Deep Learning:

Sentiment Classification on Amazon Product Reviews

Jacob Zimmer
University of Michigan BSI Class of 2019
zimmerja@umich.edu | 734-634-7274

Why this project?

- **Why build a Neural Net?**
  - Good opportunity to learn
  - Numerous potential applications
- **Why Sentiment Analysis & Why Amazon Product Reviews?**
  - Applications to past NLP work
  - Similarity to Reddit text data

Key Software, Tools, & Resources

- numPy
- Sklearn
- Keras & Tensorflow
- Plotly
- “Deep Learning with Python” by François Chollet

DATA

- Source: Kaggle & Xiang Zhang (UC-Berkeley)
  - 4 million total observations
- Amazon User Reviews: Text and Ratings
  - 3.6 million observations in training dataset
  - 400k observations in test dataset
  - Validation split of 10% from within training data

Sample Review Data

```
'__label__1 not playable since 1 hour 30 minutes: The DVD is not playable since 1 hour 30 minutes, then the picture b
ecomes stopping and going and keeps that way.\n'
```

Data Preprocessing

- Decompress bz2 file, convert to utf-8
- Split review texts from ratings
- Remove numbers and URLs from text with RegEx
- Shuffle train and test sets

Sample Review Data: Post-Preprocessing

```
('not playable since 0 hour 00 minutes: the dvd is not playable since 0 hour 00 minutes, then the picture becomes sto
pping and going and keeps that way.\n',
 0)
```

More Data Preprocessing

- Tokenize train text data
- Create integer encodings for train and test text data
- Pad train and test datasets
    - Ensures uniform shape

Building the Neural Net

Selecting NN Architecture

Convolutional Neural Net

- Generally considered ideal for image & video analysis
- Good for tasks where feature detection is more important
- Relatively faster, more efficient than LSTM
    - Conv1D layers reduce train time

Recurrent Neural Net

- Generally considered ideal for text and speech analysis
- Good for tasks where document length is important
- Loses info from old observations quickly
    - LSTM very expensive

Brief Background: Neural Nets

- Map inputs to targets via simple data transformations
- Oversimplified:
  - Output = Σ (Weight Matrix * Input) + Bias
  - Initialization values of Weight Matrix arbitrarily chosen
- Loss Function measures distance of model output from actual values
  - Guides adjustment of Weight Matrix

Constructing the Network

- Goals:
  - Maximize Validation Accuracy
  - Minimize Validation Loss
    - Measures network performance

Layer (type)	Output Shape	Param #
input_18 (InputLayer)	(None, 200)	0
embedding_18 (Embedding)	(None, 200, 64)	640000
dropout_18 (Dropout)	(None, 200, 64)	0
batch_normalization_73 (Batc	(None, 200, 64)	256
conv1d_73 (Conv1D)	(None, 200, 32)	14368
batch_normalization_74 (Batc	(None, 200, 32)	128
conv1d_74 (Conv1D)	(None, 200, 32)	3104
batch_normalization_75 (Batc	(None, 200, 32)	128
conv1d_75 (Conv1D)	(None, 200, 32)	3104
batch_normalization_76 (Batc	(None, 200, 32)	128
conv1d_76 (Conv1D)	(None, 200, 2)	66
global_average_pooling1d_18	(None, 2)	0
activation_18 (Activation)	(None, 2)	0

Total params: 661,282
Trainable params: 660,962
Non-trainable params: 320

Constructing the Network

Layer (type)	Output Shape	Param #
input_18 (InputLayer)	(None, 200)	0
embedding_18 (Embedding)	(None, 200, 64)	640000
dropout_18 (Dropout)	(None, 200, 64)	0
batch_normalization_73 (Batc	(None, 200, 64)	256
conv1d_73 (Conv1D)	(None, 200, 32)	14368
batch_normalization_74 (Batc	(None, 200, 32)	128
conv1d_74 (Conv1D)	(None, 200, 32)	3104
batch_normalization_75 (Batc	(None, 200, 32)	128
conv1d_75 (Conv1D)	(None, 200, 32)	3104
batch_normalization_76 (Batc	(None, 200, 32)	128
conv1d_76 (Conv1D)	(None, 200, 2)	66
global_average_pooling1d_18	(None, 2)	0
activation_18 (Activation)	(None, 2)	0

Total params: 661,282
Trainable params: 660,962
Non-trainable params: 320

One-Hot Encoding vs Embedding

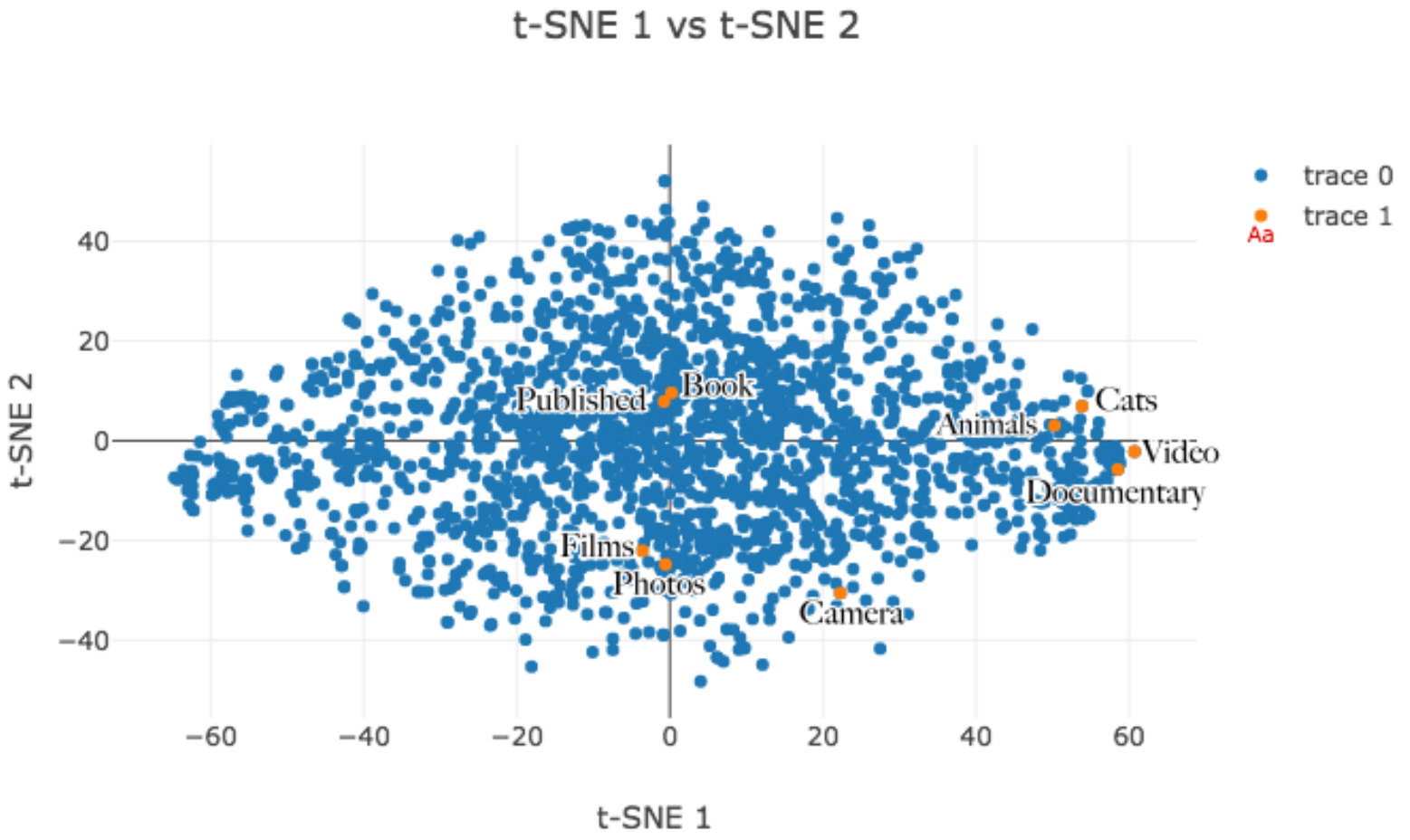
One-Hot Encoding

- Most common approach
- High-dimensional, sparse
- Requires less data, training time

Embedding Layer

- Dense vector representation
- Vectors updated *during* training
- Allows for visualizing relationships between vectorized text

Visualizing Word Embeddings

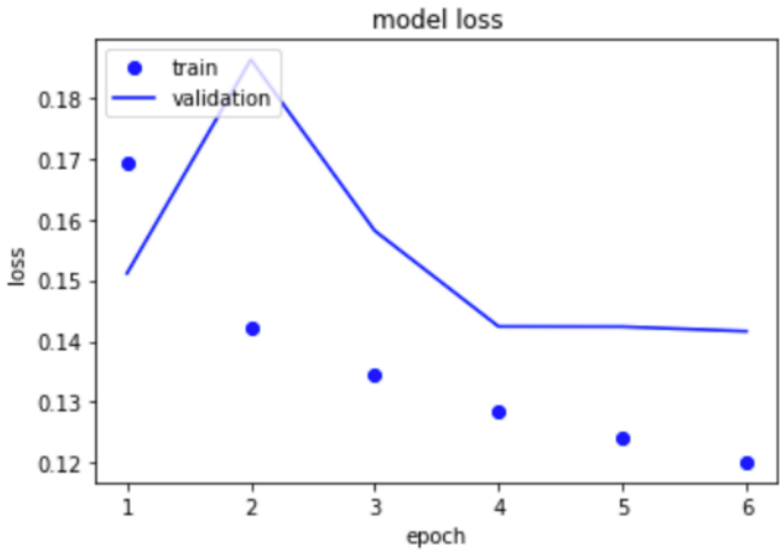


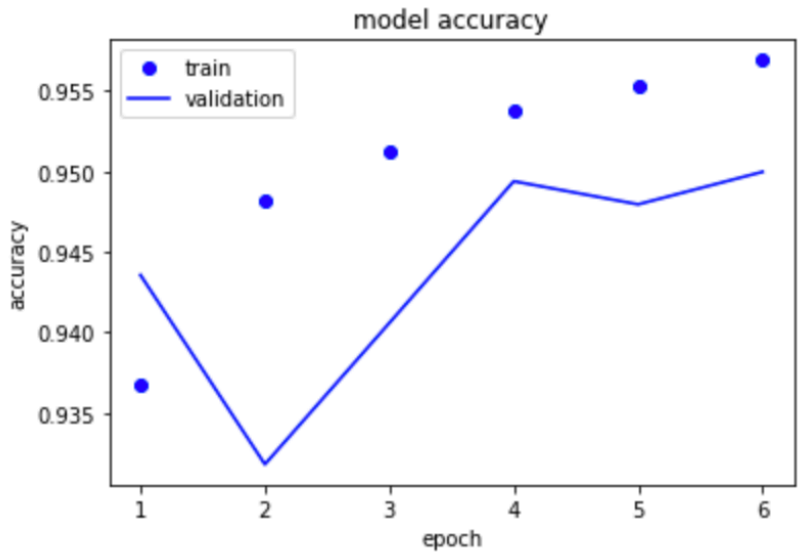
Constructing the Network

Layer (type)	Output Shape	Param #
input_18 (InputLayer)	(None, 200)	0
embedding_18 (Embedding)	(None, 200, 64)	640000
dropout_18 (Dropout)	(None, 200, 64)	0
batch_normalization_73 (Batc	(None, 200, 64)	256
conv1d_73 (Conv1D)	(None, 200, 32)	14368
batch_normalization_74 (Batc	(None, 200, 32)	128
conv1d_74 (Conv1D)	(None, 200, 32)	3104
batch_normalization_75 (Batc	(None, 200, 32)	128
conv1d_75 (Conv1D)	(None, 200, 32)	3104
batch_normalization_76 (Batc	(None, 200, 32)	128
conv1d_76 (Conv1D)	(None, 200, 2)	66
global_average_pooling1d_18	(None, 2)	0
activation_18 (Activation)	(None, 2)	0

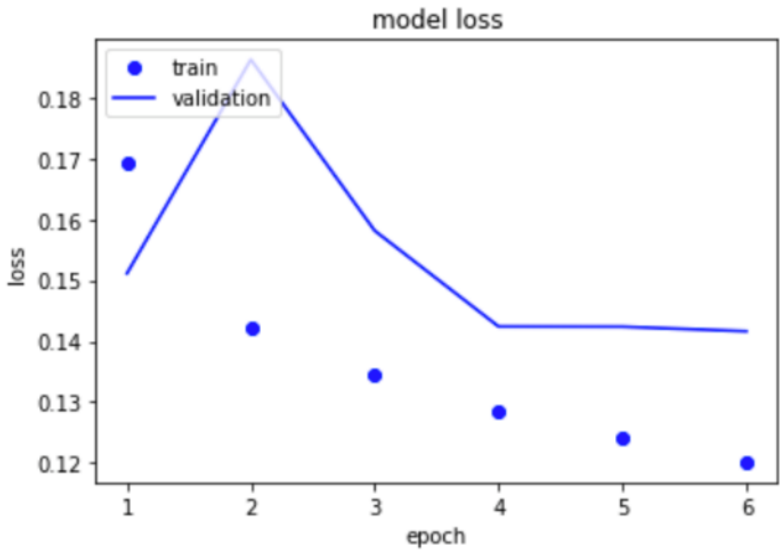
Total params: 661,282
Trainable params: 660,962
Non-trainable params: 320

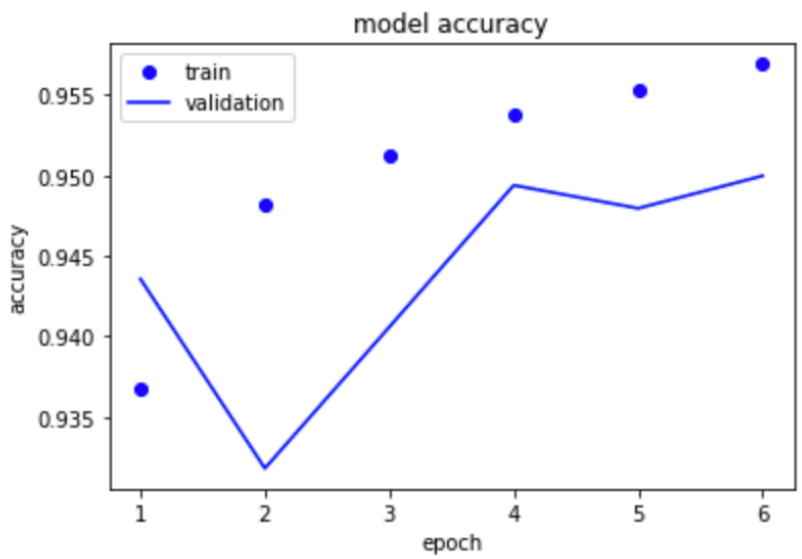
Evaluating Network Performance





Evaluating Network Performance





Final Test Dataset Loss: 0.1615

Final Test Dataset Accuracy: 93.96%

(Null Accuracy of Test Dataset: 50%)

Applying the Model

Positive Review Example	"great product: bought this for my 0 year old daughter when she started to learn to swim. a great product and works as described. it's nice because it's adjustable and all the foam pads are removeable so you can adjust as the child progresses." – **Actual Value: 1**
Negative Review Example	"problems: i have bought this player but till now i am unable to play any mp0 disc. i tried single session, multisession, various types of mp0 files etc. no effect. i have contacted samsung services, without any significant help. maybe someone can help me, how to record the cd-rom to be "visible" by this player ?" – **Actual Value: 0**
Ambiguous Review Example	"i guess for those enthusiest's that love this movie like i do...: i have a vhs to dvd copy i made long ago that looks exactly like what they did with this movie.do not get your hopes up! (has bootleg written all over it)" – **Actual Value: 0**

Applying the Model

Positive Review Example

```
ex1 = tokenizer.texts_to_sequences(train_text[good])
ex1 = pad_sequences(ex1, maxlen=maxlen, padding='post')
prediction = med_model2.predict(np.array(ex1))
class_pred = prediction.argmax(axis=-1)
print(class_pred[0])
```

Negative Review Example

```
ex2 = tokenizer.texts_to_sequences(train_text[bad])
ex2 = pad_sequences(ex2, maxlen=maxlen, padding='post')
prediction2 = med_model2.predict(np.array(ex2))
class_pred2 = prediction2.argmax(axis=-1)
print(class_pred2[0])
```

Ambiguous Review Example

```
ex3 = tokenizer.texts_to_sequences(train_text[amb])
ex3 = pad_sequences(ex3, maxlen=maxlen, padding='post')
prediction3 = med_model2.predict(np.array(ex3))
class_pred3 = prediction3.argmax(axis=-1)
print(class_pred3[0])
```

Applying the Model

Model Prediction
(1 = Positive, 0 = Negative):

Positive Review Example	"great product: bought this for my 0 year old daughter when she started to learn to swim. a great product and works as described. it's nice because it's adjustable and all the foam pads are removeable so you can adjust as the child progresses." – **Actual Value: 1**
Negative Review Example	"problems: i have bought this player but till now i am unable to play any mp0 disc. i tried single session, multisession, various types of mp0 files etc. no effect. i have contacted samsung services, without any significant help. maybe someone can help me, how to record the cd-rom to be "visible" by this player ?" – **Actual Value: 0**
Ambiguous Review Example	"i guess for those enthusiest's that love this movie like i do...: i have a vhs to dvd copy i made long ago that looks exactly like what they did with this movie.do not get your hopes up! (has bootleg written all over it)" – **Actual Value: 0**

Applying the Model

		Model Prediction (1 = Positive, 0 = Negative):
Positive Review Example	"great product: bought this for my 0 year old daughter when she started to learn to swim. a great product and works as described. it's nice because it's adjustable and all the foam pads are removeable so you can adjust as the child progresses." – **Actual Value: 1**	1
Negative Review Example	"problems: i have bought this player but till now i am unable to play any mp0 disc. i tried single session, multisession, various types of mp0 files etc. no effect. i have contacted samsung services, without any significant help. maybe someone can help me, how to record the cd-rom to be "visible" by this player ?" – **Actual Value: 0**	
Ambiguous Review Example	"i guess for those enthusiest's that love this movie like i do...: i have a vhs to dvd copy i made long ago that looks exactly like what they did with this movie.do not get your hopes up! (has bootleg written all over it)" – **Actual Value: 0**	

Applying the Model

		Model Prediction (1 = Positive, 0 = Negative):
Positive Review Example	"great product: bought this for my 0 year old daughter when she started to learn to swim. a great product and works as described. it's nice because it's adjustable and all the foam pads are removeable so you can adjust as the child progresses." – **Actual Value: 1**	1
Negative Review Example	"problems: i have bought this player but till now i am unable to play any mp0 disc. i tried single session, multisession, various types of mp0 files etc. no effect. i have contacted samsung services, without any significant help. maybe someone can help me, how to record the cd-rom to be "visible" by this player ?" – **Actual Value: 0**	0
Ambiguous Review Example	"i guess for those enthusiest's that love this movie like i do...: i have a vhs to dvd copy i made long ago that looks exactly like what they did with this movie.do not get your hopes up! (has bootleg written all over it)" – **Actual Value: 0**	

Applying the Model

		Model Prediction (1 = Positive, 0 = Negative):
Positive Review Example	"great product: bought this for my 0 year old daughter when she started to learn to swim. a great product and works as described. it's nice because it's adjustable and all the foam pads are removeable so you can adjust as the child progresses." – **Actual Value: 1**	1
Negative Review Example	"problems: i have bought this player but till now i am unable to play any mp0 disc. i tried single session, multisession, various types of mp0 files etc. no effect. i have contacted samsung services, without any significant help. maybe someone can help me, how to record the cd-rom to be "visible" by this player ?" – **Actual Value: 0**	0
Ambiguous Review Example	"i guess for those enthusiest's that love this movie like i do...: i have a vhs to dvd copy i made long ago that looks exactly like what they did with this movie.do not get your hopes up! (has bootleg written all over it)" – **Actual Value: 0**	0

Possible Improvements

- Test wider range of parameters
- Construct RNN and compare performance
- Identify important features within the data

Business Applications of Deep Learning

Advantages

- Unsupervised feature learning
- State-of-the-art solution for problem spaces within several domains
- Models can quickly & easily be adapted to new tasks

Disadvantages

- Requires a large amount of data
- Computationally expensive
- Requires extensive engineering
- Can be difficult to interpret

Business Applications of Deep Learning

- Content Recommendation Systems
- Image Detection & Object Classification
- Cybersecurity
- Customer Service
- Language translators
- Possible medical applications


THANK YOU

```
def med_functional_api(max_len, max_feat, embed_sz, lr):
    input = Input(shape=(max_len,))
    net = Embedding(max_feat, embed_sz)(input)
    net = Dropout(0.2)(net)
    net = BatchNormalization()(net)
    net = Conv1D(32, 7, padding='same', activation='relu')(net)
    net = BatchNormalization()(net)
    net = Conv1D(32, 3, padding='same', activation='relu')(net)
    net = BatchNormalization()(net)
    net = Conv1D(32, 3, padding='same', activation='relu')(net)
    net = BatchNormalization()(net)

    net = Conv1D(2, 1)(net)
    net = GlobalAveragePooling1D()(net)
    output = Activation('softmax')(net)
    model = Model(inputs = input, outputs = output)
    opt = keras.optimizers.Adam(lr=lr)
    model.compile(optimizer=opt, loss='sparse_categorical_crossentropy', metrics=['acc'])
    return model
```