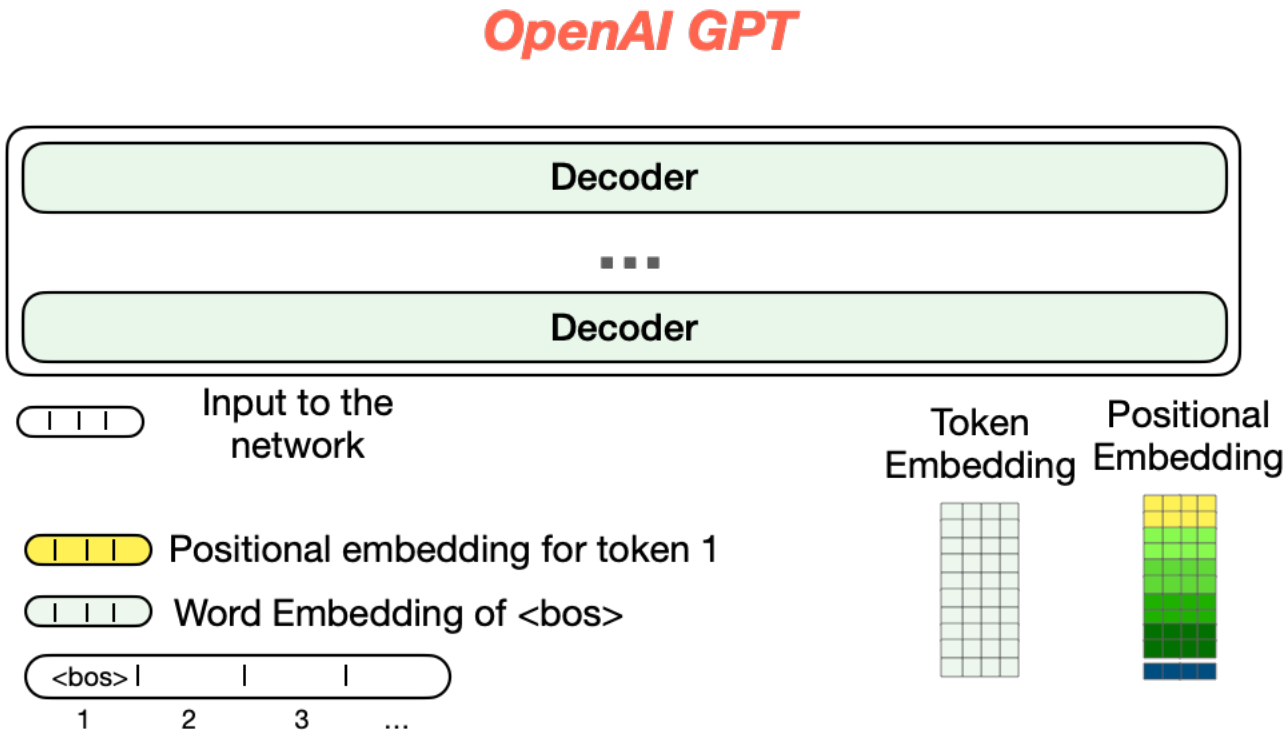
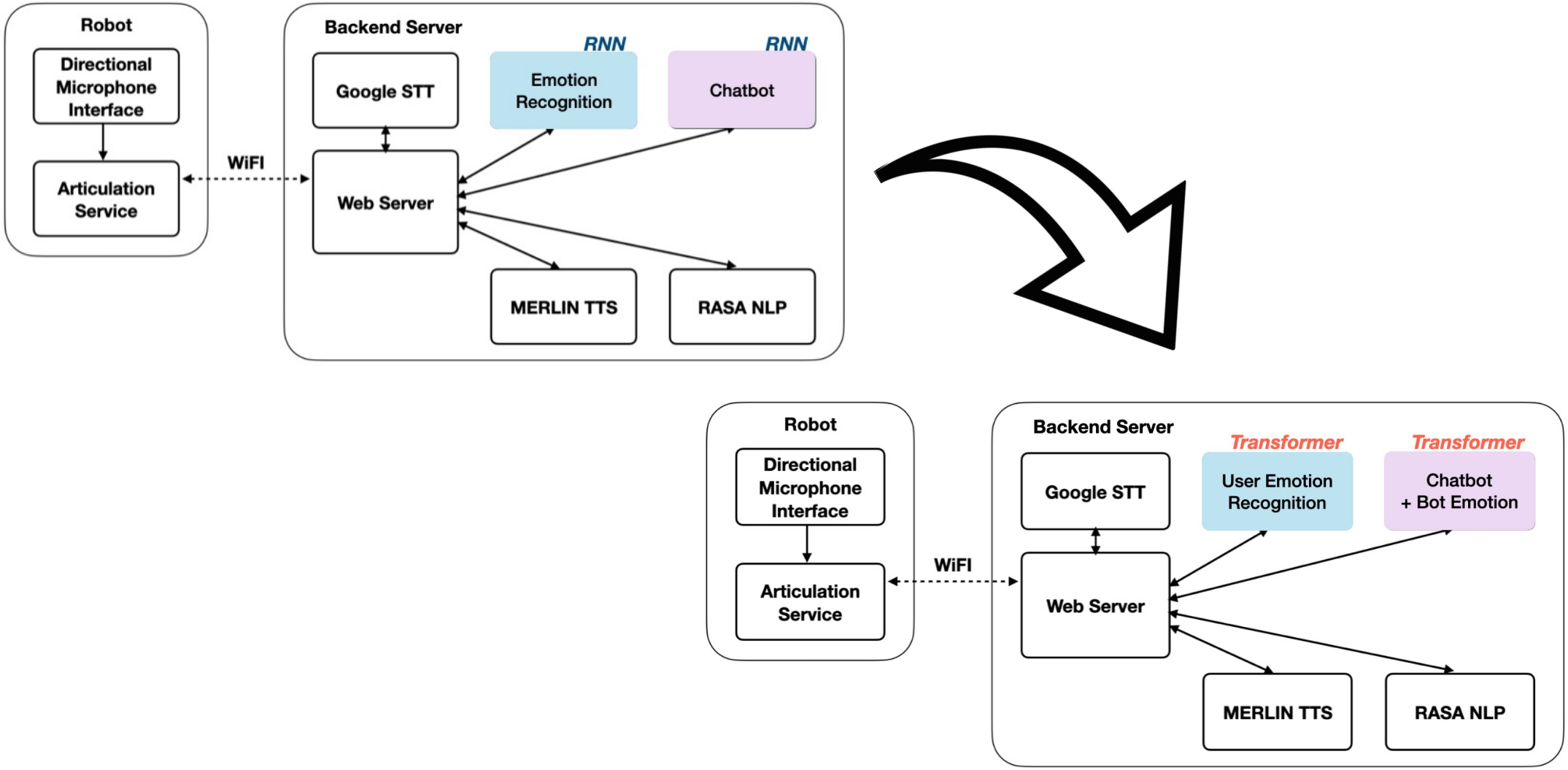
Adaptation of a Transformer based chatbot with Emotion recognition for an Interactive Robot



Big Picture: Chatbot Pipeline Update



Project Goal:

Speaker dialogue

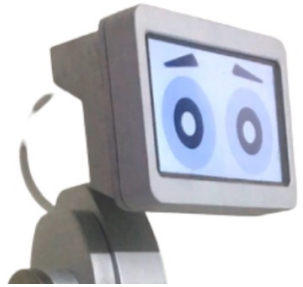
User Emotion Recognition

Speaker Emotion

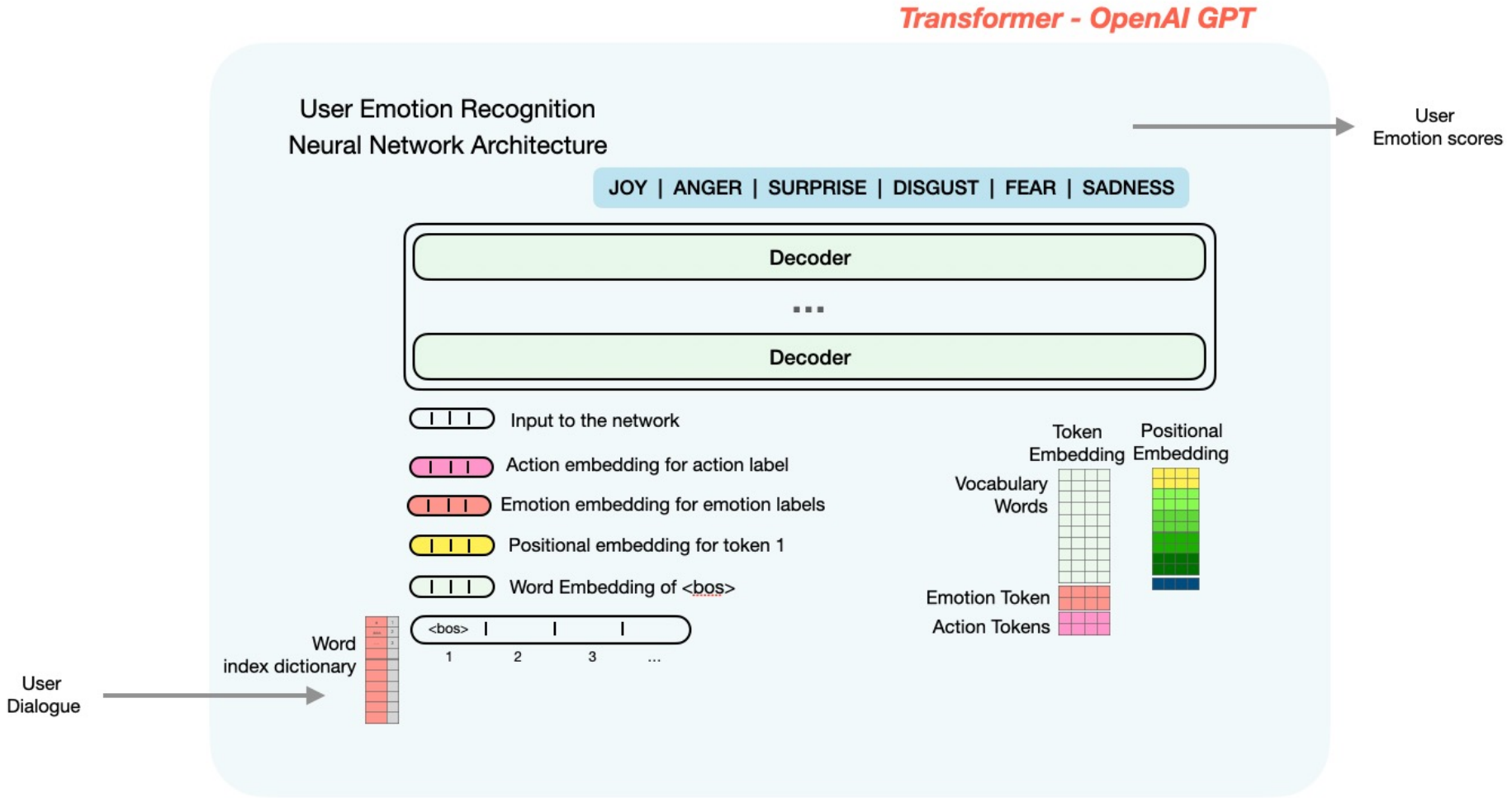
Chatbot reply and emotion

Bot Reply

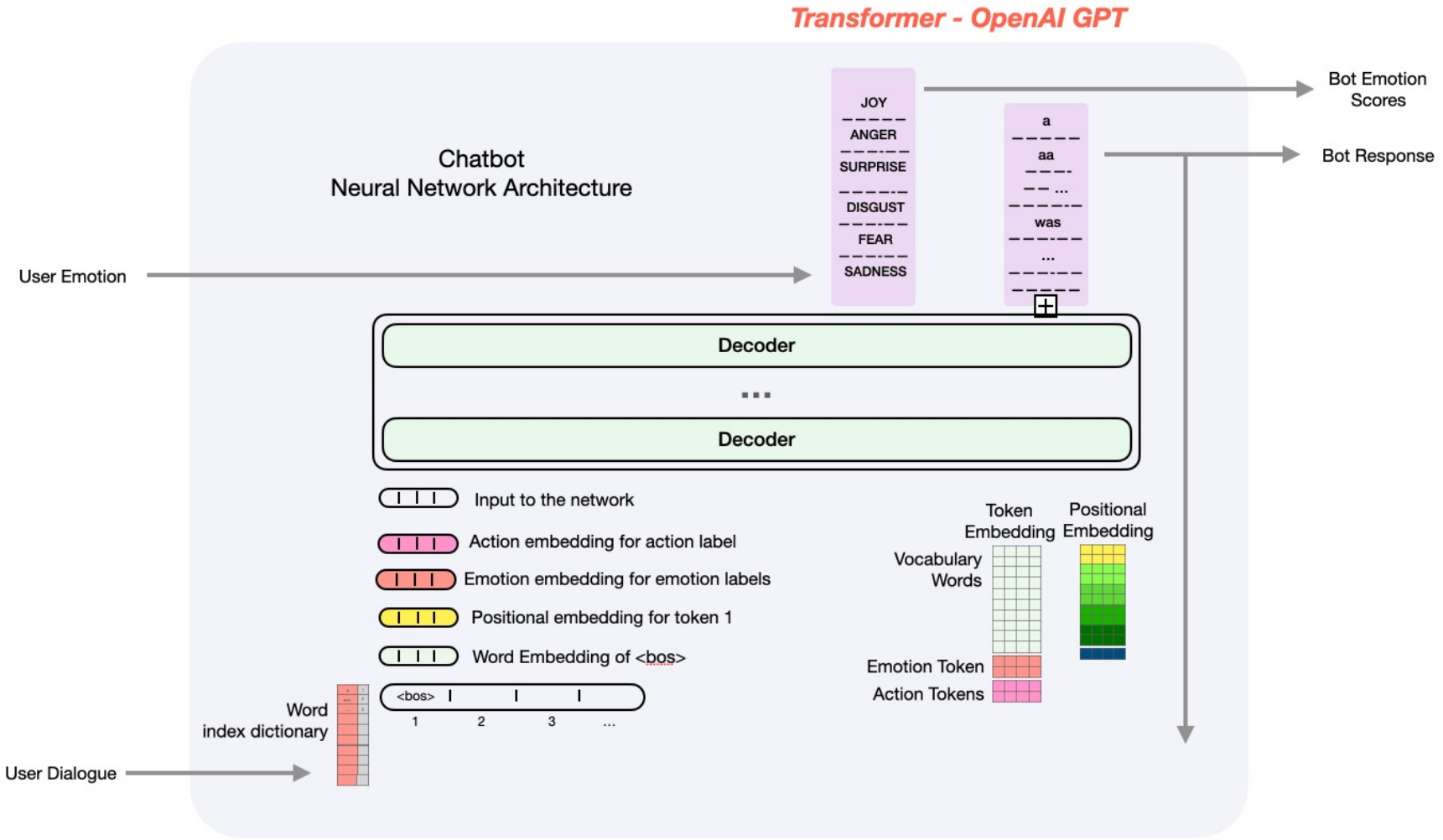
Bot Emotion



Emotion Recognition Neural network Architecture



Chatbot Neural network Architecture



Daily Dialogue Dataset

Structure

```
{ "topic" : "<attitude_and_emotion>"
        {
        "utterances":{
                "history":    [U1, U2, .., Un ],
                "emotion":  [E1, E2, .., En ],
                "act":          [I1, I2, ..,I3],
                "candidates": [U1, U2, .., Un],
                "candidates_emotions":  [E1, E2, .., En ],
                "candidates_acts": [I1, I2, .., I3],
        }

        {"utterances": ...

        }
        ....
        {"utterances": ...

        }
}

"topic" : ...
....
"topic" :...
}
```

Input : Conversation count

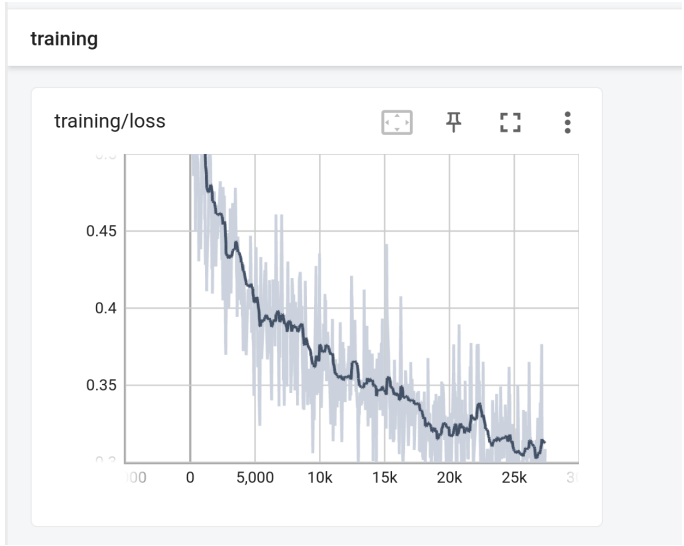
Total Dialogues	13118
Average Speaker Turns	7.9
Average Tokens per Dialogue	114.7
Average Tokens per utterance	14.6

Labels: Speaker Emotion score distribution

	Anger	disgust	Fear	Happiness	Sadness	Surprise
Count	1022	353	74	12885	1150	1823

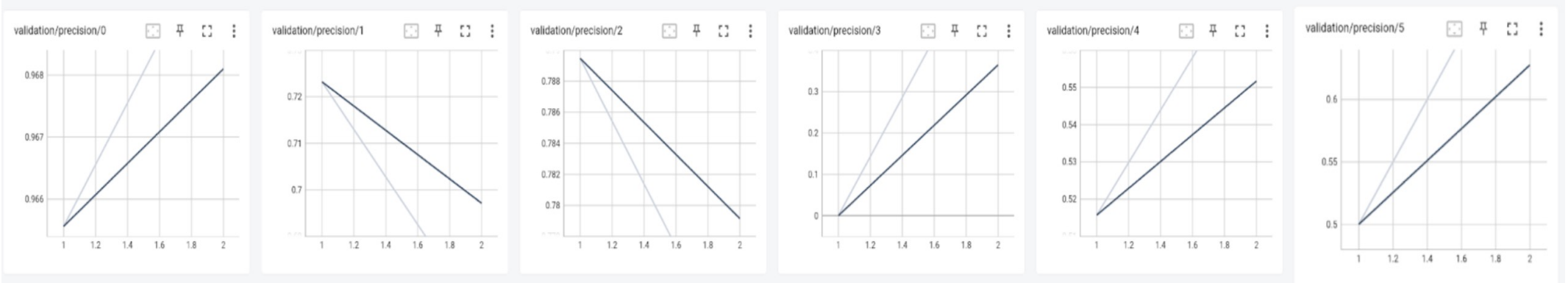
Emotion Recognition Training Results

Training loss



Confusion Matrix

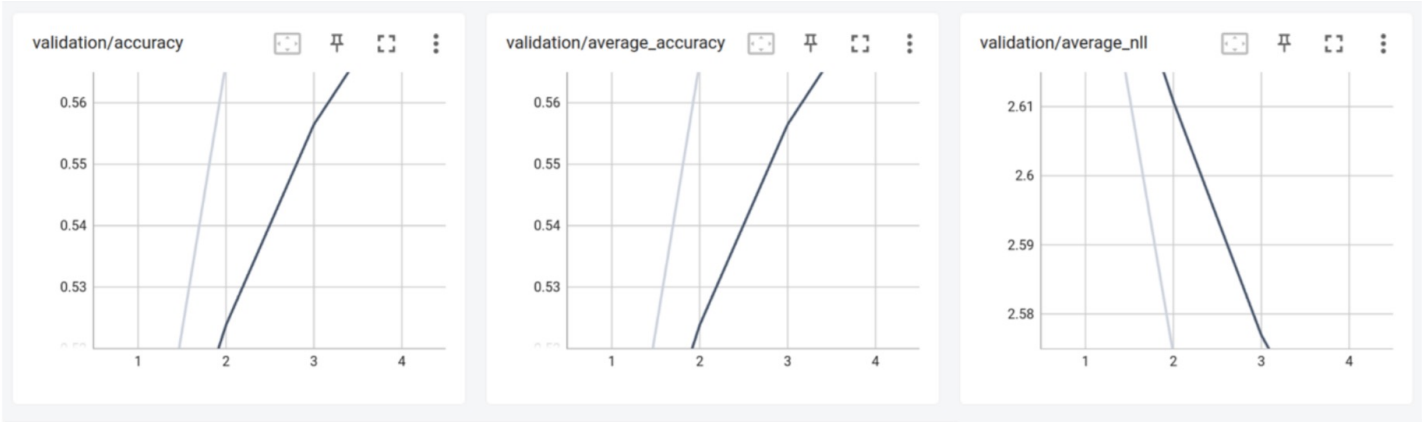
Validation precision

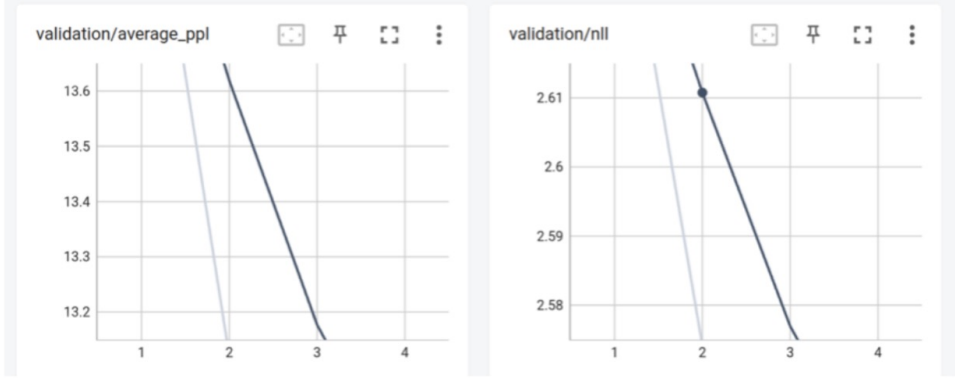


Chatbot Training Results

Training loss

Validation curves

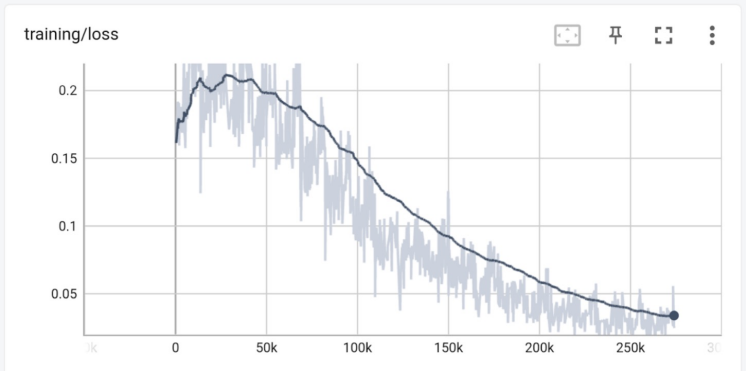




Experiments: Emotion Recognition

Finding early stop epoch count through overfitting

Training loss



2 epochs

Validation curves

Experiments: Interpreting and debugging Neural Network

```
>>> hi
HAPPINESS
>>> how are you
HAPPINESS
>>> I am very happy
HAPPINESS
>>> why are you sad?
SADNESS
>>> Look at the fireworks!
HAPPINESS
>>> This is too noisy.
ANGRY
>>> you are amazing!
HAPPINESS
>>> the movie was scary.
SURPRISE
```

Error output

```
>>> hi
HAPPINESS
>>> how are you
HAPPINESS
>>> I am very happy
HAPPINESS
>>> why are you sad?
SADNESS
>>> Look at the fireworks!
HAPPINESS
>>> This is too noisy.
ANGRY
>>> you are amazing!
HAPPINESS
>>> the movie was scary.
SURPRISE
>>> I am afraid of dark
HAPPINESS
```

Feature weights by NN (2 epochs)

Lime algorithms : 1000 - perturbations

y=HAPPINESS (probability **0.325**, score **-0.896**) top features

Contribution[?]	Feature
+0.272	Highlighted in text (sum)
-0.104	
-0.163	<BIAS>
-0.288	
-0.293	I
-0.320	

i am afraid of dark

y=SURPRISE (probability **0.141**, score **-1.943**) top features

Contribution[?]	Feature
+0.508	
+0.345	
-0.389	
-0.405	Highlighted in text (sum)
-0.470	
-0.696	<BIAS>
-0.836	I

i am afraid of dark

y=DISGUST (probability **0.162**, score **-1.782**) top features

Contribution[?]	Feature
+0.751	
+0.605	
-0.360	Highlighted in text (sum)
-0.561	
-0.662	I
-0.692	
-0.863	<BIAS>

i am afraid of dark

y=ANGRY (probability **0.061**, score **-2.858**) top features

Contribution[?]	Feature
+0.627	
+0.240	
-0.533	
-0.718	<BIAS>
-0.812	Highlighted in text (sum)
-0.829	I
-0.834	

i am afraid of dark

y=SADNESS (probability **0.233**, score **-1.341**) top features

Contribution[?]	Feature
+0.419	
+0.187	
-0.168	
-0.214	I
-0.307	Highlighted in text (sum)
-0.547	<BIAS>
-0.710	

i am afraid of dark

y=FEAR (probability **0.079**, score **-2.584**) top features

Contribution[?]	Feature
+0.785	
+0.503	
-0.639	Highlighted in text (sum)
-0.670	
-0.693	<BIAS>
-0.929	
-0.941	I

i am afraid of dark

Experiments: Interpreting the overfitted Neural Network memorizing training data

Overfitted Network Output

```
>>> hi how are you
HAPPINESS
>>> I need to go and get dinner.
HAPPINESS
>>> But, I am afraid of dark
FEAR
>>>
```

Feature weights by NN (20 epochs)

Lime algorithms : 1000 - perturbations

y=HAPPINESS (probability **0.020**, score **-3.861**) top features

Contribution?	Feature
+0.709	Highlighted in text (sum)
+0.653	
-0.395	<BIAS>
-0.549	
-1.055	
-1.257	
-1.967	I

i am afraid of dark

y=DISGUST (probability **0.000**, score **-7.601**) top features

Contribution?	Feature
-0.212	
-0.287	
-0.511	<BIAS>
-1.198	
-1.265	Highlighted in text (sum)
-1.735	
-2.393	I

i am afraid of dark

y=SURPRISE (probability **0.001**, score **-7.468**) top features

Contribution?	Feature
+0.198	
-0.250	
-0.481	<BIAS>
-0.552	
-0.828	
-2.600	I
-2.956	Highlighted in text (sum)

i am afraid of dark

y=ANGRY (probability **0.000**, score **-9.069**) top features

Contribution?	Feature
-0.184	
-0.194	
-0.515	<BIAS>
-1.274	
-1.378	
-2.310	I
-3.214	Highlighted in text (sum)

i am afraid of dark

y=SADNESS (probability **0.023**, score **-3.719**) top features

Contribution?	Feature
+1.026	
+0.130	
-0.347	Highlighted in text (sum)
-0.553	<BIAS>
-0.949	
-1.317	I
-1.708	

i am afraid of dark

y=FEAR (probability **0.956**, score **4.208**) top features

Contribution?	Feature
+2.573	
+2.095	I
+1.521	Highlighted in text (sum)
+0.093	
-0.350	<BIAS>
-0.683	
-1.041	

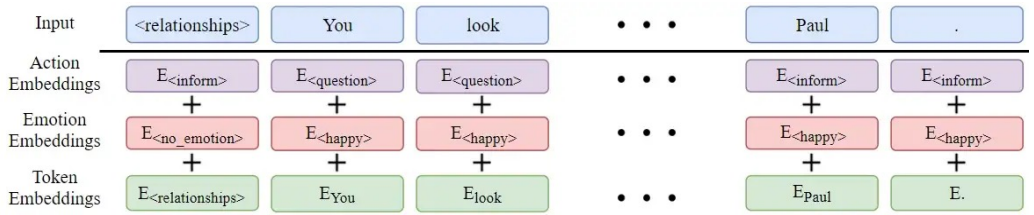
i am afraid of dark

Daily-Dialog training set search

```
351  <fear>
352  Oh , God . It ' s late . I ' m afraid I have to leave .
353  <fear>
354  Oh , God . It ' s late . I ' m afraid I have to leave .
355  <fear>
356  It must be a haunted house . Are you frightened ?
357  <fear>
358  I'm afraid of the darkness .
359  <fear>
360  I'm afraid of the darkness .
361  <fear>
362  You think it's funny . I'm terrified .
363  <fear>
```

Problems faced

1) Figuring out proper inputs to the network for training and inference



2) High Training Time : 46000 training samples

~ 20min per epoch : Emotion recognition NN
~ 3hours per epoch : Chatbot

```
Epoch [1/10]: [46791/46791] 100%|                                                                                , loss=0.382 [3:12:08<00:00]
INFO:ignite.engine.engine.Engine:Engine run complete. Time taken: 00:36:08.694
Epoch [1/10]: [46791/46791] 100%|                                                                                , loss=0.382 [3:12:08<00:00]
INFO:ignite.engine.engine.Engine:Epoch[1] Complete. Time taken: 03:12:11.990
Epoch [2/10]: [46791/46791] 100%|                                                                                , loss=0.296 [2:50:25<00:00]
```

Thank you.