Deep Learning para Visión por Computadora.

Franco Ronchetti[1,2], Facundo Quiroga[1,3], Gastón Rios[1,3], Pedro Dal Bianco[1,3], Santiago Ponte Ahon[1], Oscar Stanchi[1], Laura Lanzarini[1], Alejandro Rosete[4], Waldo Hasperué[1]

[1] Instituto de Investigación en Informática LIDI, Facultad de Informática, Universidad Nacional de La Plata, La Plata, Argentina.*
[2] Comisión de Investigaciones Científicas de la Pcia. De Bs. As. (CICPBA)
[3] Becario postgrado UNLP
[4] 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)

Contacto: fronchetti@lidi.info.unlp.edu.ar

CONTEXTO

Esta presentación corresponde a algunas de las tareas de investigación que se llevan a cabo en el III-LIDI en el marco del proyecto F025 "Sistemas inteligentes. Aplicaciones en reconocimiento de patrones, minería de datos y big data" perteneciente al Programa de Incentivos (2018-2023).

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, video y, utilizando técnicas de Aprendizaje Automático clásicas, y Aprendizaje profundo con Redes Neuronales Convolucionales, Recurrentes y Transformers. El trabajo presentado describe diferentes estrategias inteligentes para la traducción automática de la lengua de señas, particularmente enfocado en la Lengua de Señas Argentina (LSA), junto con herramientas para su aprendizaje e interpretación de resultados.

Este es un problema complejo y multidisciplinar, que presenta diversos subproblemas a resolver como el reconocimiento del intérprete que realiza una seña, la segmentación de manos, la clasificación de diferentes configuraciones y de un gesto dinámico, entre otros.

Por otro lado, la creación de un conjunto de datos apropiado para la LSA resulta esencial para la creación de modelos específicos. Uno de los subproblemas atacados en este artículo es la creación de un conjunto de datos apropiado.

Por otro lado, se está estudiando la forma de reconocer formas de mano de la Lengua de Señas con conjuntos de datos de tamaño reducido, dada la falta de datos de entrenamiento para este dominio.

Por último, se están utilizando Redes Generativas Adversarias (GANs) para aumentar bases de datos de formas de mano, con el objetivo de complementar desde otro enfoque el entrenamiento de modelos para su clasificación.

En otra línea de investigación, se está estudiando la forma en que las redes neuronales codifican la invarianza a las transformaciones y otras propiedades transformacionales, con el objetivo de poder analizar y comparar estos modelos. De esta forma se espera poder mejorar los modelos de clasificación de objetos transformados, en particular, de formas de mano. En el mismo sentido, se están estudiando técnicas de interpretabilidad de los modelos.

Palabras clave: Redes Neuronales, Redes Convolucionales, Redes Recurrentes, Visión por Computadoras, Lengua de Señas, Bases de datos, *Crowdsourcing*, Redes Generativas Adversarias, Invarianza, Equivarianza.

1. INTRODUCCION

El Instituto de Investigación en Informática LIDI (III-LIDI) tiene una larga trayectoria en el estudio, investigación y desarrollo de Sistemas Inteligentes basados en distintos métodos de Aprendizaje Automático y Redes Neuronales.

Como resultado de estas investigaciones se han diseñado e implementado técnicas originales aplicables a la clasificación y el análisis de características de objetos en imágenes, generación de imágenes para aumentación de datos, y estudio del funcionamiento de las redes neuronales. En relación con esta línea, actualmente se están desarrollando los siguientes temas:

1.1. Reconocimiento y traducción de lengua de señas

La traducción de la lengua de señas (SLT) es un campo de estudio activo que abarca la interacción humano-computadora, la visión por computadora, el procesamiento de lenguaje natural y el aprendizaje automático. En esta línea de investigación, el objetivo es el reconocimiento de gestos en vídeos de Lengua de Señas y la traducción de esta al español. Para el reconocimiento se está trabajando con la base de datos LSA64 [11]. Esta consiste en un registro de 64 señas de la Lengua de Señas Argentina. Se compararon diversas técnicas en el estado del arte del Aprendizaje Automático basadas en Redes Neuronales. Específicamente, se compararon tanto arquitecturas basadas en Redes Recurrentes y Convolucionales, como distintas estrategias de preprocesamiento para optimizar la calidad del reconocimiento. También se están analizando los modelos entrenados para comprender mejor el impacto de estas estrategias de preprocesamiento y de la forma de representación lograda por cada tipo de arquitectura [2][11]. Para la traducción, se confeccionó LSA-T[10] una base de datos masiva y realista de la Lengua de Señas Argentina, que se detalla en la sección 1.2. Este problema representa una mayor complejidad que el reconocimiento ya que implica reconocer varias señas ejecutadas una tras otra en un mismo video y luego traducir las señas identificadas al español, cuya sintaxis y semántica son distintas.

Figura 1. Activación de las capas de una red recurrente para el reconocimiento de lengua de señas en video

1.2. Creación de una base de datos para traducción de Lengua de Señas

Esta línea de investigación tiene como objetivo generar y expandir el primer conjunto de datos continuo de Lengua de Señas Argentina (LSA). Esto es de particular relevancia ya que, en el campo de la traducción de lengua de señas, es necesario contar con bases de datos específicas de la lengua que se busca traducir. No es posible utilizar modelos de traducción entrenados sobre otra lengua de señas para traducir LSA. En este marco, se desarrolló la base de datos “LSA-T”, que contiene 14.880 videos a nivel de oración de LSA extraídos del canal de YouTube CN Sordos. Contiene, además, para cada video, su traducción al español e información posicional correspondiente a cada señante. También se presentó un método para inferir al señante entre varias personas que pudieran aparecer en un mismo video, un análisis detallado de las características del conjunto de datos, una herramienta de visualización para explorar el conjunto de datos y un modelo de SLT neuronal para servir como línea de base para futuros experimentos.

Este desarrollo se lleva a cabo en el marco de una tesis doctoral financiada por la UNLP a través de una beca de postgrado con el objetivo de contribuir al avance de la traducción de la lengua de señas y, en última instancia, mejorar la integración de las personas sordas en la sociedad.

Figura 2. Captura del canal CN Sordos, cuyos videos componen LSA-T

1.3. Generación de imágenes sintéticas

Adicionalmente, se están comenzando a utilizar Redes Generativas Adversarias (GANs) y Modelos de Difusión para generar imágenes y luego videos artificiales relacionadas con la lengua de señas. Este tipo de redes permitirá, por un lado, generar videos de LSA a partir de texto, a modo de traducción y por otro, aumentar las bases de datos de formas de mano, con el objetivo de complementar desde otro enfoque el estudio de modelos y algoritmos de clasificación para bases de datos con pocos datos etiquetados. Estas investigaciones son llevadas a cabo en marco de una tesis doctoral financiada por la UNLP a través de una beca de postgrado.

1.4. Aplicación web para reconocimiento de lengua de señas y crowdsourcing

Esta línea de investigación responde, a través del desarrollo de una aplicación web, a dos objetivos específicos. Por un lado, facilitar el acceso a los modelos de traducción para ayudar en la práctica y comunicación a través de lengua de señas. Por otro, la recopilación de videos para ampliar las bases de datos con los que se entrenan dichos modelos a través del *crowdsourcing*. Esto busca subsanar el hecho de que, al disponer de bases de datos obtenidas *in the wild* (por fuera del laboratorio), muchas veces la distribución de los datos es irregular, presentando muchas señas, frases o palabras que aparecen muy pocas veces a lo largo de toda la base de datos.

En función de esto, la aplicación web sugiere a los usuarios frases puntuales a interpretar y permite la grabación de un video corto del usuario interpretando LSA a través de la cámara de cualquier dispositivo electrónico. Finalmente, muestra la traducción de dicho video a texto en castellano generado por un modelo de SLT. Dicho modelo es ejecutado en el mismo navegador a través de Tensorflow-JS y funciona a través de la información posicional extraída de los videos a través de otro modelo de la librería Mediapipe [5]. Los videos grabados, con previa autorización, se almacenan con la intención de ser utilizados para extender dichas bases de datos. Según el video, esto puede requerir de un preprocesamiento o de la verificación de un intérprete experto.

1.5. Interpretabilidad de modelos de Deep Learning

La interpretabilidad es un campo de investigación que tiene como objetivo estudiar técnicas para comprender los estímulos por los cuales modelos de IA de caja negra generan determinadas salidas. Se utiliza con un rol diagnóstico para descubrir cómo es la contribución de las capas ocultas en los modelos.

En el marco de esta línea de investigación se realizó una implementación de RISE (Randomized Input Sampling for Explanation) para Captum. RISE es un método post-hoc y local, concentrado en modelos de caja negra ya que no requiere acceso a los parámetros internos de los modelos. Este genera un mapa de importancia que, al aplicarlo a imágenes, permite visualizar las regiones de la misma que fueron importantes para la predicción de determinada clase. Captum es una librería de interpretabilidad de modelos para PyTorch que permite implementar más fácilmente los algoritmos de interpretabilidad que pueden interactuar con los modelos de PyTorch.

Para probar esta implementación, se utilizaron los conjuntos de datos EyePACS y EyeQ cuyas imágenes sirven para evaluar retinopatía diabética y la calidad de las imágenes para esta patología, respectivamente. Se realizaron pruebas con la implementación de RISE para interpretar los resultados que arroja un modelo de evaluación de calidad de imagen.

1.6. Métricas de Equivarianza

Las redes neuronales son modelos tradicionalmente considerados como de caja negra. En los años recientes, se han realizado varios esfuerzos para comprender su funcionamiento de forma tal que el mismo sea más predecible o modulable.

La invarianza y equivarianza a las transformaciones son propiedades deseables en varios modelos de redes debido a que nos permiten razonar más fácilmente respecto a su funcionamiento.

En los últimos años, varios modelos fueron propuestos para añadir invarianza a la rotación y otras transformaciones en CNNs [3]. No obstante, no está claro cómo estos modelos impactan en el aprendizaje usual de los pesos de la red.

Por este motivo, se continúa con la utilización de las métricas previamente definidas [4] para caracterizar modelos de redes neuronales típicos, ya sea desde capas muy utilizadas como Batch Normalization, Dropout, Max-Pooling, arquitecturas completas como Residual Networks, VGG o AllConvolutional, y arquitecturas especializadas como TI-Pooling.

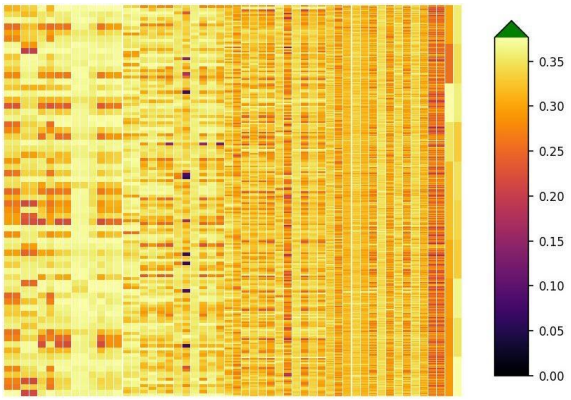

Figura 3. Invarianza por capas y unidades de una CNN con arquitectura ResNet.

2. LINEAS DE INVESTIGACIÓN Y DESARROLLO

- Redes neuronales profundas, convolucionales y transformers para análisis de secuencia.
- Invarianza y auto-equivarianza en redes neuronales.
- Reconocimiento de lenguaje de señas.
- Generación de imágenes con GANs.
- *Crowdsourcing* para la extensión de bases de datos.

3. RESULTADOS OBTENIDOS/ESPERADOS

- Comparación de modelos especializados para bases de datos con pocas muestras para la clasificación de formas de mano.
- Desarrollo de métricas de invarianza y auto-equivarianza para redes neuronales.
- Análisis de modelos para el reconocimiento y traducción de lengua de señas en video.
- Redes generativas para la creación de datos artificiales en la Lengua de Señas.
- LSA-T como primer base de datos continua de LSA.
- Sistema Web para Reconocimiento y *Crowdsourcing* de Lengua de Señas Argentina.

4. FORMACIÓN DE RECURSOS HUMANOS

El grupo de trabajo de la línea de I/D aquí presentada está formado por: 3 profesores con dedicación exclusiva, 1 investigador CIC-PBA, 3 becarios de posgrado de la UNLP con dedicación docente, 2 becarios CIN y 1 profesor extranjero.

Dentro de los temas involucrados en esta línea de investigación, en los últimos dos años se han finalizado diversas tesinas de grado, como así también trabajos de final de carrera y de especialización.

Actualmente se están desarrollando 2 tesis de doctorado, 2 tesis de especialista, 1 de mágister, 3 tesinas de grado de Licenciatura y 2 trabajos finales de Ingeniería en Computación. También participan en el desarrollo de las tareas becarios y pasantes del III-LIDI.

5. REFERENCIAS

[1] 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). 2017

[2] Cornejo Fandos, U., Rios, G., Ronchetti, F, Quiroga, F., Hasperué, W., Lanzarini, L. Recognizing Handshapes using Small Datasets, publicado en el XXV Congreso Argentino de Ciencias de la Computación (CACIC 2019, Rio Cuarto) (pp. 105-114). 2019.

[3] 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). 2018.

[4] Quiroga, F., Torrents-Barrena, J., Lanzarini, L., & Puig, D. Measuring (in) variances in Convolutional Networks. In Conference on Cloud Computing and Big Data (pp. 98-109). Springer, Cham. 2019.

[5] Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., ... & Grundmann, M.. Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172. 2009.

[6] Goodfellow I. J., Pouget-Abadie j., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y. Generative Adversarial Networks. NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems. v2. pp 2672-2680. 2014.

[7] Jaschek M., Slettebak A., Jaschek C. *Be star terminology*. Be Star Newsletter. 1981.

[8] Potash, P., Romanov, A., and Rumshisky, A. Ghostwriter: Using an lstm for automatic rap lyric generation. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 1919-1924. 2015.

[9] Young, T., Hazarika, D., Poria, S., and Cambria, E. Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 13(3):55. 2018.

[10] Dal Bianco, P., Ríos, G., Ronchetti, F., Quiroga, F., Stanchi, O., Hasperué, W., & Rosete, A. (2023, January). LSA-T: The First Continuous Argentinian Sign Language Dataset for Sign Language Translation. In Advances in Artificial Intelligence–IBERAMIA 2022: 17th Ibero-American Conference on AI, Cartagena de Indias, Colombia, November 23–25, 2022, Proceedings (pp. 293-304). Cham: Springer International Publishing.

[11] Iván Mindlin, Facundo Quiroga, Franco Ronchetti, Pedro Dal Bianco, Gastón Ríos, Laura Lanzarini, Waldo Hasperué. "A Comparison of Neural Networks for Sign Language Recognition with LSA64". JCC-BD&ET: Conference on Cloud Computing, Big Data & Emerging Topics. pp 104-117. Springer, Cham. Junio 2021.