SUBTITULACIÓ AUTOMÀTICA D'IMATGES. ESTAT DE L'ART I LIMITACIONS EN EL CONTEXT ARXIVÍSTIC

Lluis Gomez, Marçal Rusiñol, Ali Furkan Biten, Dimosthenis Karatzas
Centre de Visió per Computador, Universitat Autònoma de Barcelona

Introducció

L'ús de noves tecnologies per part de la comunitat arxivística és una realitat en gairebé tots els passos de la cadena documental, des de les tècniques de conservació i els sistemes de digitalització, fins a les eines informàtiques per l'anotació de metadades o les bases de dades de consulta que faciliten la difusió de continguts. Tot i així, el procés documental requereix encara una gran quantitat de treball manual, que pot arribar a convertir-se en un coll d'ampolla a l'hora de processar grans volums d'informació.

A Catalunya hi ha més de 300 arxius, un 78% dels quals són públics. El seu tipus inclou des d'arxius centrals administratius, històrics, comarcals, municipals o eclesiàstics entre d'altres. De tots els documents emmagatzemats i gestionats en aquests arxius, un 92,2% correspon a fotografies, que es comptaven en gairebé 34 milions d'imatges el 2015 (Dept. de Cultura 2015). Una dada que de ben segur s'ha vist incrementada a dia d'avui i que no pot fer més que créixer en els propers anys.

Aquest volum d'informació pot arribar a fer inviable, per motius de cost i temps, la tasca d'etiquetar i descriure acuradament cadascuna de les imatges amb les dades adequades. Com a conseqüència, part de la informació visual dels arxius estarà condemnada a romandre inaccessible als sistemes de cerca informàtica. En aquest sentit, un procés automàtic que pogués proposar descripcions per a fotografies històriques mitjançant l'anàlisi del seu contingut visual podria contribuir a facilitar-ne l'accés, en permetre'n la indexació de automàtica i fer-les accessibles als motors de cerca.

Als darrers anys, dins del camp de la intel·ligència artificial, ha emergit una nova línia de recerca coneguda com a subtitulació automàtica d'imatges. Combinant algoritmes dels camps de la lingüística computacional i de la visió per computador, s'han dissenyat models computacionals capaços de generar de manera automàtica frases semànticament correctes que proporcionen una descripció textual dels continguts visuals de la imatge tractada. Aquestes tècniques es podrien emprar per tractar automàticament quantitats immenses de fotografies i generar el peu de foto corresponent a cadascuna, sense intervenció humana. En aquesta comunicació presentem una introducció a l'estat de l'art en l'àmbit de la subtitulació automàtica d'imatges, exemplificant-ho amb diversos casos, i analitzant les limitacions que ens hem trobat a l'hora d'adaptar els mètodes actuals al domini de les fotografies històriques provinents d'arxius.

Estat de l'art

La recerca en mètodes de subtitulat automàtic d'imatges té una llarga història (Pan 2004), però no va ser fins el 2014, coincidint amb la publicació del conjunt de dades MS-

COCO (Lin 2014), que el camp es va revolucionar per complert amb l'aparició de nous sistemes que aprofitaven els models d'aprenentatge profund (LeCun 2015). Un d'aquests sistemes, que ha influït de forma notable en el desenvolupament de la gran majoria de mètodes posteriors, es va inspirar en el paradigma *"sequence-to-sequence"* que s'utilitza normalment en sistemes de traducció automàtica (Sutskever 2014). Tal i com il·lustra el diagrama en la Figura 1, els sistemes *"sequence-to-sequence"* (seq2seq) son aquells que converteixen seqüències d'un domini (per exemple, frases en català) a seqüències d'un altre domini (per exemple, les mateixes frases traduïdes a l'anglès) de forma automàtica.

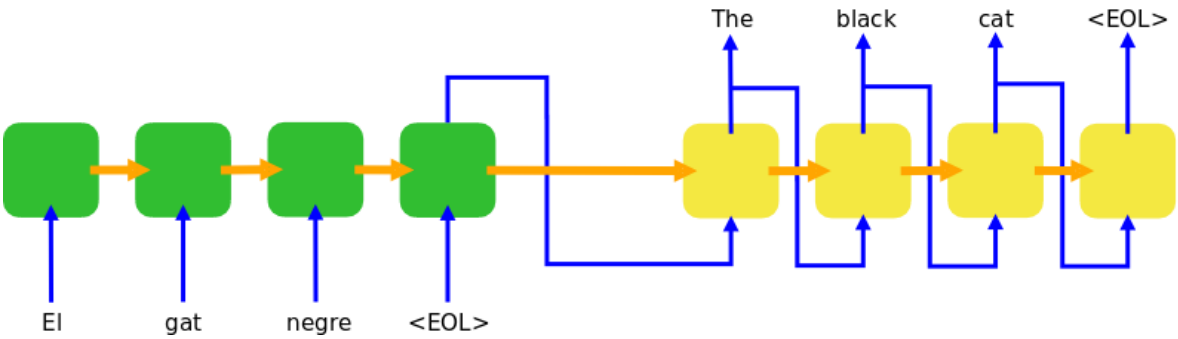

Figura 1. Els models seq2seq que s'utilitzen en traducció automàtica es composen de dues xarxes neuronals recurrents (RNN): el codificador (en verd) y el decodificador (en groc). Aquestes xarxes s'anomenen "recurrents" perquè s'apliquen repetidament, per a cada paraula de la seqüència.

En un sistema de traducció automàtica seq2seq hi trobem dues xarxes neuronals recurrents (RNN): el codificador, que s'encarrega de codificar la seqëncia d'entrada en informació numèrica, y el decodificador, que transforma aquesta informació numèrica en una seqüència de sortida. Aquestes xarxes s'anomenen "recurrents" per què el mateix model s'aplica de forma repetida per a codificar/decodificar cada una de les paraules de les quals es composen les seqüències d'entrada i sortida. La Figura 2 mostra un diagrama simplificat d'aquest model on les xarxes RNN es representen amb un sol bloc. En qualsevol cas les xarxes RNN no deixen de ser un tipus de model estadístic que en el camp de la intel·ligència artificial s'anomenen d'*aprenentatge supervisat*. L'aprenentatge supervisat és la tasca d'aprendre una funció que assigna una entrada a una sortida determinada a partir de veure molts exemples de parells entrada-sortida donats. Per exemple, a partir de una col·lecció de parells de frases en català i anglès (dades d'entrenament en que coneixem la traducció) l'algorisme d'aprenentatge supervisat analitza les dades i produeix una funció inferida, que es pot utilitzar per traduir noves frases del català a l'anglès. Les RNN s'utilitzen per tractar informació seqüencial (com ara el text) per dues raons: 1) permeten tractar dades amb mida variable (per exemple frases de diferents longituds); i 2) poden tenir memòria (no es tracta d'aplicar el model a cada paraula d'una frase de forma independent sino de condicionar la resposta final a totes les paraules en conjunt i en un ordre donat).

Figura 2. Diagrama simplificat d'un model seq2seq.

El sistema estàndard de subtitulació automàtica d'imatges, basat en un model seq2seq (Vinyals 2015), utilitza una xarxa neuronal convolucional (CNN) - de les que s'empren normalment per a tasques de classificació d'imatges - per codificar la imatge d'entrada, i un decodificador RNN que generarà la frase de sortida (subtitol).

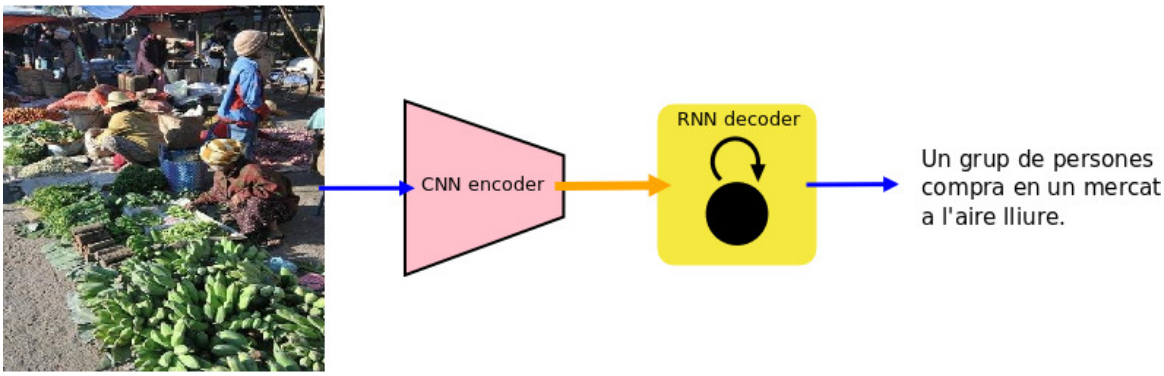

Figura 3. Sistema de subtitulació automàtica d'imatges basat en un model seq2seq.

Així doncs, un model *seq2seq* per a la subtitulació d'imatges que hagi estat entrenat amb una col·lecció de parells imatge-subtitol relativament gran serà capaç d'inferir subtítols per a noves imatges que no ha vist mai abans en el seu conjunt d'entrenament. Aquesta propietat del models estadístics s'anomena *generalització*, i és, al cap i a la fi, el que fa que aquests models tinguin utilitat real. Dit d'una altre manera, si un model de traducció automàtica no pot traduir frases diferents de les del conjunt d'entrenament serà un model inútil, doncs ja coneixíem prèviament les traduccions de les frases d'aquest conjunt. El que volem és que a partir d'un conjunt finit de frases d'exemple, el nostre model pugui fer traduccions del conjunt (infinit) de totes les frases que es poden crear amb un llenguatge determinat.

En general, per a que un sistema d'aprenentatge automàtic sigui capaç de generalitzar haurà de complir dues condicions: 1) tenir suficient complexitat de càlcul; i 2) ser entrenat amb una base de dades d'entrenament el més gran i diversa possible. Si bé la primera condició és fàcil d'acomplir avui dia amb els recents avenços en el desenvolupament de xips de càlcul numèric, la segona requereix normalment d'un esforç humà considerable per etiquetar manualment les dades d'entrenament. És per això, com ja s'ha mencionat anteriorment, que els sistemes de subtitulació automàtica no van ser una realitat fins que al 2014 es va fer públic el conjunt de dades d'entrenament MS-COCO.

El conjunt de dades de MS-COCO, amb més de 120.000 imatges i 650.000 subtítols creats de forma manual, va permetre l'entrenament de xarxes neuronals complexes per a la subtitulació automàtica d'imatges. Anteriorment, els conjunts de dades eren més

aviat de mida limitada, o contenien imatges amb subtítols ambigus i poc fiables recollits de forma semi-automàtica de xarxes socials com ara Flickr[1]. Per crear el conjunt de dades MS-COCO es va utilitzar el sistema Amazon Mechanical Turk (AMT), una plataforma de *crowdsourcing* on es realizen tasques (per exemple la subtitulació d'imatges) sobre conjunts de dades (imatges) a gran escala. Cada imatge fou lliurada a un treballador (també anomenat anotador) amb les següents instruccions:

- Descriu totes les parts importants de l'escena.
- No iniciïs les frases amb "Hi ha ...".
- No descriguis detalls sense importància.
- No descriguis les coses que podrien haver passat en el futur o el passat.
- No descriguis el que una persona podria dir o pensar.
- No donis noms a les persones.
- Les frases han de contenir almenys 8 paraules.

D'aquesta manera els autors van voler crear un conjunt de dades d'entrenament el més genèric i uniforme possible. En la Figura 4 es mostren alguns exemples d'anotacions creades en aquest procés de *crowdsourcing* manual. Cada imatge va ser anotada per 5 anotadors de mitjana per així obtenir més diversitat en els subtítols.

- *A woman holding a baby near a dog at a poolside.*
- *A women is holding a baby next to a dog, a patio table, and a pool.*
- *A lady holding an infant in her arms next to a pool with a dog in front of them.*

- *Woman taking a selfie with a giraffe in an enclosure.*
- *A young girl taking a photo of herself with a giraffe behind her.*
- *A woman takes a selfie picture with a giraffe in the background.*

- *A couple of people at the beach during a sunny day.*

1 http://www.flickr.com

- *Some people on a beach stare out into the ocean.* - *The view down waikiki beach towards the royal hawaiian.*	- *Teenage girls giggle and chat as they devour a large pepperoni pizza.* - *A group of women sitting around a table with a pizza on a pan.* - *Several girls at a table with a partly eaten pizza on a tray.*

Figura 4. Exemples d'imatges i subtítols del conjunt de dades d'entrenament MS-COCO.

En la Figura 4 val la pena destacar el cas concret de la imatge amb una escena de platja (imatge inferior esquerra): observem que un dels anotadors ha fet un subtitol molt més contextualitzat que la resta, localitzant el lloc on s'ha fet la fotografia amb molt de detall, mentre que la resta d'anotadors han creat subtítols molt més genèrics i neutres. D'aquesta diferència entre l'anotador "expert" i l'anotador "corrent" o "inexpert" en parlarem més endavant quan analitzem els resultats dels algorismes de subtitulació en imatges d'arxiu.

La Figura 5 mostra alguns dels resultats del model seq2seq de subtitulat automàtic d'imatges (Vinyals 2015) entrenat amb el conjunt de dades MS-COCO. És important destacar aquí que les imatges mostrades no han estat mai vistes abans per l'algorisme d'aprenentatge, i que, per tant, ens poden donar una idea de com de bé generalitza el model a noves imatges. De fet, hem escollit les imatges de tal manera que mostrin casos en els que el model generalitza força bé (imatges superiors), i casos on el model no és capaç de generalitzar (inferiors) i produeix un subtítol que no està relacionat amb la imatge d'entrada. En el cas del subtítol generat per la imatge del cotxe groc, podriem conjecturar sobre les possibles raons que afecten negativament al model: hi ha moltes imatges de autobusos escolars en el conjunt de dades d'entrenament? el model no ha vist imatges d'altres cotxes "disfressats"? etc. Si les imatges que volem subtitular no s'assemblen en res a les imatges del conjunt d'entrenament el model no serà capaç de produir bons subtítols. Es aquí on es manifesta la necessitat de tenir conjunts de dades molt grans i diversos.

A person riding a motorcycle on a dirt road

Two hockey players are fighting over the puck

A little girl in a pink hat is blowing bubbles

A yellow school bus parked in a parking lot

Figura 5. Exemples de subtítols generats automàticament[2] amb un model seq2seq.

A partir del model seq2seq (Vinyals 2015), altres treballs de recerca més recents han explorat diverses formes d'utilitzar els mecanismes d'atenció en les xarxes neuronals. Modelant d'alguna manera quines parts concretes de la imatge són importants per cada concepte o paraula del seu sbtítol. Altres treballs han explorat l'extensió de les descripcions o subtítols d'imatges a diferents granularitats o estils de text. Per exemple, hi ha hagut cert interès en la generació de narratives a partir d'imatges (Kiros 2015), en obtenir descripcions a nivell de paràgraf, o en la transferència d'estils (humorístics, romàntics, etc.) en els subtítols d'imatges (Mathews 2016, Gan 2017).

Aplicabilitat i limitacions en el context arxivistic

En la Figura 6 mostrem alguns resultats d'un model *seq2seq* de subtitulació automàtica d'imatges, entrenat amb el conjunt de dades MS-COCO, aplicat a fotografies provinents d'arxius històrics del portal d'internet Europeana[3].

A man and a woman are standing next to a

A dining room with a table and chairs.

2 Per una colecció més extensa d'exemples visiteu:
http://www.cs.toronto.edu/~nitish/nips2014demo/

3 http://www.europeana.eu

sign.	
	
A group of people standing around a white table.	*A store front with a bunch of signs on it.*

Figura 6. Exemples de subtítols generats automàticament amb un model seq2seq en imatges d'arxiu històric.

Observem que els subtítols generats si bé ofereixen una correcte descripció visual de la imatge, en cap cas són capaços d'oferir-ne una interpretació conceptual. No s'hi fa cap referència a l'època ni el lloc on van ser fetes les fotografies; tot i què aquestes dades es podrien arribar a inferir a partir de la informació visual. Per exemple, la descripció de la imatge superior dreta en la Figura 6 feta per un arxiver aporta un gran nombre de detalls que requereixen coneixement expert, no només visual, si no també del context històric del fons documental al que pertany la imatge: *"First Class Chinese Tea Room on the 'Ciudad de Buenos Aires' (1914)".* En canvi, els subtítols generats de forma automàtica ofereixen més aviat una descripció enumerativa, on es llisten els objectes que apareixen a la imatge i, fins a cert punt, les seves posicions relatives en l'espai. De fet, aquest comportament no ens ha de sorprendre, atès que el model ha estat entrenat amb el conjunt d'entrenament MS-COCO que hem descrit anteriorment i per tant només és capaç de generar subtítols genèrics. Amb tot i amb això, val a dir que els subtítols generats poden ser d'utilitat en alguns casos, depenent de les necessitats concretes de cada arxiu.

Ara bé, podem crear un model de subtitulació automàtica que generi subtítols amb els detalls que donaria un anotador "expert"? Evidentment, si el nostre model ha estat entrenat amb descripcions genèriques creades per anotadors "no experts", no podem esperar que sigui capaç d'identificar conceptes semàntics més específics de la imatge, com ara els noms de llocs, persones, etc. o el seu context històric. Els models actuals *seq2seq* no incorporen aquesta informació i només descriuen continguts visuals, sense contextualitzar-los en un instant històric. Per tant, si volem models capaços de crear subtítols amb aquest nivell de detall necessitarem crear nous models i conjunts d'entrenament que ens ho permetin.

Una base de dades de parells imatges-subtítols creada a partir d'imatges d'arxiu i amb unes normes d'anotació similars a les que utilitzen els arxivers ens permetria en principi entrenar models de subtitulació automàtica més rics i adequats a les necessitats reals dels arxius. Aquestes normes d'anotació varien en funció de les necessitats de cada centre, però seran en general molt diferents a les que es van utilitzar per a la base de dades MS-COCO. Per posar un exemple, algunes de les normes que es poden donar

als arxivers per a la descripció d'imatges en documents àudio-visuals poden ser (Bailac 1994):

- Utilitzar frases curtes i concises, però explicatives.
- Indicar al començament de cada nou pla el seu codi de descripció.
- Normalitzar els noms i sigles.
- Utilitzar sinònims d'un mateix terme per facilitar la recerca.
- Donar el nom complet dels personatges que apareixen, i el càrrec per qual intervenen en les imatges.

La bona notícia en aquest sentit és que ja existeix un gran nombre de dades que han estat anotades per arxivers seguint normes com aquestes, i amb les que es pot crear aquesta base de dades de subtitulació "experta". És doncs d'esperar que la col·laboració entre arxivers i experts en visió per computador ens porti en els propers anys a desenvolupar models de subtitulació automàtica que s'adaptin millor a les necessitats reals dels arxius.

Agraïments

Aquest treball ha estat finançat en part pel projecte TIN2014-52072-P, el programa CERCA de la Generalitat de Catalunya, el projecte aBSINTHE de la Fundació BBVA, el programa H2020 Marie Skłodowska-Curie d'accions de la Unió Europea, acord de subvenció No 712949 (TECNIOspring PLUS), i l'Agència per a la competitivitat de l'empresa del Govern de Catalunya (ACCIO). Agraïm a NVIDIA Corporation la seva donació de la GPU Titan Xp, que s'utilitza en el marc d'aquesta investigació.

Bibliografia

Bailac, Montserrat. Català, Montserrat. "Anàlisi de documents audiovisuals : la descripció de les imatges." 1994 - Imatge i Recerca : Jornades Antoni Varés (3es : 1994 : Girona)

Dept. de Cultura. "Estadístiques Culturals de Catalunya". Març 2017.

Gan, Chuang, et al. "Stylenet: Generating attractive visual captions with styles." Proc IEEE Conf on Computer Vision and Pattern Recognition. 2017.

Kiros, Ryan, et al. "Skip-thought vectors." Advances in neural information processing systems. 2015.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." nature 521.7553 (2015): 436.

Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context." European conference on computer vision. Springer, Cham, 2014.

Mathews, Alexander Patrick, Lexing Xie, and Xuming He. "SentiCap: Generating Image Descriptions with Sentiments." AAAI. 2016.

Pan, Jia-Yu, et al. "Automatic image captioning." Multimedia and Expo, 2004. ICME'04.

2004 IEEE International Conference on. Vol. 3. IEEE, 2004.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.