

# Segmentación de artículos periodísticos con transferencia de aprendizaje en redes neuronales convolucionales

Adolfo Simaz Bunzel<sup>1,2</sup>, Agostina Filócomo<sup>1,3</sup>, Ezequiel A. Pássaro<sup>1</sup>

<sup>1</sup>Facultad de Ciencias Astronómicas y Geofísicas, UNLP, Paseo del Bosque S/N, La Plata, Argentina

<sup>2</sup>Instituto Argentino de Radioastronomía (CCT La Plata, CONICET; CICPBA; UNLP), C.C.5, (1894) Villa Elisa, Buenos Aires, Argentina

<sup>3</sup>Departamento de Investigación en Ciencias Exactas e Ingeniería, UNRN - Sede Atlántica, Don Bosco y Leloir S/N, Viedma, Argentina

29 de mayo de 2024

## 1 INTRODUCCIÓN

La extracción de información de documentos basados en imágenes es una herramienta muy poderosa para la digitalización de los mismos. Para ello, se utilizan técnicas de *machine learning* específicas como lo es el uso de redes neuronales convolucionales. En el presente informe describimos los pasos seguidos para la extracción de información de diversos recortes de periódicos con un segmentador de instancias *Mask R-CNN*<sup>1</sup>.

Para facilitar la reproducción de los resultados aquí mostrados todos los programas desarrollados tienen asociado un enlace a un repositorio con versionado de código. Se ha hecho especial hincapié en las licencias de código abierto de las herramientas utilizadas.

El informe se organiza de la siguiente manera: en la Sección 2 se presenta una descripción de los datos utilizados. En la Sección 3 se detallan los métodos aplicados sobre el conjunto de datos antes mencionado para su segmentación. Finalmente, la Sección 4 contiene un análisis de los resultados encontrados, junto con una breve discusión sobre los mismos.

## 2 DATOS

El primer paso de la tarea fue la producción de una muestra de datos para ser utilizado en el entrenamiento de las redes neuronales. Para ello, creamos un conjunto de datos en formato *Common Objects in Context* (COCO)<sup>2</sup> a partir de los 501 archivos en formato JSON producidos por los participantes durante la fase de etiquetado comunitario y recopilados por la organización de la competencia. Posteriormente, realizamos un filtrado de la muestra para garantizar la consistencia de la información, dando como resultado un total de 297 imágenes con sus respectivas anotaciones.

En el proceso notamos que, pese a haber entregado en tiempo y forma las 60 anotaciones que le fueron asignadas a este grupo, las mismas no se encontraban presentes en los datos recopilados. Por tal motivo, añadimos al conjunto de datos otras 200 imágenes con sus respectivas anotaciones (las que nos fueron asignadas originalmente, más 140), sumando un total de 497 imágenes.

Así, de esta etapa de procesamiento obtenemos un conjunto de datos a partir del cual entrenamos las redes neuronales, que describimos en la Sección 3. Todas las herramientas para reproducir el conjunto de datos mencionado se encuentran en el repositorio de entrenamiento: <https://github.com/epassaro/diaxi-training-tools>.

## 3 MÉTODOS

El objetivo del trabajo consiste en la segmentación automática de artículos periodísticos para su posterior procesamiento con un software de reconocimiento óptico de caracteres (OCR, por sus siglas en inglés), *tesseract*<sup>3</sup>. Para ello, decidimos hacer *transfer learning* sobre un modelo de segmentación basado en el análisis de documentos, *PrimaLayout*<sup>4</sup>. En lo que a su entrenamiento respecta, utilizamos una librería de algoritmos de detección y segmentación, *detectron2*<sup>5</sup>, desarrollada por *Facebook Research*. Cabe mencionar que el conjunto de datos utilizado en la fase de entrenamiento tuvo que ser modificado para incluir las máscaras de segmentación que no fueron provistas inicialmente, y luego separarlo en subconjuntos de *entrenamiento* y *validación*.

Para una explicación detallada del entrenamiento del modelo, referirse al repositorio de entrenamiento: <https://github.com/epassaro/diaxi-training-tools>.

De esta fase resulta un análisis asociado a la precisión en la detección de las diferentes secciones de una nota periodística sobre el conjunto de *validación*, basado en las regiones de la nota que las redes neuronales detectan, denominadas *bounding boxes*. Los resultados de esta fase se encuentran descriptos con mayor detalle en la Sección 4.

## 4 RESULTADOS

Debido al rol central que tiene el modelo para realizar la segmentación automática, dividimos la presentación de los resultados asociados al entrenamiento del modelo, y su posterior aplicación en el esquema provisto por la organización.

### 4.1 Fase de entrenamiento

Aquí presentamos un breve análisis asociado a la calidad de las predicciones hechas por el modelo previamente explicado, respecto a las *bounding boxes* que delimitan las diferentes secciones a obtener de una nota periodística. Para ello, utilizamos una métrica ampliamente utilizada en el campo de la segmentación en redes neuronales convolucionales, denominada intersección sobre la unión (IoU, por sus siglas en inglés), que cuantifica la similitud entre dos *bounding boxes*: la predicha por el modelo entrenado y la provista como valor verdadero. De esta manera, cuanto más se aproxime a 1 esta cantidad,

<sup>1</sup> <https://arxiv.org/abs/1703.06870>

<sup>2</sup> <https://cocodataset.org/>

<sup>3</sup> <https://tesseract-ocr.github.io/>

<sup>4</sup> <https://layout-parser.readthedocs.io/en/latest/notes/modelzoo.html>

<sup>5</sup> <https://detectron2.readthedocs.io/en/latest/>

