Being Funny Is Hard
Humor Response Generation

W266 - NLP with Deep Learning
Final Project

Tyler Ryu , Tsung-Chin Han

Background & Introduction

Humor is a staple of human communication

- Jokes are used as fillers in communication to keep listeners engaged
- NLP demand is rapidly growing (Alexa, Google Assistant, Siri) and with it recreational use cases are growing as well (Asking Siri to tell a joke)
- A need to make "fun" NLP responses exist!

Goal

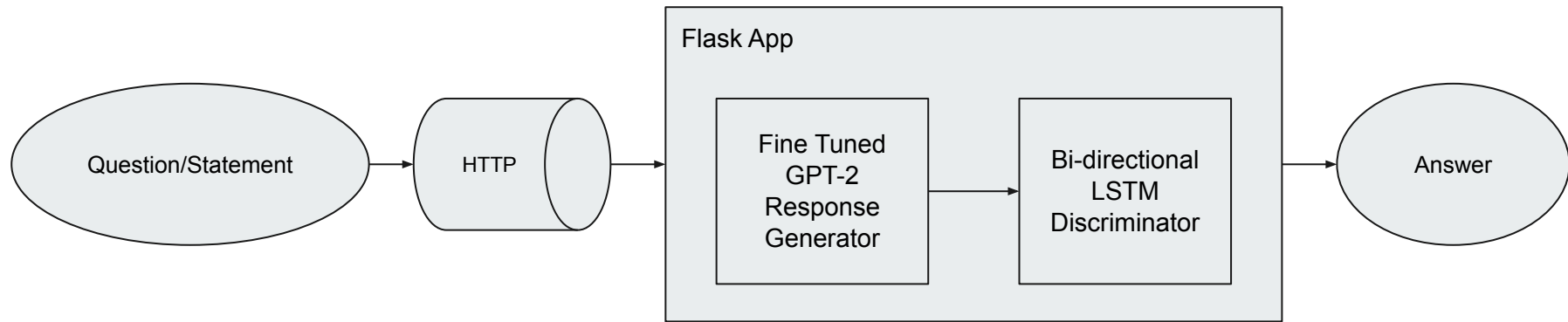Why did the chicken cross the road?

Yo mama: (Burps) What chicken?

Chickens often live next to roads and therefore have a chance it may have crossed unknowingly

I like ice cream

Perfect Answer: Subvert expectation, but should also make sense
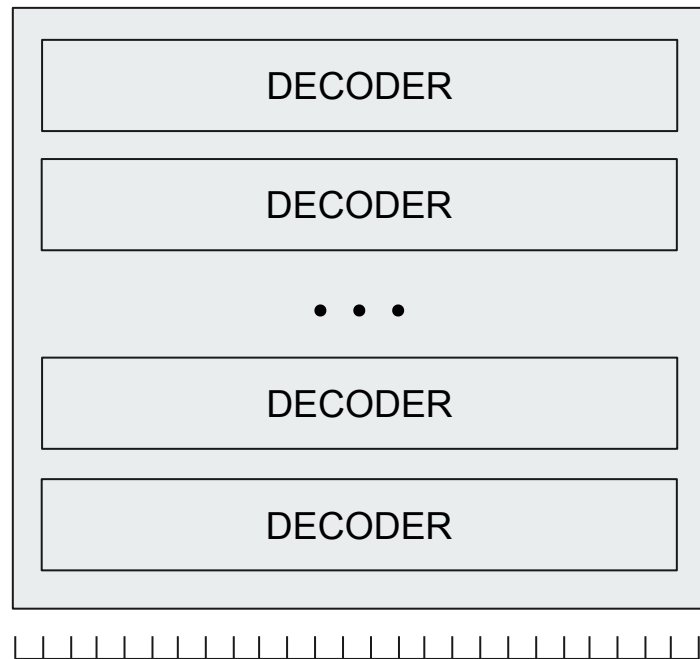
Our Implementation

GPT-2

- Why GPT-2?
  - Some responses are "improv" like and GPT-2s auto-regressiveness helps recreate this effect
  - Easy to fine tune and produces natural results

24 DECODER BLOCKS

DECODER

DECODER

• • •

DECODER

DECODER

Model Dimensionality: 1024

Fine Tuned GPT-2

- 345M Parameter Model
  - While larger parameter models perform better, they are more expensive to train
- Trained on 84,000 + input + response pair from Reddit r/Jokes
- 3000 Epochs
- 47/100 Sample Question-Answer pairing had human concensus

Output Samples

What's the best thing about living in Switzerland?	**Option 1:** It's the cheese. **Option 2:** They are all so close! **Option 3:** I'm the only person there who knows what a good German is.

Output Samples

| Why was the orchestra teacher angry?	Because she said he had no legs!
Where do you find a dog with no legs?	In the bathroom.
How do you make a tissue dance?	You make it go round with all of its legs, and you slap the person with the wrong leg."
Those ads actually ran.	The new ad features a smiling man and a handsome woman both lying on a bed. The woman is excited. "Do I look fat?" "Yes, do you?" "No, it's just you're so pretty that if I was with you I would have to shave my legs."

Output Samples

How do you encourage a bear to eat cheese?	Option 1:
	Option 2:
	Option 3:

Bidirectional LSTM on top of GloVe

- Pre-train GloVe embedding + Bidirectional LSTM + dense layers

- Trained on 19,4553 **(input + response, score)** pair from Reddit data

- 3.5 M Parameters with 10 Epochs

- Source Data

  - Turning each score to binary labels

  - Formatting altered to include title and body tags ("title: " + input + "body: " + response)

- Tested on the Fine-Tuned GPT-2 output pairs (requested from Flask app)

LSTM Training Results
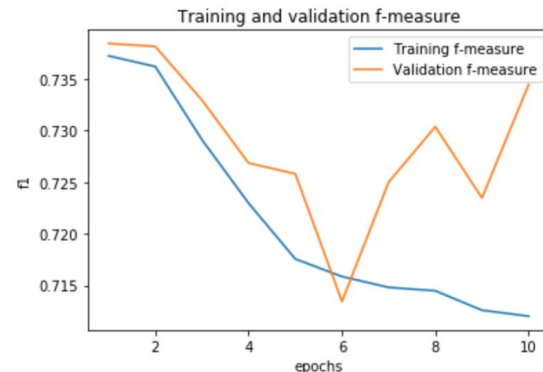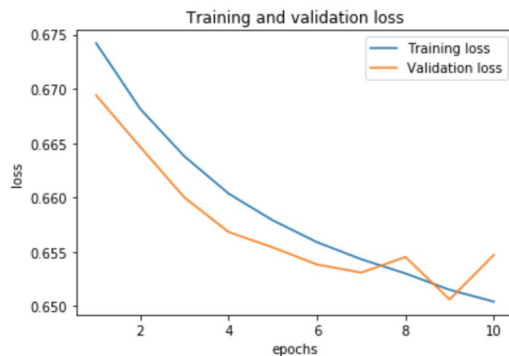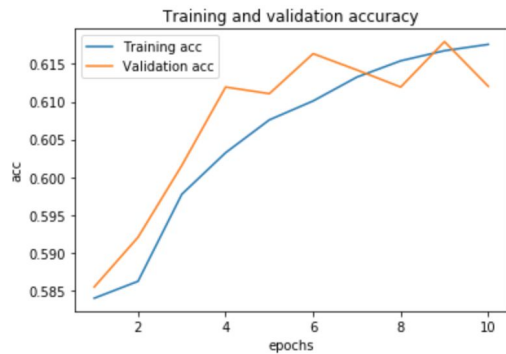
Output

	jokes	is_funny	proba
0	title: Why is it so hard to break up with a Ja...	0	0.474580
1	title: Why did the chicken cross the road? \nb...	1	0.687178
2	title: Why did the producers of 007 films use ...	1	0.515699
3	title: I feel so bad for the necrophiliac's si...	0	0.463882
4	title: What's the difference between you and a...	1	0.716760

Human Evaluation Comparison

- Survey - each joke we generate 3 pair of reponses - choose the best option

- Model's Predictions vs. Human's Evaluation

  - 8 participants; more than half of the people agree on question-answering paring

  - Correlations with the human consensus

  - Compared to the machine: 21% accuracy out of 47 pairs

Conclusion & Future Study

- Pending - More survey submissions expected

- Future - Use bigger versions of GPT-2 (1.5B)

- Future - Improve classification (Random Forests & Gradient Boosting, 65%, see ref)

- Future - Study the ethical implications of biases that are accumulated during pre-training

Thank You !

Reference

Classification Benchmark: https://medium.com/datadriveninvestor/the-nlp-of-humor-20deb591f953