Universidad del Valle de Guatemala
Data Science
Sección 20

Laboratorio No. 4

Carlos Edgardo López Barrera 21666

Brian Anthony Carrillo Monzon - 21108

Guatemala, 08 de septiembre del 2024

Entrenamiento y resultados modelo simple

- Embedding Layer: Se convierten las palabras en vectores de 128 dimensiones, usando las 20,000 palabras más frecuentes del dataset IMDB.
- LSTM Layer: Tiene 128 unidades con dropout=0.2 y recurrent_dropout=0.2 para regularizar el modelo y reducir el riesgo de sobreajuste.
- Dense Layer (Output): Una sola neurona con activación sigmoid para realizar la clasificación binaria (positivo/negativo).
- Optimizer: Adam.
- Loss Function: binary_crossentropy.

```
✓ 18m 35.0s

Epoch 1/15
782/782 - 64s - 82ms/step - accuracy: 0.7748 - loss: 0.4690 - val_accuracy: 0.8167 - val_loss: 0.4039
Epoch 2/15
782/782 - 58s - 74ms/step - accuracy: 0.8748 - loss: 0.3059 - val_accuracy: 0.8222 - val_loss: 0.3926
Epoch 3/15
782/782 - 58s - 74ms/step - accuracy: 0.9133 - loss: 0.2198 - val_accuracy: 0.8360 - val_loss: 0.4064
Epoch 4/15
782/782 - 58s - 74ms/step - accuracy: 0.9415 - loss: 0.1572 - val_accuracy: 0.8304 - val_loss: 0.4883
Epoch 5/15
782/782 - 59s - 76ms/step - accuracy: 0.9584 - loss: 0.1155 - val_accuracy: 0.8216 - val_loss: 0.5274
Epoch 6/15
782/782 - 58s - 74ms/step - accuracy: 0.9720 - loss: 0.0781 - val_accuracy: 0.8218 - val_loss: 0.6785
Epoch 7/15
782/782 - 63s - 80ms/step - accuracy: 0.9815 - loss: 0.0556 - val_accuracy: 0.8063 - val_loss: 0.7687
Epoch 8/15
782/782 - 62s - 79ms/step - accuracy: 0.9882 - loss: 0.0377 - val_accuracy: 0.8083 - val_loss: 0.8335
Epoch 9/15
782/782 - 61s - 77ms/step - accuracy: 0.9906 - loss: 0.0291 - val_accuracy: 0.8158 - val_loss: 0.8955
Epoch 10/15
782/782 - 90s - 116ms/step - accuracy: 0.9924 - loss: 0.0238 - val_accuracy: 0.8046 - val_loss: 0.9312
Epoch 11/15
782/782 - 94s - 120ms/step - accuracy: 0.9949 - loss: 0.0159 - val_accuracy: 0.8052 - val_loss: 1.0034
Epoch 12/15
782/782 - 98s - 125ms/step - accuracy: 0.9944 - loss: 0.0154 - val_accuracy: 0.8122 - val_loss: 1.1609
Epoch 13/15
...
Epoch 14/15
782/782 - 102s - 130ms/step - accuracy: 0.9974 - loss: 0.0098 - val_accuracy: 0.8062 - val_loss: 1.2600
Epoch 15/15
782/782 - 96s - 123ms/step - accuracy: 0.9974 - loss: 0.0084 - val_accuracy: 0.8064 - val_loss: 1.2009
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

```
782/782 - 29s - 37ms/step - accuracy: 0.8064 - loss: 1.2009
Pérdida de la Prueba: 1.2009077072143555
Exactitud de la Prueba (Test accuracy): 0.8063600063323975
```

Entrenamiento y resultados modelo modificado

- Embedding Layer: Se convierten las palabras en vectores de 128 dimensiones usando las 50,000 palabras más frecuentes.
- LSTM Layers:
    - Primera capa LSTM con 128 unidades, configurada con return_sequences=True para devolver la secuencia completa.
    - Segunda capa LSTM con 64 unidades y dropout=0.3.
- Additional Features Input: Dos características adicionales (longitud de la crítica y proporción de palabras positivas/negativas) se procesan por separado y se concatenan con la salida de las capas LSTM.
- Dense Layers: Dos capas densas adicionales con activación relu, con 64 y 32 unidades respectivamente, y dropout=0.5.
- Output Layer: Neurona final con activación sigmoid para la clasificación binaria.
- Optimizer: Adam.
- Loss Function: binary_crossentropy.

```
✓ 100m 3.8s

Epoch 1/15
782/782 - 202s - 259ms/step - accuracy: 0.7449 - loss: 0.5151 - val_accuracy: 0.8049 - val_loss: 0.4288
Epoch 2/15
782/782 - 204s - 261ms/step - accuracy: 0.8656 - loss: 0.3261 - val_accuracy: 0.7731 - val_loss: 0.4906
Epoch 3/15
782/782 - 195s - 249ms/step - accuracy: 0.9154 - loss: 0.2283 - val_accuracy: 0.8565 - val_loss: 0.3430
Epoch 4/15
782/782 - 288s - 369ms/step - accuracy: 0.9578 - loss: 0.1271 - val_accuracy: 0.8475 - val_loss: 0.4049
Epoch 5/15
782/782 - 1109s - 1s/step - accuracy: 0.9762 - loss: 0.0743 - val_accuracy: 0.8354 - val_loss: 0.5464
Epoch 6/15
782/782 - 189s - 241ms/step - accuracy: 0.9832 - loss: 0.0543 - val_accuracy: 0.8440 - val_loss: 0.6007
Epoch 7/15
782/782 - 1366s - 2s/step - accuracy: 0.9919 - loss: 0.0273 - val_accuracy: 0.8447 - val_loss: 0.6779
Epoch 8/15
782/782 - 189s - 242ms/step - accuracy: 0.9927 - loss: 0.0240 - val_accuracy: 0.8400 - val_loss: 0.6954
Epoch 9/15
782/782 - 693s - 886ms/step - accuracy: 0.9950 - loss: 0.0156 - val_accuracy: 0.8429 - val_loss: 0.5532
Epoch 10/15
782/782 - 138s - 177ms/step - accuracy: 0.9952 - loss: 0.0155 - val_accuracy: 0.8329 - val_loss: 0.7074
Epoch 11/15
782/782 - 130s - 167ms/step - accuracy: 0.9960 - loss: 0.0136 - val_accuracy: 0.8397 - val_loss: 0.7408
Epoch 12/15
782/782 - 123s - 158ms/step - accuracy: 0.9967 - loss: 0.0110 - val_accuracy: 0.8430 - val_loss: 0.8283
Epoch 13/15
...
Epoch 14/15
782/782 - 931s - 1s/step - accuracy: 0.9976 - loss: 0.0079 - val_accuracy: 0.8297 - val_loss: 1.0295
Epoch 15/15
782/782 - 125s - 160ms/step - accuracy: 0.9974 - loss: 0.0080 - val_accuracy: 0.8395 - val_loss: 0.8289
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

```
782/782 - 29s - 37ms/step - accuracy: 0.8395 - loss: 0.8289
Pérdida de la prueba: 0.8288770914077759
Exactitud de la prueba: 0.8394799828529358
```

Características Adicionales Seleccionadas y Razón de su Elección

1. **Longitud de la Crítica**: Se incrementó la longitud del tamaño de la reseña. Esto debido a que en algunos casos, las críticas más largas podrían contener más información relevante o matices, lo que podría influir en la polaridad general del sentimiento.
2. **Proporción Positiva/Negativa de Palabras (Pos/Neg Ratio)**: Se contó la cantidad de palabras que coinciden con una lista básica de palabras positivas y negativas, lo que permitió calcular una proporción simple entre palabras positivas y negativas. Esto lo incluimos, ya que el análisis de sentimientos depende de la polaridad de las palabras utilizadas. Por lo tanto, si una crítica contiene más palabras positivas, es más probable que el sentimiento general sea positivo, y viceversa.

Arquitectura del Modelo y Razones Detrás de sus Elecciones

Entrada de Secuencias:

- Embedding Layer: Convierte las palabras en vectores de 128 dimensiones, usando las 50,000 palabras más frecuentes del dataset IMDB. Esto representa la entrada de texto secuencial.

Capas LSTM:

- Primera Capa LSTM:
    - 128 unidades.
    - return_sequences=True
    - Sin dropout directo en esta capa.
- Segunda Capa LSTM:
    - 64 unidades.
    - return_sequences=False
    - dropout=0.3

Entrada de Características Adicionales:

- Dos características:
    - Longitud de la crítica: La longitud total de cada reseña.
    - Proporción positiva/negativa: Diferencia entre la cantidad de palabras positivas y negativas, dividida por el número total de palabras en la crítica.
- Estas características son normalizadas utilizando StandardScaler

Concatenación de Entradas:

- La salida de la segunda capa LSTM se concatena con las características adicionales normalizadas, formando una entrada combinada que lleva tanto información secuencial como estadística sobre la crítica.

Capas Densas:

- Primera Capa Densa:
    - 64 unidades con activación relu
    - Dropout=0.5
- Segunda Capa Densa:
    - 32 unidades con activación relu.

Capa de Salida:

- Neurona con activación sigmoide

Optimización y Función de Pérdida:

- Optimizer: Adam
- Loss Function: binary_crossentropy

Resultados Obtenidos y Comparación

- Resultados del Primer Modelo (Simple):
    - Precisión: 80.64%
    - Pérdida: 1.2009
- Resultados del Segundo Modelo (Modificado):
    - Precisión: 83.95%
    - Pérdida: 0.8289

El segundo modelo presenta una mejora notable en la precisión, alcanzando un 83.95% frente al 80.64% del primero, además de una reducción en la pérdida de 0.8289 frente a 1.2009. Estos resultados sugieren que la adición de características adicionales, como la longitud de la crítica y la proporción de palabras positivas/negativas, junto con una arquitectura más profunda, permitió al modelo generar predicciones más precisas y generalizadas. Esta combinación de información secuencial y características adicionales le permitió capturar patrones que el modelo simple basado solo en LSTM no logró, lo que explica su superioridad tanto en precisión como en pérdida.