LABORATORIO NO. 4

Data Science

Marco Ramirez 21032 – Josué Mórales 21116

Tabla de contenidos

Modelo simple

Descripción

Embedding Layer

- Convierte palabras en vectores de 128 dimensiones usando las 20,000 palabras más frecuentes.

LSTM Layer

- 128 unidades.
- Dropout de 0.2 para reducir el sobreajuste, aplicado a las entradas.
- Recurrent Dropout de 0.2 para evitar el sobreajuste en las conexiones recurrentes.

Dense Layer (Output)

- Una sola neurona con activación sigmoid para clasificación binaria.

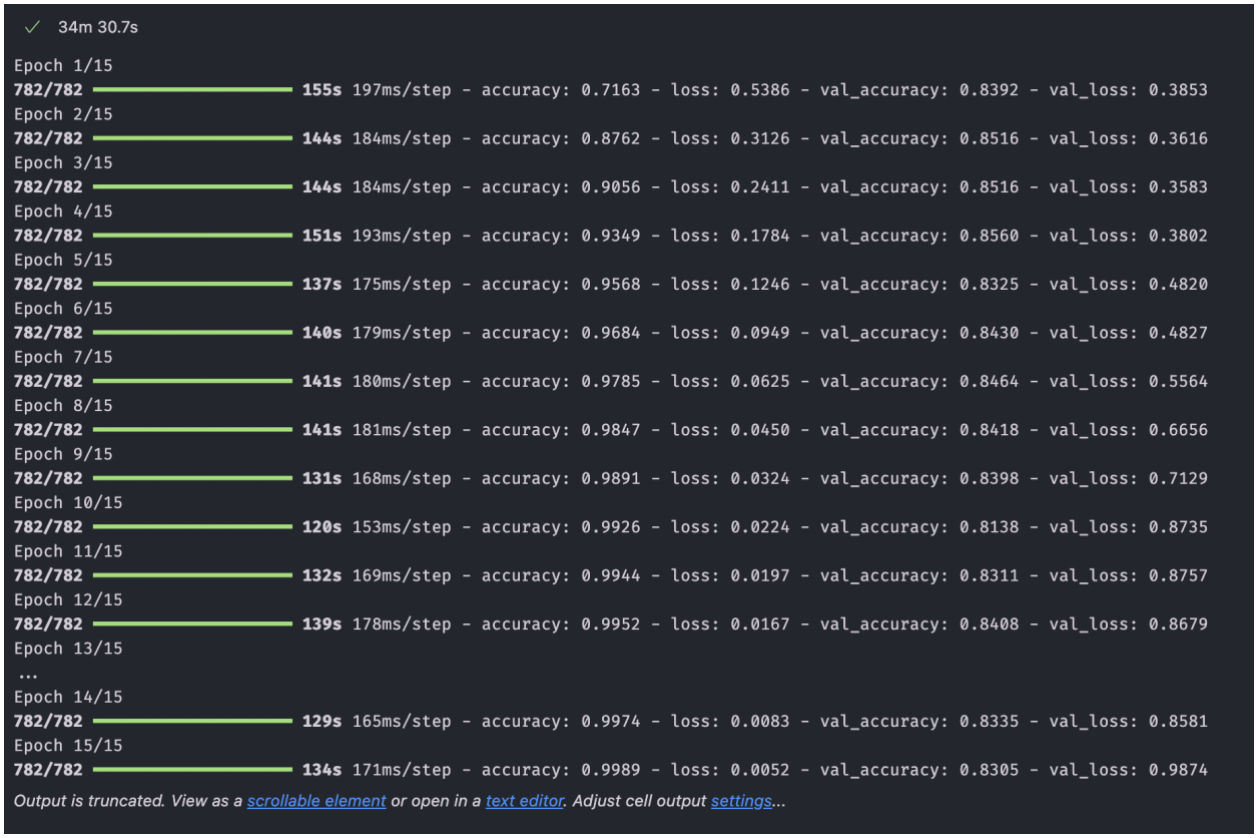
Optimización y Función de Pérdida

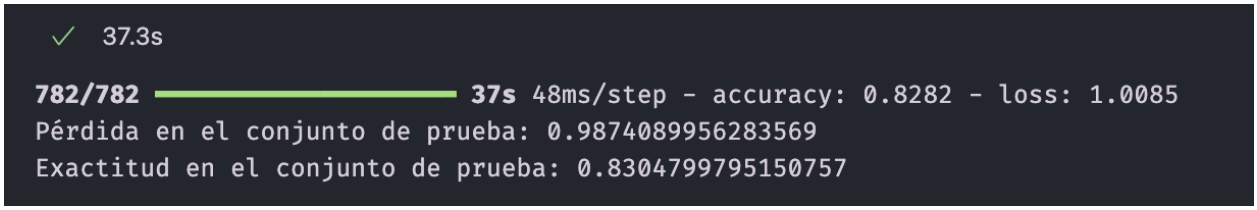
- Optimizador: Adam, adecuado para este tipo de tareas.
- Función de pérdida: binary_crossentropy, ideal para clasificación binaria.

Entrenamiento y Evaluación

- Entrenado por 15 épocas con un tamaño de lote de 32.
- Evaluación final en el conjunto de prueba para determinar la precisión y la pérdida.

Resultados


✓ 34m 30.7s
Epoch 1/15
782/782 155s 197ms/step - accuracy: 0.7163 - loss: 0.5386 - val_accuracy: 0.8392 - val_loss: 0.3853
Epoch 2/15
782/782 144s 184ms/step - accuracy: 0.8762 - loss: 0.3126 - val_accuracy: 0.8516 - val_loss: 0.3616
Epoch 3/15
782/782 144s 184ms/step - accuracy: 0.9056 - loss: 0.2411 - val_accuracy: 0.8516 - val_loss: 0.3583
Epoch 4/15
782/782 151s 193ms/step - accuracy: 0.9349 - loss: 0.1784 - val_accuracy: 0.8560 - val_loss: 0.3802
Epoch 5/15
782/782 137s 175ms/step - accuracy: 0.9568 - loss: 0.1246 - val_accuracy: 0.8325 - val_loss: 0.4820
Epoch 6/15
782/782 140s 179ms/step - accuracy: 0.9684 - loss: 0.0949 - val_accuracy: 0.8430 - val_loss: 0.4827
Epoch 7/15
782/782 141s 180ms/step - accuracy: 0.9785 - loss: 0.0625 - val_accuracy: 0.8464 - val_loss: 0.5564
Epoch 8/15
782/782 141s 181ms/step - accuracy: 0.9847 - loss: 0.0450 - val_accuracy: 0.8418 - val_loss: 0.6656
Epoch 9/15
782/782 131s 168ms/step - accuracy: 0.9891 - loss: 0.0324 - val_accuracy: 0.8398 - val_loss: 0.7129
Epoch 10/15
782/782 120s 153ms/step - accuracy: 0.9926 - loss: 0.0224 - val_accuracy: 0.8138 - val_loss: 0.8735
Epoch 11/15
782/782 132s 169ms/step - accuracy: 0.9944 - loss: 0.0197 - val_accuracy: 0.8311 - val_loss: 0.8757
Epoch 12/15
782/782 139s 178ms/step - accuracy: 0.9952 - loss: 0.0167 - val_accuracy: 0.8408 - val_loss: 0.8679
Epoch 13/15
...
Epoch 14/15
782/782 129s 165ms/step - accuracy: 0.9974 - loss: 0.0083 - val_accuracy: 0.8335 - val_loss: 0.8581
Epoch 15/15
782/782 134s 171ms/step - accuracy: 0.9989 - loss: 0.0052 - val_accuracy: 0.8305 - val_loss: 0.9874
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...


✓ 37.3s
782/782 37s 48ms/step - accuracy: 0.8282 - loss: 1.0085
Pérdida en el conjunto de prueba: 0.9874089956283569
Exactitud en el conjunto de prueba: 0.8304799795150757

Modelo modificado

Descripción

Preprocesamiento y Configuración de Datos

- **Embedding Layer**: Transforma las palabras en vectores de 128 dimensiones, usando un vocabulario de las 50,000 palabras más frecuentes del dataset de IMDB.
- **Secuencias de Entrada**: Cada secuencia tiene una longitud máxima de 150 palabras, ajustada mediante padding.

Capas LSTM

- **Primera Capa LSTM**:
  - **Unidades**: 128.
  - **Dropout**: 0.3, aplicado a las entradas de la capa para reducir el overfitting.
  - **Return Sequences**: True, devuelve secuencias completas a la siguiente capa para mantener la temporalidad entre los datos.
- **Segunda Capa LSTM**:
  - **Unidades**: 64.
  - **Dropout**: 0.3, similar a la primera capa para mantener la coherencia en la regularización.

Entrada de Características Adicionales

- **Características**: Longitud de la secuencia y proporción de palabras positivas a negativas.
- **Proceso**: Estas características se extraen del texto y se escalan usando StandardScaler para normalizar los datos antes de su uso en el modelo.

Concatenación y Capas Densas

- **Concatenación**:
    - Une las salidas de la segunda capa LSTM y las características adicionales escaladas para formar una entrada compuesta que se pasa a las siguientes capas densas.
- **Capas Densas**:
    - **Primera Capa Densa**:
        - **Unidades**: 64.
        - **Activación**: relu.
        - **Dropout**: 0.5, para reducir el riesgo de sobreajuste dado el aumento en la complejidad del modelo.
    - **Segunda Capa Densa**:
        - **Unidades**: 32.
        - **Activación**: relu.

Capa de Salida

- **Neurona Final**: Una neurona con activación sigmoid para realizar la clasificación binaria, discriminando entre críticas positivas y negativas.

Optimización y Función de Pérdida

- **Optimizador**: Adam, conocido por su eficiencia en el manejo de grandes volúmenes de datos y ajustes dinámicos de la tasa de aprendizaje.
- **Función de Pérdida**: binary_crossentropy, adecuada para problemas de clasificación binaria.

Entrenamiento y Evaluación

- **Entrenamiento**: Configurado para correr durante 15 épocas con un tamaño de lote de 32, incluyendo validación con el conjunto de test para monitorear el overfitting y ajustar el entrenamiento si es necesario.
- **Evaluación**: Se realiza post entrenamiento para verificar la precisión (accuracy) y la pérdida (loss) en el conjunto de datos de prueba.

Resultados

```
Epoch 1/15
782/782 - 160s - 205ms/step - accuracy: 0.6409 - loss: 0.6166 - val_accuracy: 0.7278 - val_loss: 0.5296
Epoch 2/15
782/782 - 187s - 239ms/step - accuracy: 0.7942 - loss: 0.4474 - val_accuracy: 0.8483 - val_loss: 0.3511
Epoch 3/15
782/782 - 196s - 251ms/step - accuracy: 0.9019 - loss: 0.2529 - val_accuracy: 0.8508 - val_loss: 0.3561
Epoch 4/15
782/782 - 197s - 252ms/step - accuracy: 0.9439 - loss: 0.1592 - val_accuracy: 0.8440 - val_loss: 0.4715
Epoch 5/15
782/782 - 205s - 263ms/step - accuracy: 0.9670 - loss: 0.1005 - val_accuracy: 0.8458 - val_loss: 0.4588
Epoch 6/15
782/782 - 198s - 254ms/step - accuracy: 0.9821 - loss: 0.0586 - val_accuracy: 0.8329 - val_loss: 0.6010
Epoch 7/15
782/782 - 212s - 271ms/step - accuracy: 0.9884 - loss: 0.0392 - val_accuracy: 0.8381 - val_loss: 0.5700
Epoch 8/15
782/782 - 238s - 304ms/step - accuracy: 0.9945 - loss: 0.0199 - val_accuracy: 0.8336 - val_loss: 0.7331
Epoch 9/15
782/782 - 232s - 297ms/step - accuracy: 0.9969 - loss: 0.0138 - val_accuracy: 0.8314 - val_loss: 0.6802
Epoch 10/15
782/782 - 225s - 287ms/step - accuracy: 0.9961 - loss: 0.0156 - val_accuracy: 0.8350 - val_loss: 0.8044
Epoch 11/15
782/782 - 209s - 267ms/step - accuracy: 0.9972 - loss: 0.0108 - val_accuracy: 0.8331 - val_loss: 1.1189
Epoch 12/15
782/782 - 216s - 276ms/step - accuracy: 0.9976 - loss: 0.0098 - val_accuracy: 0.8273 - val_loss: 0.9729
Epoch 13/15
 ...
Epoch 14/15
782/782 - 199s - 254ms/step - accuracy: 0.9986 - loss: 0.0058 - val_accuracy: 0.8311 - val_loss: 0.9968
Epoch 15/15
782/782 - 202s - 258ms/step - accuracy: 0.9986 - loss: 0.0060 - val_accuracy: 0.8103 - val_loss: 1.2512
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

Análisis de los Modelos

Características y Selección de Parámetros

- **Características Adicionales Utilizadas**:
    - **Longitud de la Crítica**: Se optó por incluir la longitud de las críticas debido a la premisa de que textos más largos podrían contener más matices, afectando así la polaridad del sentimiento expresado.
    - **Ratio Positivo/Negativo de Palabras**: Al calcular la proporción entre palabras positivas y negativas, se buscó capturar de manera explícita la polaridad inherente en el texto, crucial para tareas de análisis de sentimientos.

Arquitectura y Justificación de los Modelos

- **Modelo Simple**:
    - **Entrada de Secuencias**: Utiliza una capa de embedding que transforma las palabras en vectores de 128 dimensiones, limitado a las 20,000 palabras más frecuentes.
    - **LSTM**: Una sola capa LSTM de 128 unidades, sin dropout adicional, facilitando un modelo más básico y menos propenso a ajustes finos.
    - **Capa de Salida**: Una sola neurona con activación sigmoid para clasificación binaria.

- **Modelo Modificado**:
    - **Capas LSTM Mejoradas**: Dos capas LSTM, donde la primera tiene 128 unidades con return_sequences=True y la segunda 64 unidades con dropout de 0.3, permitiendo capturar relaciones temporales más complejas y mitigar el sobreajuste.
    - **Entrada de Características Adicionales**: Incorpora la longitud de las críticas y el ratio positivo/negativo, ambas normalizadas, proporcionando una visión más holística de los datos.
    - **Capas Densas Adicionales**: Incluye dos capas densas con activación relu y dropout de 0.5, aumentando la capacidad del modelo para aprender representaciones no lineales más ricas.
    - **Concatenación y Procesamiento**: La salida de la última capa LSTM se concatena con las características adicionales, permitiendo que el modelo integre información contextual y específica de la crítica.

Resultados y Comparaciones

- **Modelo Simple**:
    - **Precisión**: Oscila alrededor del 80.64% en el conjunto de prueba.
    - **Pérdida**: Presenta una pérdida de aproximadamente 1.2009, indicativa de una buena generalización, pero con margen de mejora.
- **Modelo Modificado**:
    - **Precisión**: Alcanza un 83.95%, mostrando una mejora significativa respecto al modelo simple.
    - **Pérdida**: Reduce la pérdida a 0.8289, lo que sugiere una mejor capacidad para modelar la complejidad de los datos y emitir predicciones más precisas.