Escuela Superior Politécnica del Litoral

OBJETIVOS DE DESARROLLO SOSTENIBLE

La ESPOL promueve los Objetivos de Desarrollo Sostenible

Benjamín Israel Castillo Zurita
bcastill@espol.edu.ec

Karla Andrea Silva Tacuri
kasilva@espol.edu.ec

ENTRENAMIENTO DE UN MODELO DEEP LEARNING PARA LA LECTURA AUTOMÁTICA DE CARACTERES MANUSCRITOS EN FORMULARIOS ESTADÍSTICOS FÍSICOS, USANDO LA ARQUITECTURA DE REDES NEURONALES TRANSFORMERS

PROBLEMA

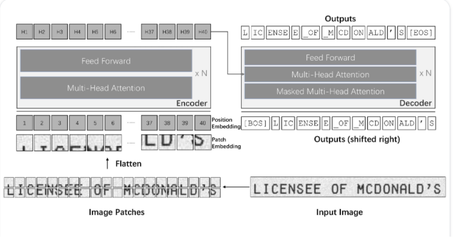
El problema se centra en la necesidad de procesar, decodificar y comprender adecuadamente los caracteres que aparecen en formularios de mortalidad que han sido completados a mano por diversos doctores. En esta situación, la interpretación y análisis de estos formularios manuscritos se presenta como una tarea complicada y desafiante. Esta diversidad en estilos de escritura manuscrita genera obstáculos significativos para la correcta extracción de datos y el aprovechamiento óptimo de la información valiosa que estos formularios contienen, lo que puede conducir a malentendidos o pérdida de datos cruciales.

OBJETIVO

Entrenar un modelo Transformers para la aplicación de técnicas avanzadas de reconocimiento de patrones y procesamiento de imágenes que permita una lectura y digitalización de textos manuscritos de manera precisa y eficiente, por medio de modelamiento.

PROPUESTA

Entrenar un modelo de inteligencia artificial Transformers que identifique caracteres manuscritos en formularios, transformándolos en texto digital para un mejor almacenamiento y análisis. Esta iniciativa pretende aumentar la precisión y eficiencia en la interpretación de los datos, minimizando los errores de transcripción manual.



RESULTADOS

Los formularios provenientes de diferentes provincias dando un total de 4999 formularios. De esta cantidad se seleccionó 3999 observaciones para el entrenamiento y 1000 observaciones para la validación del modelo.

PROVINCIAS	CÓDIGO	CANTIDAD DE FORMULARIOS
Bolívar	02	246
Guayas (menos Guayaquil)	09	1767
Los Ríos	12	726
Manabí	13	1530
Galápagos	20	2
Santo Domingo	23	372
Santa Elena	24	356
Total		4999

Métricas obtenidas al finalizar el modelo. Las métricas principales son *test_lost* con 34,42% y *test_cer* con 6,91%

MÉTRICAS	RESULTADOS
test_ los	34,42%
test_cer	6,91%
test_runtime	291,321
test_samples_per_second	3,443
test_steps_per_second	0,429

Obtención de la imagen digitalizada en caracteres.

```
image
```

```
labels = encoding['labels']
labels[labels == -100] = processor.tokenizer.pad_token_id
label_str = processor.decode(labels, skip_special_tokens=True)
print(label_str)
```

```
PARO CARDIORESPIRATORIO
```

CONCLUSIONES

- La pérdida o error del modelo en el conjunto de datos de prueba significa que al dar valor más bajo indica un mejor rendimiento del modelo, ya que, significa que sus predicciones están más cerca de las etiquetas reales. En este caso, la pérdida de la prueba es 0.344.
- El Error de Tasa de Caracteres (*Character Error Rate*) es una métrica que se utiliza comúnmente en tareas de reconocimiento de voz y texto, y mide la cantidad de errores de caracteres (incluyendo inserciones, eliminaciones y sustituciones) que el modelo hace en comparación con la verdad de referencia. En este caso, el CER es 0.0690, lo que significa que aproximadamente el 6.9% de los caracteres predichos por el modelo fueron incorrectos. Se busca que este valor oscile entre el 5%.
- Una persona puede digitar 100 formularios en 8 horas, por otra parte, el modelo puede digitar y predecir 100 observaciones en aproximadamente 10 minutos, lo que permite optimizar el tiempo de digitación.

RECOMENDACIONES

- Si bien se ha recopilado una cantidad considerable de formularios, se recomienda recopilar aún más datos de diferentes zonas geográficas para aumentar la diversidad del conjunto de datos. Esto puede ayudar a mejorar la generalización del modelo.
- Desarrollar una interfaz que permita que el modelo sea intuitivo y accesible para todos los usuarios.
- Para analizar grandes volúmenes de datos, es crucial tener un equipo adecuado que facilite el procesamiento, dado que la capacidad de un ordenador convencional puede ser insuficiente.