Aprendizaje automático y visión por computadora

L. Lanzarini[1], C. Estrebou[1], F. Ronchetti[1,3], F. Quiroga[1,4], G. Camele[1,5], G. Rios[1,5], U. Cornejo Fandos[1,5], B. Rey[1,5], A. Rosete[6]

[1] Instituto de Investigación en Informática LIDI, Facultad de Informática, UNLP, La Plata, Argentina.*
[2] Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina
[3] Becario postdoctoral UNLP [4] Becario postgrado UNLP [5] Pasante LIDI
[6] Universidad Tecnológica de La Habana "José Antonio Echeverría" (CUJAE), La Habana, Cuba

*** Centro asociado de la Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CIC)**

laural@lidi.info.unlp.edu.ar

CONTEXTO

Esta presentación corresponde a las tareas de investigación que se llevan a cabo en el III LIDI en el marco del proyecto "Sistemas inteligentes. Aplicaciones en reconocimiento de patrones, minería de datos y big data" perteneciente al Programa de Incentivos (2018-2021) y del proyecto PITAP-BA "Computación de Alto Desempeño, Minería de Datos y Aplicaciones de Interés Social en la Provincia de Bs.As." evaluado y subsidiado por la Comisión de Investigaciones Científicas de la Provincia de Bs.As. (2017-2019).

RESUMEN

Esta línea de investigación se centra en el estudio y desarrollo de Sistemas Inteligentes para la resolución de problemas de reconocimiento de patrones en imágenes y video, utilizando técnicas de Aprendizaje Automático clásicas, junto con las nuevas Redes Neuronales Convolucionales y Aprendizaje profundo. El trabajo presentado describe diferentes casos de aplicación en visión por computadora.

Uno de los principales problemas desarrollados es el reconocimiento de lengua de señas. Este es un caso que presenta diversas aristas a atacar como el reconocimiento del intérprete, la segmentación de manos, la clasificación de diferentes configuraciones y de un gesto dinámico, entre otros problemas.

En esta área se llevan a cabo diversas investigaciones. Por un lado, se están desarrollando dos librerías para potenciar la carga, el manejo y procesamiento de bases de datos de lengua de señas, así como de imágenes de formas de mano correspondientes a las mismas.

Por otro lado, se está estudiando cómo mejorar el reconocimiento de formas de manos de la lengua de señas con Redes Neuronales Convolucionales. Un problema importante dentro de esa área es el de otorgar invarianza a la rotación para reconocer formas de mano. A partir de ese requerimiento, se han desarrollado análisis sobre la forma en que las redes neuronales convolucionales aprenden la invarianza a la rotación y otras transformaciones afines, no sólo en el dominio de las formas de mano sino en general.

Siguiendo con la línea de reconocimiento de patrones en video, se realizó un estudio de diferentes modelos de clasificación de peatones en la vía pública. El objetivo de esta investigación se centró en el análisis de la generalización de los modelos, midiendo cómo se comportaban luego de ser entrenado con un conjunto de datos y ser evaluado con otro.

Por último, se está desarrollando un proyecto de I+D+I para identificar diferentes atributos faciales en rostros humanos como emociones y otras características.

Palabras clave: Aprendizaje Automático, Visión por Computadoras, Redes Neuronales, Lengua de Señas, Reconocimiento de Peatones.

1. INTRODUCCION

El Instituto de Investigación en Informática LIDI tiene una larga trayectoria en el estudio, investigación y desarrollo de Sistemas Inteligentes basados en distintos métodos de Aprendizaje Automático. Los resultados obtenidos han sido medidos en la solución de problemas pertenecientes a distintas áreas.

En el III LIDI, desde hace varios años se viene trabajando en el procesamiento de señales de audio y video. Como resultado de estas investigaciones se han diseñado e implementado técnicas originales aplicables al reconocimiento tanto de gestos dinámicos como de detección de patrones en videos en diferentes problemas. En relación con esta línea, actualmente se están desarrollando los siguientes temas:

1.1. Reconocimiento de Lengua de Señas

El reconocimiento de la lengua de señas es un campo de investigación relativamente nuevo cuyo objetivo final es traducir de la lengua de señas a una lengua escrita. Esto implica poder procesar un video en donde una persona habla en lengua de señas, y reconocer la posición de su cuerpo, su cara y sus manos, la expresión de su rostro, la forma de sus manos y también la de sus labios si la seña requiere pronunciar la palabra para desambiguar. Con esa información, se debe reconocer la seña realizada, para luego a partir de una secuencia de señas generar una traducción a una lengua escrita [1,2].

El problema del reconocimiento de la lengua de seña es similar al del reconocimiento del habla, con la dificultad agregada de necesitar un procesamiento multimodal y la falta de bases de datos orientadas al entrenamiento de un sistema de reconocimiento. Además, las bases de datos disponibles varían ampliamente en el formato, metadatos, anotaciones y forma de obtención. Por estos motivos, el campo de la lengua de señas no ha logrado desarrollar sistemas con un nivel de desempeño similar al alcanzado por el reconocimiento del habla.

En esta área, en donde se ha terminado una tesis doctoral y otra está en progreso, se está trabajando en crear una librería para unificar y simplificar la carga y el preprocesamiento y etiquetado de bases de datos de lengua de seña dinámicas[1]. Se proveerá soporte para cargar una base de datos de Lengua de Señas Argentina desarrollada previamente [2], y otras más.

La librería a desarrollar buscar proveer una interfaz unificada para cargar y unir bases de datos, de modo de poder ampliar el repertorio de señas y de ese modo poder entrenar modelos más potentes. Al mismo tiempo, facilitará el ingreso a la investigación en el área, incrementando el nivel y la cantidad de los desarrollos.

1.2. Clasificación de formas de mano de la Lengua de Seña

Siguiendo con la temática de la sección anterior, una de las líneas investigadas es la clasificación de formas de manos. Las lenguas de señas utilizan un conjunto finito de formas de mano, que, en combinación con movimientos de las manos y el cuerpo, y expresiones faciales, se utilizan para señar [1].

En base a estudios previos [1,3], una etapa fundamental en el reconocimiento de la lengua de señas es la clasificación de estas formas de mano, y por ende un área prioritaria para mejorar el reconocimiento.

Las bases de datos de formas de manos para lengua de señas presentan problemas similares a las bases de datos completas de lengua de señas. Por ende, también se está trabajando en una librería para unificar su carga y preprocesamiento[2]. Esto es más importante aún de cierto modo para las formas de mano, porque si bien las lenguas de señas son regionales y un modelo entrenado para un país o región no sirve para otro, el conjunto de formas de mano es universal para todos los seres humanos, si bien cada lengua utiliza subconjuntos distintos.

Por otro lado, las Redes Neuronales Convolucionales (CNN, por sus siglas en

[1]https://github.com/midusi/sign_language_datasets

[2]https://github.com/midusi/handshape_datasets/

inglés) son actualmente el estado del arte en el área. En un estudio previo, se realizaron experimentos comparando diversas arquitecturas de CNN para clasificar formas de manos, pero sólo con dos bases de datos, LSA16 y RWTH [4]. Actualmente, se está extendiendo este análisis a varias bases de datos mediante el uso de la librería mencionada anteriormente.

1.3. Invarianza a la rotación en Redes Convolucionales

Una de las propiedades deseables en un modelo de clasificación de formas de mano es que sea invariante a la rotación, debido a que la forma de mano se considera la misma independientemente de la orientación que tenga.

Esto motivó el estudio de distintos modelos y técnicas que permitan que una CNN sea invariante a la rotación y otras transformaciones afines. En los últimos años, varios modelos han sido propuestos para añadir invarianza a la rotación en CNNs [6].

No obstante, no existe una comparación sistemática de dichos modelos, en especial contra un método simple y efectivo como la generación de datos transformados (*data augmentation)*. Por ende, se realizó una comparación de los modelos más emblemáticos del área con y contra *data augmentation*. Para la comparación, se utilizaron dos modelos de CNN: una red convolucional simple, y otra completamente convolucional. Los modelos fueron entrenados y probados con MNIST y CIFAR10, las dos bases de datos de clasificación más conocidas, con el objetivo de que los resultados sean fáciles de interpretar. Encontramos que estos modelos no ofrecen un desempeño significativamente superior vs *data augmentation,* aunque si pueden reducir el número de parámetros del mismo. Además, se analizó la posibilidad de aplicar modelos entrenados sin invarianza y reentrenarlos para que obtengan dicha invarianza. Se encontró que, sorpresivamente, las capas convolucionales son capaces de reentrenarse para ofrecer invarianza, y no solo las capas densas o completamente conexas pueden hacerlo, no obstante el hecho de que individualmente una capa convolucional no puede ser nunca invariante a la rotación[6].

Actualmente, se está ampliando la investigación en esta área, desarrollando un método para cuantificar la (in)varianza de cada unidad en una CNN de forma de comprender mejor como se codifica la invarianza en las redes convolucionales.

1.4. Detección de peatones

En el marco de la investigación sobre detección de objetos en video, se realizó un proyecto sobre detección de peatones en la vía pública. El objetivo estuvo centrado en el estudio de la transferencia de aprendizaje de los modelos entrenados para validar su efectividad al utilizarlos en entornos reales. Se eligieron tres bases de datos para entrenar los modelos de clasificación: INRIA, DAIMLER y TUD-Brussels, sumando más de 5.000 imágenes en total.

Luego, se analizaron diferentes descriptores con el fin de determinar su utilidad para incrementar la capacidad predictiva del modelo. Se estudiaron los Histograma de Gradientes Orientados (HOG) y Patrones Binarios Locales (LBP), siendo los más utilizados para estas tareas en el estado del arte. En tercer lugar, se entrenó una Máquina de Vectores de Soporte (SVM) con el fin de clasificar las imágenes de entrenamiento.

Se estableció un protocolo de experimentación para entrenar un modelo con cada conjunto de datos y evaluarlo en todos los restantes. Del mismo modo, se realizó un entrenamiento con diferentes conjuntos de datos mezclados para verificar si esto enriquecía la generalización del modelo entrenado.

Los resultados obtenidos mostraron que, si bien cada conjunto de datos presentaba escenas del mundo real, existen diferencias significativas que hacen que un modelo entrenado con un conjunto de imágenes no funcione apropiadamente con otro. Por otro lado, se encontró que al entrenar un modelo

con la combinación de diferentes bases de datos se genera una mayor transferencia de aprendizaje, aunque no siempre ayuda al entrenamiento de un conjunto de datos particular. Los resultados de esta investigación se encuentran publicados en [7].

1.5. Clasificación de atributos faciales

En el marco del proyecto de investigación, desarrollo e innovación "Análisis de Atributos Faciales" evaluado y financiado por la Facultad de Informática de la UNLP y continuando la línea de trabajos anteriores de reconocimiento y localización de objetos en video, se está trabajando en el desarrollo de aplicaciones capaces de reconocer atributos faciales en rostros humanos, como pueden ser emociones u otras características.

Para llevar esto a cabo, se están analizando diferentes bases de datos utilizadas en el estado del arte, como Facial Expressions in the Wild que contiene más de 80 mil imágenes obtenidas de internet con etiquetas de emociones básicas como *alegría, tristeza* o *enojo.* Otro conjunto de datos analizado es Large-scale CelebFaces Attributes (CelebA) Dataset, que contiene más de 200 mil imágenes de personalidades reconocidas con diferentes etiquetas sobre el rostro como el color del cabello, si tiene bigote, si usa anteojos, etc.

Como modelos de clasificación se están analizando diferentes arquitecturas de CNNs actuales, como VGG, Inception o ResNet. Si bien el grupo de trabajo ya ha realizado trabajos con estas arquitecturas, aún resta aplicarlas al procesamiento en tiempo real de rostros humanos.

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

- Estudio de técnicas para obtener invarianza a transformaciones afines en clasificadores de imágenes.
- Representación y clasificación de configuraciones de manos para el lenguaje de señas.
- Representación y clasificación de señas dinámicas.
- Estudio y desarrollo de modelos de Redes Convolucionales Profundas aplicados a problemas de visión por computadora.
- Estudio de la generalización de los modelos de detección de peatones.
- Análisis de diferentes representaciones aplicados a la clasificación de emociones y atributos faciales.

3. RESULTADOS OBTENIDOS/ESPERADOS

- Desarrollo de dos librerías para potenciar la investigación en reconocimiento de lengua de señas y clasificación de formas de mano.
- Comparación de varios modelos de redes convolucionales con invarianza a las transformaciones afines.
- Análisis de los mecanismos de las redes neuronales para adquirir invarianza a las transformaciones afines.
- Modelos de clasificación de peatones y sistema de detección en tiempo real.
- Sistema de clasificación de emociones humanas en rostros y descripción de atributos faciales.

4. FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo de la línea de I/D aquí presentada está formado por: 2 profesores con dedicación exclusiva, 2 becarios de investigación UNLP con dedicación docente, 1 becario CIN, 3 tesistas y 1 profesor extranjero.

Dentro de los temas involucrados en esta línea de investigación, en los últimos dos años se han finalizado 2 tesis de doctorado, 1 tesis de especialización, y 7 tesinas de grado de Licenciatura.

Actualmente se están desarrollando 1 tesis de doctorado, 1 tesis de especialista y 5 tesinas

de grado de Licenciatura. También participan en el desarrollo de las tareas becarios y pasantes del III-LIDI.

5. REFERENCIAS

[1] Ronchetti F., Quiroga F., Estrebou C., Lanzarini L., Rosete A. *Sign languague recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language*. Publicado en Ibero-American Conference on Artificial Intelligence IBERAMIA 2016 (pp. 338-349)

[2] Ronchetti, F., Quiroga, F., Estrebou, C.A., Lanzarini. *Handshape recognition for Argentinian Sign Language using ProbSom*. Journal of Computer Science & Technology, vol. 16, Nº 1, págs. 1-5, ISSN 1666-6038, 2016.

[3] Ronchetti, F., Quiroga, F., Estrebou, C. A., Lanzarini, L.C., Rosete, A. . *LSA64: An Argentinian Sign Language Dataset*, publicado en el XXII Congreso Argentino de Ciencias de la Computación (CACIC 2016) (pp. 794-803).

[4] Koller, O., Ney, H., Bowden, R. . *Deep Hand: How to Train a CNN on 1 Million Hand Images When Your Data Is Continuous and Weakly Labelled.* Computer Vision and Pattern Recognition Conference (CVPR 2016) (pp. 3793-3802).

[5] Quiroga, F., Antonio, R., Ronchetti, R., Lanzarini, L., Rosete, A. *A Study of Convolutional Architectures for Handshape Recognition applied to Sign Language,* publicado en el XXIII Congreso Argentino de Ciencias de la Computación (CACIC 2017) (pp. 13-22).

[6] Quiroga F., Ronchetti F., Lanzarini L., Fernandez-Bariviera A. *Revisiting Data Augmentation for Rotational Invariance in Convolutional Neural Networks*. International Conference on Modeling and Simulation in Engineering, Economics and Management (MS'2018 GIRONA) (en prensa).

[7] Camele G, Quiroga F., Ronchetti F., Hasperué W., Lanzarini L. *Transferencia de Aprendizaje para Clasificación de Peatones.* XXIV Congreso Argentino de Ciencias de la Computación. CACIC 2018. Tandil. Octubre 2018.