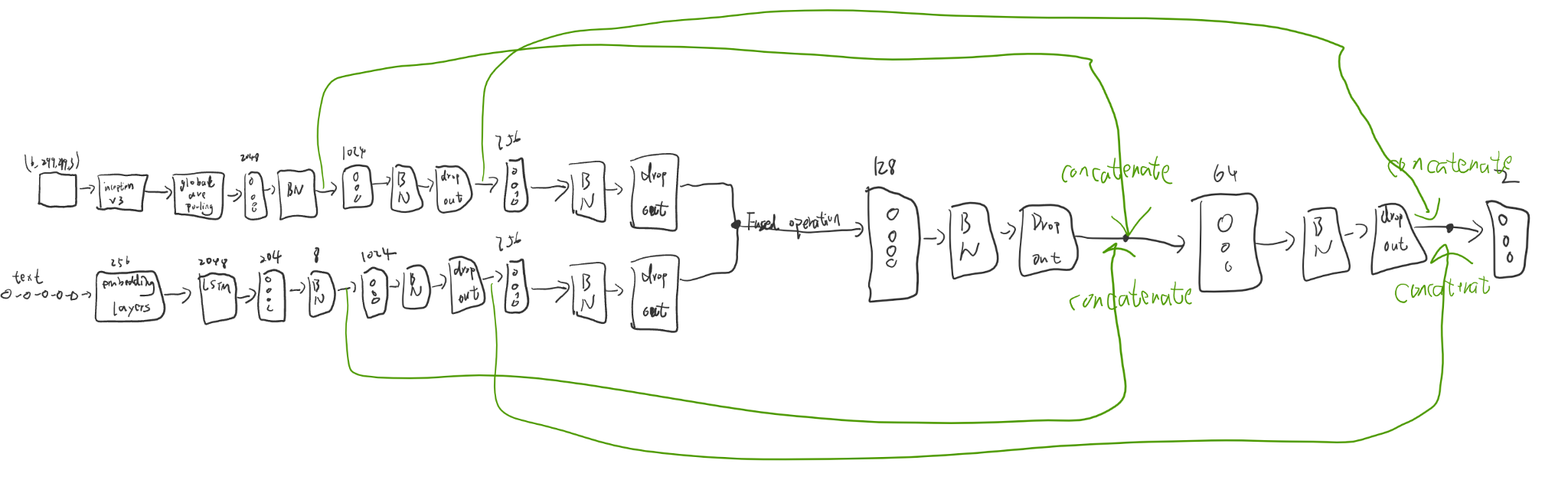
* **Middle Fusion: Deep Fusion with Concatenation or Addition or Dot-Product**

****

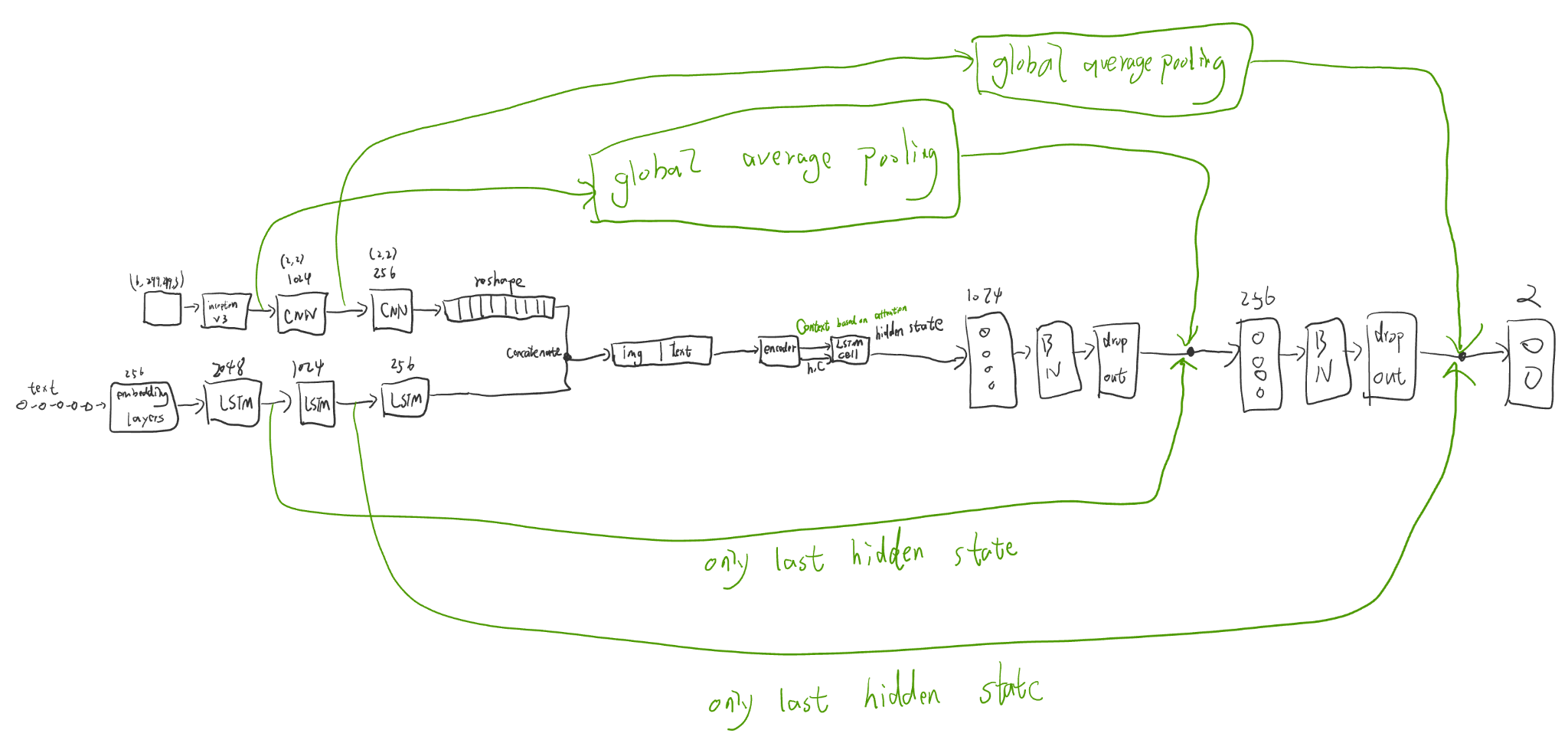
* **Middle Fusion: Deep Fusion with Attention**

****

* **Middle Fusion: Short-cut with Addition, Dot-Product, Concatenation**

****

* **Middle Fusion: Short-cut with attention**

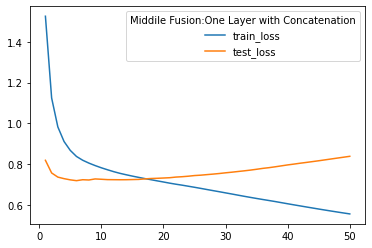
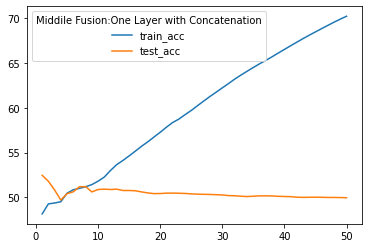
****

**Result**

X-axis: epoch

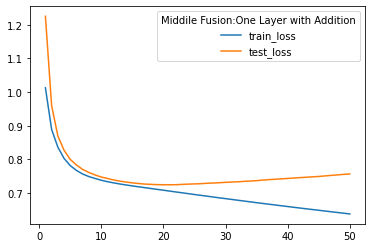
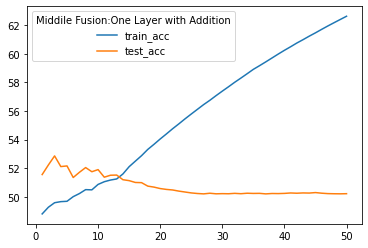
Y-axis: Accuracy or Loss

1. **Middle Fusion: One Layer with Concatenation**



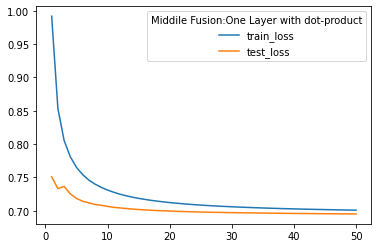
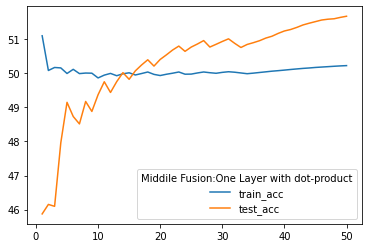
Epoch 50, Loss: 0.5547, Accuracy: 70.25, Test loss: 0.8378, Test accuracy: 49.95

1. **Middle Fusion: One Layer with Addition**

****

Epoch 50, Loss: 0.6377, Accuracy: 62.61, Test loss: 0.7566, Test accuracy: 50.23

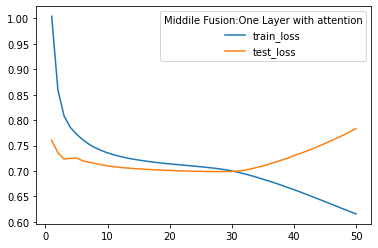
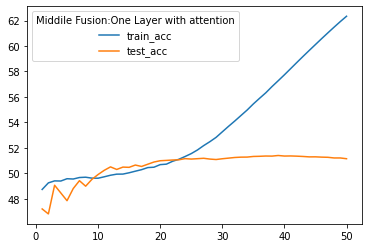
1. **Middle Fusion: One Layer with Dot-Product**

****

Epoch 50, Loss: 0.7008, Accuracy: 50.23, Test loss: 0.6952, Test accuracy: 51.68

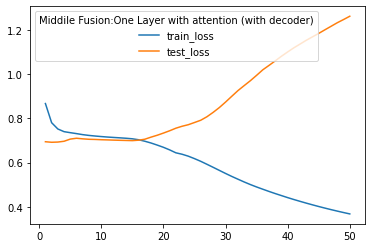
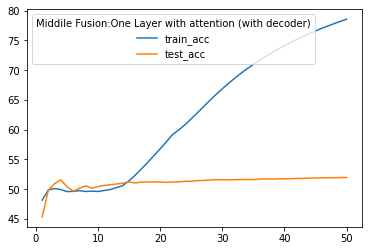
1. **Middle Fusion: One Layer with attention**

* Version 1: without decoder



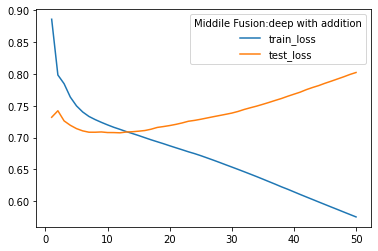
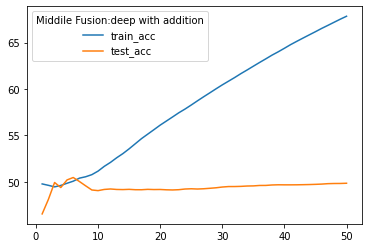
Epoch 50, Loss: 0.6156, Accuracy: 62.34, Test loss: 0.7834, Test accuracy: 51.15

* Version 2: with decoder



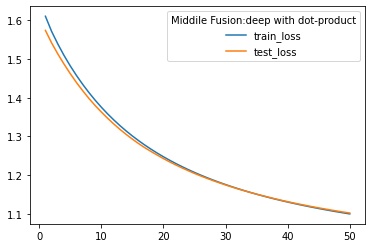
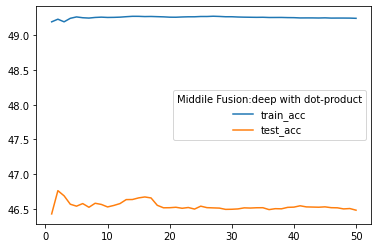
Epoch 50, Loss: 0.3681, Accuracy: 78.54, Test loss: 1.2611, Test accuracy: 51.95

1. **Middle Fusion: Deep with Addition**

****

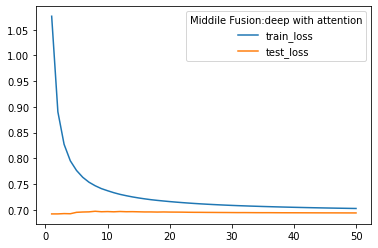
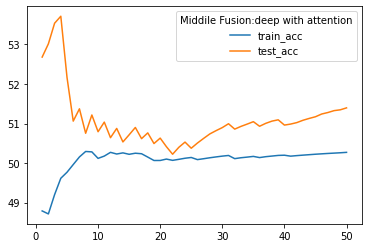
Epoch 50, Loss: 0.5751, Accuracy: 67.85, Test loss: 0.8024, Test accuracy: 49.85

1. **Middle Fusion: Deep with Dot-Product**

****

Epoch 50, Loss: 1.1000, Accuracy: 49.24, Test loss: 1.1023, Test accuracy: 46.48

1. **Middle Fusion: Deep with Attention**

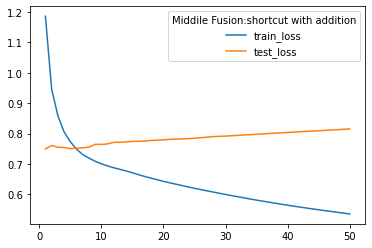
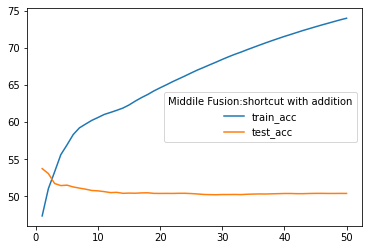
****

Epoch 50, Loss: 0.7023, Accuracy: 50.27, Test loss: 0.6937, Test accuracy: 51.40

1. **Middle Fusion: Deep with concatenation (Splitting into two branches is non-sense)**

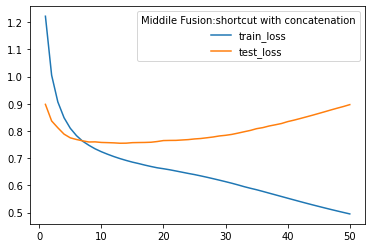
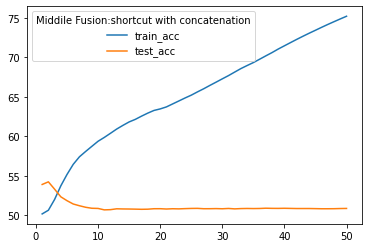
Concatenating feature to text and then split them back to their models, which is not meaningful

1. **Middle Fusion: Short-cut with Addition**

****

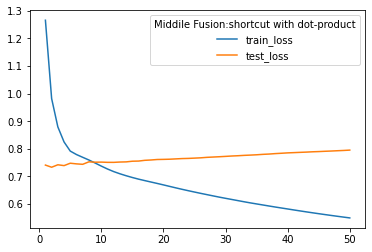
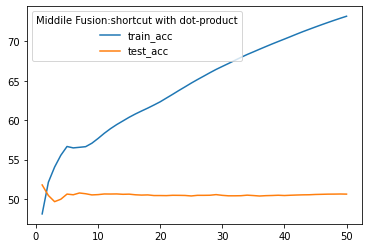
Epoch 50, Loss: 0.5358, Accuracy: 73.96, Test loss: 0.8155, Test accuracy: 50.33

1. **Middle Fusion: Short-cut with Concatenation**

****

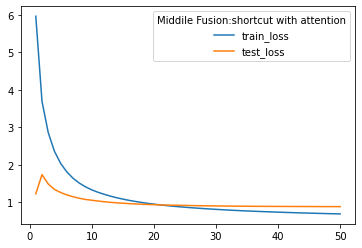
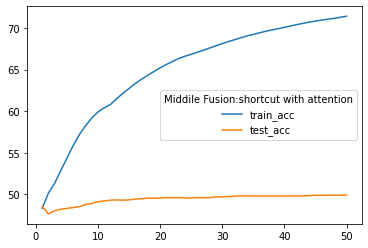
Epoch 50, Loss: 0.4949, Accuracy: 75.22, Test loss: 0.8968, Test accuracy: 50.87

1. **Middle Fusion: Short-cut with Dot-Product**

****

Epoch 50, Loss: 0.5494, Accuracy: 73.18, Test loss: 0.7951, Test accuracy: 50.62

1. **Middle Fusion: Short-cut with Attention**

****

Epoch 50, Loss: 0.6873, Accuracy: 71.42, Test loss: 0.8813, Test accuracy: 49.92