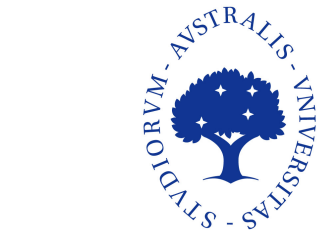

UNIVERSIDAD
AUSTRAL

Maestría en Explotación de Datos y Gestión del Conocimiento

Proyecto de Tesis

Franco Lianza

Julio 2022

Tema

Título del trabajo

Clasificación de los digitos escritos en los telegramas de las elecciones legislativas en Santa Fe mediante técnicas de adaptación de dominio.

Resumen

Resumen del área sobre la que se realizará el trabajo.

La rama de *Computer Vision* se encarga desarrollar herramientas para reconocer patrones complejos en imágenes en múltiples dominios. Se ha extendido exponencialmente a lo largo del tiempo, llegando a un punto en el cual se pueden detectar todo tipo de objetos con una precisión óptima [1].

Las técnicas desarrolladas en el área precisan de un gran volumen de datos para su entrenamiento. Esto implica que es de suma importancia de tener disponibles las *labels* (etiquetas) de los datos que se van a utilizar para entrenar los modelos. El etiquetado de los datos es una tarea costosa, ineficiente y hasta a veces resulta inviable de realizar [2].

Aún teniendo los *labels*, puede ocurrir que el *dataset* (conjunto de datos) donde se va a utilizar el modelo resulte diferente al que se utilizó para entrenarlo. Por mencionar, un modelo de detección de rostros entrenado en una etnia demográfica particular funcionará de manera errónea si se lo aplica a otra. Este fenónmeno se conoce como *dataset bias* o *dataset shift* (sesgo en los datos). Dicho de otra manera, un modelo entrenado en un *dataset* puede no generalizar correctamente debido al *dataset bias*. Algunos autores afirman que el sesgo es un problema que no se puede evitar al momento de crear un *dataset* [3].

La detección de dígitos en los telegramas de elecciones en Argentina podría llevarse a cabo mediante un modelo entrenado en *datasets* de dígitos públicos como el *MNIST* [4]. Como no existe una única forma de escribir, el modelo estará sesgado a reconocer dígitos escritos de forma similar a los que se encontraban en el *dataset* de entrenamiento. No será capaz de generalizar lo aprendido en un dominio distinto.

En trabajos anteriores, se aplican distorsiones al conjunto de entrenamiento para aumentar la cantidad de datos de entrenamiento y de esta forma el modelo pueda generalizar y aplicarse a los telegramas de elecciones de la Ciudad de Buenos Aires [5]. En el presente trabajo se utilizarán técnicas referidas al *transfer learning*, específicamente de *domain adaptation* para resolver el problema.

El *transfer learning* (transferencia de aprendizaje) es un área del *machine learning* que se encarga de almacenar el conocimiento ganado en un problema y aplicarlo a otro. *Domain adaptation* (adaptación de dominio) consiste en la habilidad de aplicar un modelo entrenado en uno o mas dominios de origen (*source domain*) en un uno distinto pero relacionado (*target domain*). La figura 1 muestra un ejemplo dos *datasets* de dígitos pero con dominios diferentes. Un modelo entrenado con *MNIST* no generalizará al conjunto *SVHN* por más que ambos sean dígitos.

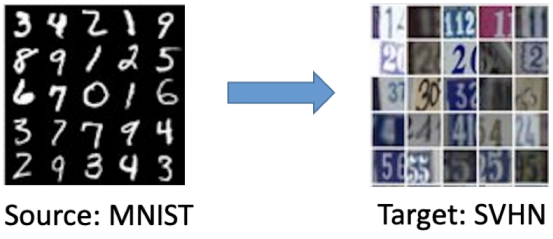

Figura 1: Ejemplo dominios diferentes: MNIST y SVHN

Director o Tutor

El nombre del director o tutor de la tesis o trabajo, si éste ya hubiese aceptado la tarea.

Dr. Leandro Bugnon

Motivación e importancia del campo

Explicar la o las motivaciones que llevan a realizar el trabajo planteado y su importancia.

La principal motivación es obtener un modelo que, basado en un *dataset* similiar ya etiquetado, pueda interpretar los dígitos de los telegramas de las elecciones de Santa Fe sin necesitar el costoso proceso de etiquetar los datos. De esta manera, el modelo podría utilizarse en futuras elecciones a fin de automatizar la digitalización de los mismos reduciendo considerablemente los costos operacionales y reduciendo los posibles errores humanos.

Problemas no resueltos

Problemas no resueltos detectados en el área y que el trabajo a realizar pretende resolver.

La digitalización de los telegramas de las elecciones sigue siendo de forma manual, lo que genera lentitud y desconfianza en el proceso. Detectar los votos de manera automática y reducir la intervención humana, aumentará la eficiencia de las elecciones y la confianza en ellas.

Por otra parte, cada una de las técnicas existentes de *domain adaptation* pretende aumentar el poder de generalización del modelo de una forma diferente. Debido a la variedad ténicas que existen, no es claro cúal conviene utilizar en la digitalización de los telegramas.

Objetivo del trabajo

Explicar claramente el objetivo del trabajo, especificando su alcance y limitaciones.

El objetivo del trabajo consiste en desarrollar un método para realizar adaptación de dominio de un modelo aplicado a la digitalización de los telegramas de las elecciones legislativas de Santa Fe sin que necesite el etiquetado de los mismos.

Requerimientos y desafíos

Requerimientos y desafíos que plantea el trabajo a realizar.

Debido a que se posee tanto los datos como el poder de cómputo, el proyecto presenta como principal desafío analizar e implementar distintas técnicas de *domain adaptation* a un modelo de reconocimiento de dígitos. Se deberán estudiar diferentes conceptos de los cuales dependen las mismas, como las *Generative Adversarial Networks (GANs)* [6, 7].

Metodología

Metodología a emplear para el desarrollo del trabajo.

Se hará revisión del estado del arte para después continuar con el preprocesamiento de los datos. Los telegramas serán descargados desde la página oficial del estado argentino. Luego se extraerán los dígitos de los votos de cada telegrama utilizando la librería *OpenCV* [8]. Una vez obtenido los dígitos, se realizarán ciclos de selección de técnica de *domain adaptation*, entrenamiento de distintas redes convolucionales (LeNet [4], ResNet [9], etc), implementación de la adaptación y evaluación. Finalmente, se seleccionará el mejor modelo adaptado al problema.

Plan de Trabajo

Especificar las distintas tareas a realizar con los tiempos que se estime que deberían insumir. Indicar las fechas estimadas de inicio y finalización del trabajo.

El trabajo se realizará a partir de las siguientes etapas:

- Estudio del estado del arte: estudiar las diferentes técnicas de *domain adaptation*.
- Extracción datos: recolectar los telegramas y extraer los dígitos de los votos.
- Limpieza de datos: detectar votos incorrectos (por ejemplo letras) y tratarlos para reducir lo máximo posible el ruido en los datos.
- Experimentación: realizar ciclos de entrenamiento de modelos y adaptarlos mediante alguna de las técnicas disponibles.
- Elaboración de reportes: generar reportes que sinteticen los experimentos realizados para compararlos.

- Redacción de tesis.

Se estima que el proyecto comienza en Agosto del 2022 y finaliza en Diciembre del 2022. La figura 2 presenta el cronograma propuesto.

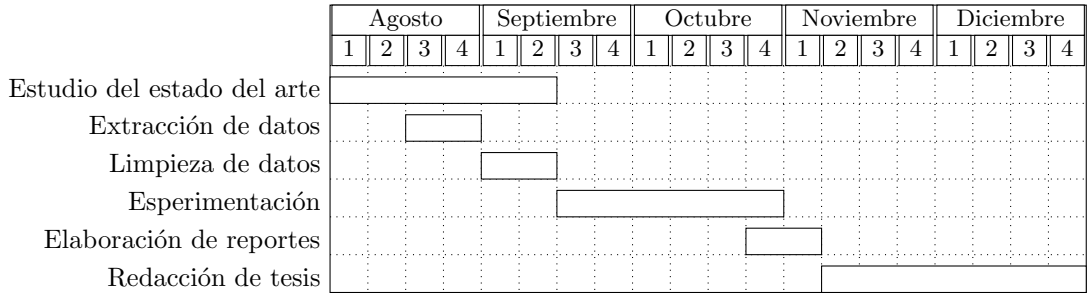

Figura 2: Distribución de las tareas del proyecto.

Bibliografía

[1] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[2] Pedro Miguel Lima de Sousa Reis. Data labeling tools for computer vision: a review. 2022.

[3] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[4] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[5] Walter Marcelo Lamagna. *Lectura artificial de números manuscritos en datos abiertos de elecciones legislativas en la Ciudad de Buenos Aires*. PhD thesis, Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales, 2016.

[6] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[8] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.