UNIVERSIDAD AUSTRAL

TESIS DE MAESTRÍA

Taming dataset bias via Domain Adaptation

Autor:
Franco LIANZA

Supervisor:
Dr. Leandro BUGNON

12 de agosto de 2022

UNIVERSIDAD AUSTRAL

Resumen

Facultad de Ingeniería

Magister en Explotación de Datos y Gestión del Conocimiento

Taming dataset bias via Domain Adaptation

by Franco LIANZA

The Thesis Abstract is written here (and usually kept to just this page). The page is kept centered vertically so can expand into the blank space above the title too...

Acknowledgements

The acknowledgments and the people to thank go here, don't forget to include your project advisor...

Índice general

Índice de figuras

Índice de cuadros

For/Dedicated to/To my…

Capítulo 1

Introducción

1.1. Elecciones en Argentina

En Argentina se celebran elecciones cada 2 años a excepción de las presidenciales que se realizan cada 4 años. En Argentina se realizan tres tipos de elecciones:

- Elecciones nacionales, para elegir a las autoridades federales del país: el Poder Ejecutivo, constituido por el Presidente y el vicepresidente y el Congreso Nacional, formado por Senadores y Diputados.
- Elecciones provinciales y de la Ciudad de Buenos Aires o locales, para elegir a las autoridades de cada provincia: los poderes ejecutivos de las provincias y sus legislaturas.
- Elecciones municipales, regidas por las leyes y procedimientos de cada provincia.

Si bien emitir el sufragio es diferente en cada una de ellas, generalmente consta de ingresar a un cuarto oscuro, elegir el candidato que se desea y depositar el voto en una urna. Al finalizar la jornada, las autoridades de mesas recuentan los votos y llenan una planilla a mano alzada donde se resume la cantidad de votos obtenidos por cada candidato o partido político. Dicha planilla es escaneada y enviada a traves de un telegrama correo argentino al centro de cómputo para su procesamiento. Una vez allí, se contabilizan en un sistema informático una por una. Por la metodología de contabilizacíon, idealmente lo escrito a mano en el telegrama y lo computado en el sistema es lo mismo. Sin embargo, como esta tarea es realizada por personas, es plausible pensar que pueden haber errores en dicho proceso.

(TODO: ver de agregar un diagrama aca del proceso) (TODO: ver de agregar una imagen de un telegrama en el apendice o anexo)

A su vez, durante la jornada electoral existe una ansiedad generalizada para ir sabiendo los resultados parciales y finales de la misma, por lo que se debe contratar a una gran cantidad personas destinadas al centro de cómputo. En las elecciones legislativas del 2021 se gastaron unos $17.000 millones de pesos de los cuales $4.000 millones de pesos fueron destinados a sueldos para el personal[1]. Mejorar el proceso manual de contabilización de los telegramas supondrá un ahorro considerable en el presupuesto de las elecciones,

[1] Fuente: El cronista

agilizará la obtención de los resultados y aportará transparencia al proceso en general.

(TODO: buscar alguna noticia de ver si algun pais ya digitaliza automaticamente o khe)

1.2. Clasificación de dígitos (TODO: ???)

La clasificación de dígitos escritos a mano lleva resuelto hace un tiempo con una performance óptima. LeCun et al., 1998 propone una arquitectura de red neuronal con múltiples capas y clasifica correctamente el dataset *MNIST* con ella. Se podría proponer un modelo similar para esto.

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* LeCun et al., 1998. 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 Lamagna, 2016. En el presente trabajo se utilizarán técnicas referidas al *transfer learning*, específicamente de *domain adaptation* para resolver el problema.

1.3. Deep Learning

Cuando se habla de *Deep learning*, se hace referencia a una serie de algoritmos de *machine learning* que son capaces de utilizar múltiples capas de procesamiento de forma que puedan aprender representaciones de los datos con diferentes niveles de abstracción (LeCun, Bengio e Hinton, 2015). Estos algoritmos, denominados redes neuronales profundas (o DNNs por sus siglas en inglés), poseen la capacidad de encontrar variables que expliquen la naturaleza del comportamiento de los datos.

Los modelos obtenidos a partir del *deep learning* han demostrado tener gran capacidad de aprendizaje para todo tipo de problemas, como ser *computer vision* (Szeliski, 2010; Redmon et al., 2016), procesamiento del lenguaje natural (Devlin et al., 2018), reconocimiento del habla (Hannun et al., 2014), juegos (Silver et al., 2016), generación de imágenes a partir de descripciones (Ramesh et al., 2022), entre otros.

Aunque la utilidad de estos modelos se encuentra demostrada y día a día so utilizados en diferentes ámbitos de la vida, presentan un gran problema: la enorme cantidad de datos que requieren para su entrenamiento. La mayoría de los modelos que mejores métricas de performance presentan, necesitan millones de datos en sus datasets de entrenamiento. Esto implica que, para que los mismos sean de utilidad, es de suma importancia los procesos de recolección y etiquetado de los datos. La eficacia de los modelos queda

altamente relacionada con la calidad de los datos que se posean o se logren conseguir. Particularme, el etiquetado de los datos es una tarea costosa, ineficiente y hasta a veces resulta inviable de realizar (Reis, 2022).

Una posible solución a este problema consiste en emular la capacidad que tienen los humanos de adquirir conocimiento relevante en un área y aplicarlo en otra similar (Thrun y Pratt, 1998). Es decir, poder *transferir* lo aprendido. En el caso del *deep learning*, lo que se busca es que la red aprenda representaciones lo suficientemente generales para que despues sean utilizados en el entrenamiento de una tarea similar. Esto busca acortar los tiempos de entrenamiento, mejorar las predicciones y hacer los modelos más robustos.

Capítulo 2

Estado del Arte

2.1. Reconocimiento de dígitos

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

2.2. Redes Neuronales

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam ultricies lacinia euismod. Nam tempus risus in dolor rhoncus in interdum enim tincidunt. Donec vel nunc neque. In condimentum ullamcorper quam non consequat. Fusce sagittis tempor feugiat. Fusce magna erat, molestie eu convallis ut, tempus sed arcu. Quisque molestie, ante a tincidunt ullamcorper, sapien enim dignissim lacus, in semper nibh erat lobortis purus. Integer dapibus ligula ac risus convallis pellentesque.

2.3. Domain Adaptation

While pre-training on large-scale datasets can gain transferable knowledge in deep models, performing task adaptation with the target data is still necessary for most applications, as the target task is usually different from the pre-training task. When the labeled data for the target task is not enough, domain adaptation from a related source domain with labeled data to boost the performance on the target domain is also necessary in many applications.

Apéndice A

Frequently Asked Questions

A.1. How do I change the colors of links?

The color of links can be changed to your liking using:
`\hypersetup{urlcolor=red}`, or
`\hypersetup{citecolor=green}`, or
`\hypersetup{allcolor=blue}`.
If you want to completely hide the links, you can use:
`\hypersetup{allcolors=.}`, or even better:
`\hypersetup{hidelinks}`.
If you want to have obvious links in the PDF but not the printed text, use:
`\hypersetup{colorlinks=false}`.

Bibliografía

Devlin, Jacob et al. (2018). «Bert: Pre-training of deep bidirectional transformers for language understanding». En: *arXiv preprint arXiv:1810.04805*.

Hannun, Awni et al. (2014). «Deep speech: Scaling up end-to-end speech recognition». En: *arXiv preprint arXiv:1412.5567*.

Lamagna, Walter Marcelo (2016). «Lectura artificial de números manuscritos en datos abiertos de elecciones legislativas en la Ciudad de Buenos Aires». Tesis doct. Universidad de Buenos Aires. Facultad de Ciencias Exactas y Naturales.

LeCun, Yann, Yoshua Bengio y Geoffrey Hinton (2015). «Deep learning». En: *nature* 521.7553, págs. 436-444.

LeCun, Yann et al. (1998). «Gradient-based learning applied to document recognition». En: *Proceedings of the IEEE* 86.11, págs. 2278-2324.

Ramesh, Aditya et al. (2022). «Hierarchical text-conditional image generation with clip latents». En: *arXiv preprint arXiv:2204.06125*.

Redmon, Joseph et al. (2016). «You only look once: Unified, real-time object detection». En: *Proceedings of the IEEE conference on computer vision and pattern recognition*, págs. 779-788.

Reis, Pedro Miguel Lima de Sousa (2022). «Data Labeling tools for Computer Vision: a Review». En.

Silver, David et al. (2016). «Mastering the game of Go with deep neural networks and tree search». En: *nature* 529.7587, págs. 484-489.

Szeliski, Richard (2010). *Computer vision: algorithms and applications*. Springer Science & Business Media.

Thrun, Sebastian y Lorien Pratt (1998). *Learning to learn*.