MARÇAL RUSIÑOL, LLUÍS GÓMEZ

Avances en clasificación de imágenes en los últimos diez años

Perspectivas y limitaciones en el ámbito de archivos fotográficos históricos

Advances in the classification of images in the last ten years

Prospects and constraints in the field of photographic historical archives

Marçal Rusiñol, marcal@cvc.uab.es
Lluís Gómez, lgomez@cvc.uab.es
Centre de Visió per Computador, Universitat Autònoma de Barcelona

Citación: Rusiñol, Marçal; Gómez, Lluís (2018). "Avances en clasificación de imágenes en los últimos diez años. Perspectivas y limitaciones en el ámbito de archivos fotográficos históricos". *Tábula,* n. 21, pp. 161-174

Recibido: 28-07-2018. Aceptado: 12-10-2018

Resumen analítico / Analitic summary

En este artículo presentaremos un resumen de los últimos avances en el tema de la clasificación de imágenes. Veremos cómo gracias a la aparición de bases de datos y competiciones de clasificación de imágenes a gran escala, las redes neuronales convolucionales emergieron y revolucionaron por completo el campo de la visión por computador. Nos centraremos en analizar cuáles son las limitaciones a presente del estado del arte cuando deseamos aplicar dichas metodologías al campo de los archivos fotográficos históricos.

VISIÓN POR COMPUTADOR | RECONOCIMIENTO DE PATRONES | INTELIGENCIA ARTIFICIAL | APRENDIZAJE COMPUTACIONAL | APRENDIZAJE PROFUNDO | REDES NEURONALES CONVOLUCIONALES | CLASIFICACIÓN DE IMÁGENES

In this paper we present an overview of the latest achievements on the topic of image classification. We will see how thanks to the large scale image classification datasets and competitions such as Pascal VOC and ImageNet, deep convolutional neural networks took rise and revolutionized the computer vision field. We will then focus on analyzing which are the current limitations of the state of the art when applying such methods to the particular field of historic photographic archives.

COMPUTER VISION | PATTERN RECOGNITION | ARTIFICIAL INTELLIGENCE | MACHINE LEARNING | DEEP LEARNING | CONVOLUTIONAL NEURAL NETWORKS | IMAGE CLASSIFICATION

Introducción.
La visión por computador y el análisis de documentos

La visión por computador es la disciplina científica enmarcada dentro del campo de la inteligencia artificial y del reconocimiento de patrones centrada en el diseño de algoritmos computacionales capaces de analizar imágenes digitales, interpretar sus contenidos, entenderlas y devolver de manera estructurada información relevante que pueda ser tratada por un computador. De la misma manera que los humanos usamos nuestro sistema de visión para poder comprender el entorno que nos rodea, la visión por computador tiene como objetivo reproducir el mismo efecto para que los ordenadores sean capaces de interpretar los contenidos de una imágen y actuar en consecuencia.

La visión por computador apareció a finales de la década de los sesenta, en universidades que en la época empezaban a investigar el campo de la inteligencia artificial. Queriendo dotar a los robots de un comportamiento inteligente, se empezó a intentar imitar el sistema visual de los humanos. A primera vista se creyó que dicho problema se podía atacar fácilmente, y, en 1966, investigadores del Massachusetts Institute of Technology (MIT) intentaron "resolver" el problema a través de un proyecto de verano donde se conectó una cámara a un ordenador y se intentó implementar los mecanismos computacionales para que el computador "describiese lo que veía" (Papert, 1996). Rápidamente se evidenció que el problema de la visión por computador no era tan sencillo de abordar. En la década de los setenta, se cimentaron las primeras bases para muchos de los algoritmos de visión por ordenador y del reconocimiento de patrones estructurados que existen hoy, incluyendo la extracción de contornos en imágenes, el etiquetado

de líneas o la estimación de movimiento al analizar flujos de vídeo (Szeliski, 2010). Hasta que a finales de la década de los noventa, el campo de la visión por computador se empezó a cruzar con otros campos de la informática como los gráficos por ordenador o el aprendizaje computacional (Sebe, 2005).

A día de hoy, la visión por computador es una tecnología clave habilitadora en sectores como la automoción, los deportes, el entretenimiento, la electrónica de consumo, la robótica, la fabricación avanzada, la salud o la seguridad, que constantemente lanzan nuevos productos y servicios. Se estima que el mercado de visión por computador crecerá un 40% anualmente, alcanzando ingresos de $ 50,000M en 2022 (Tractica, 2016). La visión por ordenador no deja de ser una ciencia extremadamente aplicada, y por eso hoy en día aparece en campos muy diversos de nuestro día a día. Por ejemplo, la visión por computador se usa extensivamente en entornos industriales, para realizar controles de calidad en líneas de producción. En el ámbito de la conducción autónoma de vehículos, tan de moda últimamente, gracias a los avances de gigantes como Google o Tesla, los coches van equipados con varias cámaras que registran el entorno en tiempo real. Encontramos técnicas de visión por computador en el reconocimiento de gestos, presente por ejemplo en los últimos modelos de videoconsolas como Nintendo Wii o Kinect. La visión por computador aparece también en el sector médico, donde se usa como asistencia a los facultativos para el diagnóstico mediante análisis de imágenes médicas provenientes de rayos X, TACs, etc. En los algoritmos de reconocimiento facial integrados en nuestras cámaras compactas o teléfonos móviles. O simplemente cada vez que realizamos una búsqueda de imágenes en los motores de búsqueda en internet como Google o Bing.

Existen por otro lado, aplicaciones de la visión por computador más cercanas al ámbito archivístico. Dentro del campo de la visión por ordenador, el análisis de documentos aborda el problema de reconocer de manera automática el contenido de documentos (ya sea texto impreso, texto escrito a mano o elementos gráficos). Se puede considerar que el origen del análisis de documentos surge en los años sesenta, cuando aparecen los primeros sistemas de reconocimiento óptico de caracteres (OCR). Aunque resulta curioso que las primeras invenciones de sistemas de reconocimiento de texto datan de mucho antes de la aparición de la visión por computador. En 1914, Emanuel Goldberg ya desarrolló una máquina que leía caracteres y los convertía en códigos de telégrafo (Herbert, 1982). Los sistemas de OCR permiten reconocer caracteres impresos y fuentes de texto tanto provenientes de computadoras como de máquinas de escribir, y codificarlos en formato electrónico para que un ordenador pueda interpretarlos (Cao 2014). Los sistemas de OCR integran un modelo óptico de la forma de las letras y un modelo lingüístico sobre las probabilidades de que estas se combinen según el lenguaje de escritura. Por tanto, los programas de OCR reconocen agrupaciones de píxeles como letras y, en un nivel superior, validan las

interpretaciones conjuntas para acabar transformando una imagen en un archivo editable de palabras. El software de OCR ha evolucionado mucho y tiene hoy en día buenas prestaciones, especialmente en documentos impresos y con digitalización de calidad. Las aplicaciones de ofimática y los escáneres domésticos suelen incorporar softwares de OCR que nos permiten transcribir automáticamente los documentos cotidianos para tratarlos a continuación con procesadores de textos. Comercialmente, grandes corporaciones como Nuance (OMNIPAGE), ABBYY (FineReader) o Google (Tesseract) ofrecen buenos sistemas de OCR, sin embargo, estos sistemas tienen todavía restricciones, y la investigación en análisis de documentos tiene que avanzar para poder ofrecer soluciones a gran escala. Un caso de gran interés son los documentos manuscritos (Tulyakov, 2014), (Frinken, 2014), y en particular los documentos históricos, que pueden estar degradados, escritos en lenguas antiguas o presentarse de manera no estructurada.

Entorno a estas fuentes documentales históricas, la informática y las humanidades convergen en el ámbito de las humanidades digitales, un área emergente y interdisciplinaria. Los documentos custodiados en archivos históricos, administrativos o eclesiásticos contienen información muy valiosa, que recoge la memoria histórica de la sociedad. La digitalización masiva de estos fondos permite construir archivos digitales de imágenes que a menudo son accesibles a través de los portales web de las instituciones que los custodian. Este acceso, sin embargo, no suele ir más allá de navegadores que permiten visualizar las imágenes de manera lineal, página a página. En estos casos, las tecnologías de visión por computador y de análisis de imágenes de documentos pueden resultar beneficiosas. Pero en este artículo nos alejaremos de los fondos documentales, entendidos como documentos textuales, y nos centraremos en los archivos fotográficos. Emplazamos al lector interesado en el uso de técnicas de visión para archivos textuales a leer el artículo donde se analizaba esta problemática (Fornés 2016).

Clasificación de imágenes

Uno de los problemas centrales en visión por computador, y, por ende de los más estudiados, es la clasificación de imágenes. El problema de clasificación de imágenes se puede definir como la tarea de asignar una o varias etiquetas predefinidas de un conjunto fijo de categorías a la imagen que se está analizando dependiendo de sus contenidos. Podemos ver en la Figura 1, un ejemplo de qué imágenes se asignan a ciertas clases (en este caso concreto: “gato”, “perro”, “taza” y “gorro”) dependiendo de sus contenidos.

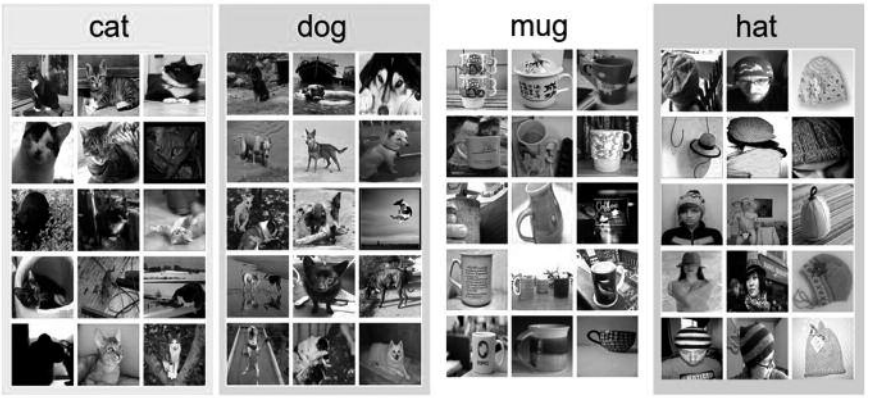

Figura 1. Ejemplo de clasificación de imágenes por contenido

La tarea de clasificar imágenes en base a los conceptos visuales contenidos en ellas es una tarea relativamente trivial para el ser humano. De todas maneras, desde la perspectiva de un algoritmo de visión por computador que deba realizar la misma tarea de manera automática, se presentan varios desafíos:

- *Variación del punto de vista*. Una sola instancia de un objeto tridimensional se puede orientar de muchas maneras con respecto a la cámara. Los algoritmos deben pues aprender la apariencia de dichos objetos de manera invariante a su posición.
- *Variación de escala*. Las clases visuales a menudo muestran variación en su tamaño dependiendo desde donde se haya tomado la fotografía y del uso de zoom. Los algoritmos de clasificación deben ser invariantes al tamaño de los contenidos con tal de reconocerlos tanto si aparecen en primer plano ocupando toda la fotografía como si aparecen de fondo y no son más que un mero detalle.
- *Deformaciones*. Muchos objetos de interés no son cuerpos rígidos y pueden deformarse de manera extrema.
- *Oclusiones*. Los objetos de interés pueden aparecer sólo en parte debido a tener algún otro elemento que lo tape al situarse entre el objeto de interés y la cámara. A veces, solo una pequeña porción de un objeto podría ser visible, pero aún así es deseable que los algoritmos de visión sean capaces de clasificar estas imágenes.
- *Condiciones de iluminación*. Las fotografías se pueden tomar en condiciones lumínicas drásticamente diferentes, y, por ende, los objetos que aparecen en ellas pueden variar enormemente su distribución de colores.

- *Fondo*. Los objetos de interés pueden camuflarse en su entorno, lo que los hace difíciles de identificar.
- *Variación intraclase*. Las clases de objetos de interés a menudo pueden ser relativamente amplias. Tomemos como ejemplo una silla. Hay muchos tipos diferentes de sillas, cada uno con su propia apariencia y que pueden llegar a ser muy diferentes entre sí, pero aún así se desea que el algoritmo de visión trabaje a un nivel de abstracción alto y las clasifique a todas con la misma etiqueta.

Las competiciones PASCAL VOC y ImageNet

En el campo de la visión por computador, la manera de poder evaluar el rendimiento de los diferentes métodos que se proponen en la literatura suele ser mediante el uso de bases de datos anotadas públicas. Estas bases de datos vienen partidas en dos conjuntos disjuntos. Las imágenes contenidas en uno de ellos se utilizan como muestras para "instruir" a los algoritmos de aprendizaje computacional sobre qué es lo que deben reconocer en las imágenes y etiquetar en consecuencia, y las imágenes del otro conjunto se usan para poner a prueba el algoritmo, cuantificar cuál es su tasa de error, y poder así comparar de manera justa y en igualdad de condiciones diferentes métodos. Para ser útiles, dichas bases de datos deben ser representativas de los datos reales que uno se puede encontrar al aplicar el algoritmo, y, deben contener un volumen de datos suficiente para poder realizar correctamente los entrenamientos de los métodos de aprendizaje computacional, así como tener suficientes muestras para que estos puedan llegar a conclusiones que sean estadísticamente significativas.

Una década atrás, la base de datos más utilizada en el campo de la clasificación de imágenes era la base de datos PASCAL VOC (Everingham, 2010), (Everingham, 2015), con la cual, una vez cada año desde 2007 hasta 2012 se organizaba una competición donde los participantes debían enviar sus métodos y se publicaba un ranking de los métodos más prometedores año trás año.

La base de datos PASCAL VOC cambiaba edición tras edición, pero siempre mantenía su definición de clases. En particular, se habían definido 20 clases a las cuales asignar cada una de las imágenes, y contaba con unas 5.000 imágenes etiquetadas para el entrenamiento de los algoritmos (training set), y otras 5.000 para la fase de pruebas (test set). Las clases estaban organizadas en diferentes grupos: **Vehículos**: Avión, Bicicleta, Barco, Autobús, Motocicleta, Coche y Tren. **Objetos:** Botella, Silla, Mesa, Planta, Sofá y Televisor. **Animales:** Perro, Gato, Pájaro, Vaca, Caballo y Oveja, y finalmente la clase **Persona**.

En su última edición en 2012, el mejor método presentado obtuvo un 82% de acierto en el conjunto de test.

Durante esa época, mientras la comunidad de visión por computador se centraba en intentar mejorar los algoritmos de visión, aprendizaje y clasificación, para poder obtener mejores resultados, independientemente de los datos usados en fase de entrenamiento, Li Fei-Fei, investigadora de la Universidad de Illinois, se dio cuenta de una limitación de este enfoque: el mejor algoritmo no funcionaría si los datos sobre los que aprendía no reflejaban el mundo real. Su solución fue construir una mejor base de datos, que revolucionaría el campo de la visión. En 2009, el equipo de Li Fei-Fei presentó ImageNet, una base de datos inmensa de más de 3 millones de imágenes y 5.000 categorías diferentes (Deng, 2009). En 2010, se organizó una competición sobre un subconjunto de dicha base de datos: 200.000 imágenes entre entrenamiento y test, y 1.000 categorías diferentes (Russakovsky, 2015).

Auge del *Deep Learning*

Entre 2010 y 2012 la comunidad fué abandonando el uso de PASCAL VOC y centrándose en ImageNet, al percatarse que sin necesidad de realizar mejoras sustanciales en los métodos y algoritmos de base, aquellos producían resultados muy aceptables en una tarea a priori mucho más compleja, simplemente por el "poder" del volúmen de los datos.

En aquellos años, la tendencia era aplicar un algoritmo de visión por computador que se conocía como *Bag of Visual Words*, que consistía en describir las imágenes en base a contar cuantas apariciones tenían diferentes descriptores locales (Jun, 2007). Dichos descriptores se juntaban con modelos de clasificadores estadísticos para poder clasificar las imágenes de entrada.

Pero, tan sólo dos años después de la primera competición de ImageNet, en 2012, sucedió algo aún más grande. De hecho, si el auge de la inteligencia artificial que estamos viviendo hoy pudiera atribuirse a un solo evento, sería el anuncio de los resultados del desafío ImageNet 2012.

Geoffrey Hinton, Ilya Sutskever y Alex Krizhevsky de la Universidad de Toronto presentaron una arquitectura de red neuronal convolucional profunda llamada AlexNet (Krizhevsky, 2012), que todavía se utiliza en la investigación hasta la fecha, que superó el segundo mejor resultado por un margen de 10.8 puntos porcentuales. Podemos ver en la Figura 2 unos resultados cualitativos de la red AlexNet sobre la base de datos ImageNet, y en la Figura 3, un esquema de la arquitectura AlexNet.

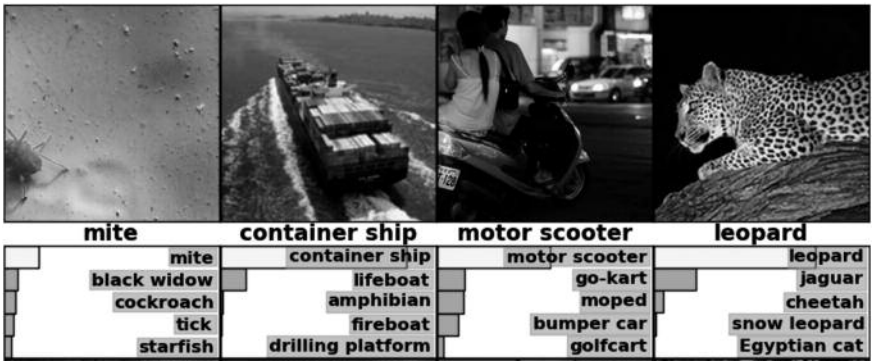

Figura 2. Resultados de clasificación obtenidos por AlexNet en la base de datos Imagenet. El algoritmo devuelve las 5 clases más probables ordenadas de mayor a menor. Como podemos observar, incluso las categorías que no están en primera posición tienen cierto sentido semántico en relación a los contenidos visuales de las imágenes tratadas

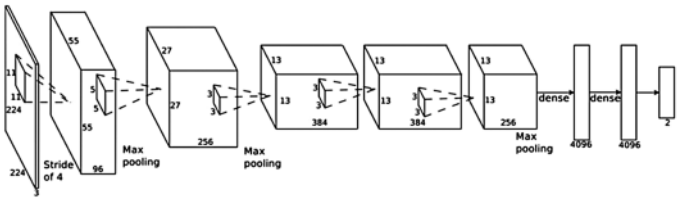

Figura 3. Esquema de la arquitectura de red neuronal convolucional AlexNet que en cierta medida revolucionó el campo de la visión por computador

Tal fue el impacto causado por el anuncio de estos resultados, que en tan sólo un año, todos los participantes en la competición de 2013 habían abandonado por completo los métodos basados en *Bag of visual words* y todos sin excepción usaban métodos de *Deep Learning* basados en redes neuronales convolucionales (CNN por sus siglas en inglés). Si en 2012 con AlexNet el grupo del profesor Hinton fue el primero en obtener unos resultados por debajo del 25% de error en ImageNet, en 2013 casi todos los participantes estaban también por debajo de ese 25%. Como podemos ver en la Figura 4, tan sólo cinco años después de ese punto de ruptura, la mayoría de métodos en la edición del 2017 están por debajo del 5% de error, llegando incluso a superar el rendimiento de clasificación humano.

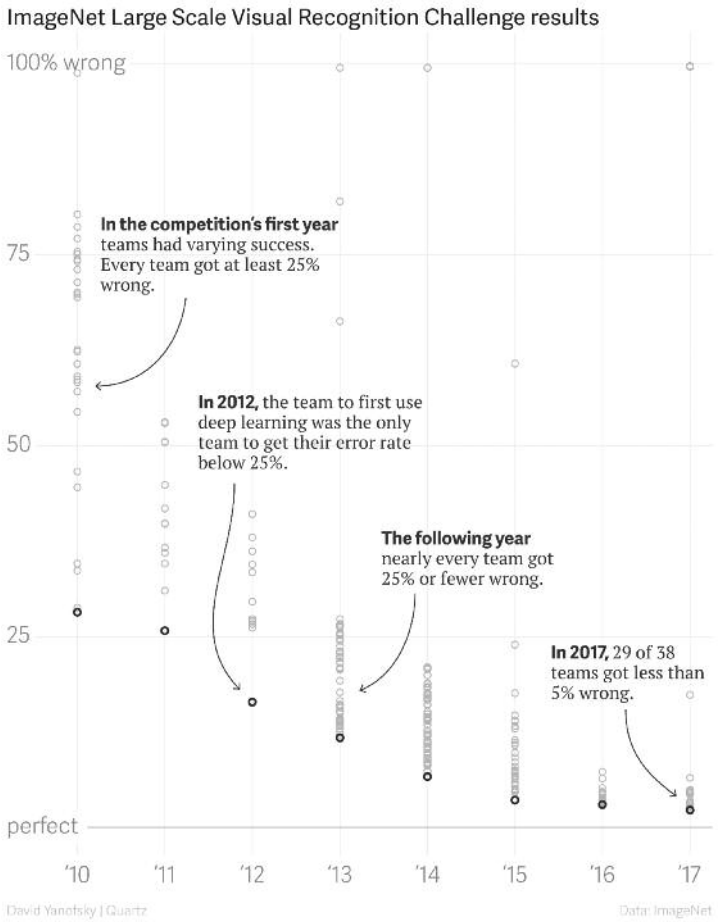

Figura 4. Distribución de las tasas de acierto en ImageNet a lo largo de los años. Se puede observar la gran diferencia entre el primer concursante y el resto en el año 2012, y como en tan sólo un año, casi todos los participantes habían adoptado el uso de las redes neuronales y estaban casi todos por debajo del 25% de error. En tan sólo cinco años, se ha reducido el error en esta base de datos en 20 puntos. (Imagen reproducida de Quartz)

A lo largo de estos cinco últimos años, caben destacar las arquitecturas listadas en la Tabla 1. A día de hoy son las más usadas en varios ámbitos y han demostrado su poder de generalidad. Hay que tener en cuenta que no sólo es importante el descenso en la tasa de error en clasificación, sino también lo complejas que pueden llegar a ser estas arquitecturas de redes neuronales. Podemos

ver que para estas arquitecturas estamos hablando de modelos matemáticos que tienen decenas de millones de parámetros. Esta complejidad influye de manera negativa en varios aspectos. Por un lado en el tamaño de dichos modelos, que pueden llegar a ocupar cientos de Megabytes. En segundo lugar, en el coste computacional. Cuantos más grandes y más parámetros tenga el modelo más tiempo tardará en clasificar imágenes de entrada, y, de manera más preocupante, más tardará el modelo en poder aprender y acabar de entrenar. Los modelos más grandes pueden llegar a necesitar varias semanas para aprender a clasificar imágenes. Finalmente, resulta curioso observar como no siempre existe una correlación entre modelos más complejos y grandes y su rendimiento. GoogleNet es netamente más pequeña que VGG16, y aún así logra resultados mejores.

Tabla 1. Estado del arte en redes neuronales convolucionales para la clasificación de imágenes en ImageNet

MÉTODO	PORCENTAJE DE ERROR EN IMAGENET	NÚMERO DE PARÁMETROS DEL MODELO
AlexNet (Krizhevsky, 2012)	15.3%	60 millones
VGG16 (Simonyan, 2015)	7.3%	138 millones
GoogleNet (Szegedy, 2014)	6.7%	5 millones
Inception v3 (Szegedy, 2015)	3.58%	23 millones
ResNet (He, 2015)	3.57%	25 millones

Aplicación en el ámbito de archivos fotográficos históricos

Ahora bién, ¿se podrían usar estas tecnologías en el ámbito de archivos fotográficos? ¿Resultarían interesantes? ¿Cuáles serían sus limitaciones?

En primer lugar, creemos que esta tecnología tiene un gran potencial de aplicación en el marco de los archivos fotográficos. Evidentemente, si se quiere tener una colección de imágenes bien curada, el uso de herramientas automáticas

siempre dejará lugar a dudas. Pero, por otro lado, estas herramientas pueden resultar muy convenientes no tanto para etiquetar de manera automática y sin ningún tipo de supervisión humana colecciones enteras de fotografías, sino más bien para tratar de asistir al archivero en el proceso de anotación de metadatos de un fondo de fotografías históricas, facilitando así este complejo proceso.

Hay que tener en cuenta también otro factor a la hora de ponderar si la tecnología actual podría ser útil en el campo archivístico, y es el de la naturaleza diferente de las imágenes a tratar. En ImageNet, las imágenes provienen en su mayoría de redes sociales como puede ser Flickr. Evidentemente este no és el mismo tipo de imágenes que encontraremos en archivos históricos. Resulta evidente entonces que con tal de realizar la prueba para ver si las arquitecturas de redes neuronales actuales pueden responder de igual manera frente a este nuevo escenario, se necesitaría una base de datos relevante y de gran volumen proveniente de los archivos para poder realizar nuevos entrenamientos *ad-hoc* a este dominio y poder evaluar su rendimiento.

Pero el punto que consideramos más crítico, y que seguramente requiera esfuerzos de investigación más allá del simple uso de arquitecturas ya conocidas sobre una nueva base de datos, recae sobre qué tipo de clases se definirían en el campo de los archivos históricos. Los metadatos que se asocian a los fondos fotográficos tienen que permitir dar acceso a los contenidos de estos fondos, es decir, tienen que describir los contenidos de las imágenes para posteriormente facilitar su búsqueda. Ahora bien, estos contenidos no tienen la misma naturaleza que los contenidos anotados en ImageNet. En ImageNet se anotan contenidos puramente visuales, como si en esta imágen aparece o no aparece un cierto objeto. En cambio, en el campo archivístico no sólo se anotan objetos, sino que muchas veces se anota el contenido de las imágenes teniendo en cuenta un cierto contexto histórico conocido. Es decir, no se anota que en una fotografía aparece una persona en una calle, sino que se anota que se trata de una fotografía donde aparece, por ejemplo, "Gaudí paseando por las Ramblas de Barcelona".

Dotar a las redes neuronales de este contexto histórico para que puedan producir este tipo de anotaciones de manera automática, no es una tarea simple. Aunque a día de hoy parece lejana la posibilidad de tener una herramienta automática que sea capaz de producir este tipo de anotaciones, el campo del *Deep Learning* ya ha virado hacia esta dirección y puede que empecemos a ver este tipo de resultados a corto plazo.

Agradecimientos

Este trabajo ha sido financiado en parte por el proyecto TIN2014-52072-P, el programa CERCA de la Generalitat de Catalunya, el proyecto aBSINTHE de la

Fundación BBVA y el programa H2020 Marie Skłodowska-Curie de acciones de la Unión Europea, grant agreement No 712949 (TECNIOspring PLUS) y la Agencia para la competitividad de la empresa del Gobierno de Catalunya (ACCIO). Agradecemos a NVIDIA Corporation su donación de la GPU Titan Xp, usada en el marco de esta investigación.

Bibliografía

CAO, Huaigu (2014) "Machine-Printed Character Recognition". *Handbook of Document Image Processing and Recognition*. Springer Verlag. p. 331-358.

DENG, Jia, DONG, Wei, SOCHER, Richard, LI-JIA, Li, KAI, Li, FEI-FEI, Li (2009). "ImageNet: A Large-Scale Hierarchical Database". *Conference on Computer Vision and Pattern Recognition*.

EVERINGHAM, Mark, ESLAMI, S. M. Ali , VAN GOOL, Luc, WILLIAMS, Christopher K. I., WINN, John, ZISSERMAN, Andrew (2015). "The PASCAL Visual Object Classes Challenge: A Retrospective". *International Journal of Computer Vision*. v. 111, n. 1, p. 98-136.

EVERINGHAM, Mark, VAN GOOL, Luc, WILLIAMS, Christopher K.I., WINN, John, ZISSERMAN, Andrew (2010). "The PASCAL Visual Object Classes (VOC) Challenge". *International Journal of Computer Vision*. v. 88, n. 2, p.303-338.

FORNÉS, Alicia, LLADÓS, Josep, RAMOS, Oriol, RUSIÑOL, Marçal (2016). "La Visió per Computador com a Eina per a la Interpretació Automàtica de Fonts Documentals." *Lligall, Revista Catalana d'Arxivística*. v. 39. p. 20-46.

FRINKEN, Volkmar, BUNKE, Horst (2014). "Continuous Handwritten Script Recognition". *Handbook of Document Image Processing and Recognition*. Springer Verlag. p. 391-425.

HE, Kaiming, ZHANG, Xiangyu, REN, Shaoqing, SUN, Jian (2015). "Deep Residual Learning for Image Recognition". ArXiV.

HERBERT. F., Schantz, (1982). "The history of OCR, optical character recognition." *Recognition Technologies Users Association*.

JUN, Yang, JIANG, Yu-Gang, HAUPTMAN, Alexander G., NGO, Chong-Wah (2007). "Evaluating bag-of-visual-words representations in scene classification." *International workshop on multimedia information retrieval*, p. 197-206.

KRIZHEVSKY, Alex, SUTSKEVER, Ilya, HINTON, Geoffrey E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks". *Advances in Neural Information Processing Systems*. p. 1097-1105.

PAPERT, Seymour (1966). "The Summer Vision Project". *MIT AI Memos*.

RUSSAKOVSKY, Olga, DENG, Jia, SU, Hao, KRAUSE, Jonathan, SATHEESH, Sanjeev, MA, Sean, HUANG, Zhiheng, KARPATHY, Andrej, KHOSLA, Aditya, BERNSTEIN, Michael, BERG, Alexander C., FEI-FEI, Li (2015). "ImageNet Large Scale Visual Recognition Challenge". *International Journal of Computer Vision*. v. 115, n. 3, p. 211-252.

SEBE, Nicu, COHEN, Ira, GARG, Ashutosh, S. HUANG, Thomas (2005). "Machine Learning in Computer Vision". *Springer Science & Business Media*.

SIMONYAN, Karen, ZISSERMAN, Andrew (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". ArXiV.

SZEGEDY, Christian, LIU, Wei, JIA, Yangqing, SERMANET, Pierre, REED, Scott, ANGUELOV, Dragomir, ERHAN, Dumitru, VANHOUCKE, Vincent, RABINOVICH, Andrew (2014) "Going Deeper with Convolutions". ArXiV.

SZEGEDY, Christian, VANHOUCKE, Vincent, IOFFE, Sergey, SHLENS, Jonathon, WOJNA, Zbigniew (2015). "Rethinking the Inception Architecture for Computer Vision". ArXiV.

SZELISKI, Richard (2010). "Computer Vision: Algorithms and Applications". *Springer Science & Business Media*.

Tractica Reports (2016) "Computer Vision Technologies and Markets".

TULYAKOV, Sergey, GOVINDARAJU, Venu (2014). "Handprinted Character and Word Recognition". *Handbook of Document Image Processing and Recognition*. Springer Verlag. p. 359-389.