

Descripció automàtica d'arxius audiovisuals: NeuralTalk, un model de video-to-text aplicat a l'arxiu de RTVE

[Versión castellana]

 VIRGINIA BAZÁN GIL
Responsable de projectes
Fons Documental de RTVE
virginia.bazan@rtve.es
RICARDO GUERRERO GÓMEZ-OLMEDO
Científic de dades
BEEVA
ricardo.guerrero@beeva.com

Opcions

<meta />

Metadades

DOI: <https://dx.doi.org/10.1344/BiD2018.41.6>

Citació recomanada

Bazán Gil, Virginia; Guerrero Gómez-Olmedo, Ricardo (2018). "Descripció automàtica d'arxius audiovisuals : NeuralTalk, un model de video-to-text aplicat a l'arxiu de RTVE". *BiD: textos universitaris de biblioteconomia i documentació*, núm. 41 (desembre) . <<http://bid.ub.edu/41/bazan.htm>>. DOI: <http://dx.doi.org/10.1344/BiD2018.41.6> [Consulta: 20-12-2018].

Resum

Objectiu: determinar la maduresa dels sistemes de *video-to-text* per a la descripció automàtica d'imatges en un arxiu de televisió.

Metodologia: es fa una prova de concepte mitjançant un sistema de *video-to-text* desenvolupat ad hoc. La prova es va articular en tres fases o iteracions diferents entre juny de 2016 i gener de 2017. En les dues primeres iteracions el sistema va analitzar un nombre determinat de continguts procedents de l'arxiu de RTVE, les descripcions es van valorar per establir la taxa d'encert del sistema o, en altres paraules, com de propera era aquesta descripció a la que podia haver subministrat un ésser humà. En una tercera fase, i prèviament a l'anàlisi dels continguts, es va entrenar el sistema utilitzant tècniques d'aprenentatge profund amb l'objectiu de millorar els resultats.

Resultats: els resultats obtinguts posen de manifest que es tracta d'una tecnologia prometedora, si bé resulta fonamental aprofundir més en els mecanismes que serien necessaris per a la seva posada en producció en els arxius de televisió.

Resumen

Objetivo: determinar la madurez de los sistemas de *video-to-text* para la descripción automática de imágenes en un archivo de televisión.

Metodología: se realiza una prueba de concepto mediante un sistema de *video-to-text* desarrollado *ad hoc*. La prueba se articuló en tres fases o iteraciones distintas entre junio de 2016 y enero de 2017. En las dos primeras iteraciones el sistema analizó un número determinado de contenidos procedentes del archivo de RTVE, las descripciones se valoraron para establecer la tasa de acierto del sistema o, en otras palabras, cómo de cercana era dicha descripción a la que podía haber suministrado un ser humano. En una tercera fase, y previamente al análisis de los contenidos, se entrenó al sistema utilizando técnicas de aprendizaje profundo con el objetivo de mejorar los resultados.

Resultados: los resultados obtenidos ponen de manifiesto que se trata de una tecnología prometedora, si bien resulta fundamental profundizar más en los mecanismos que serían necesarios para su puesta en producción en los archivos de televisión.

Abstract

Objective: To assess the deep learning capability of a video captioning model for automated image description in a television archive.

Methodology: Our proof of concept tested an ad hoc video-captioning model in three iterations between June 2016 and January 2017. In the first and second iterations the model was used to analyse a selection of content from the archives of the Spanish Radio and Television Corporation (RTVE) and the descriptions it generated were evaluated to determine the model's success rate, i.e., how close it came to providing human-like image descriptions. In the third iteration and before the content was analysed, the model was trained using deep learning techniques to optimise the results.

Results: The results indicate that the model has potential, although further development will be required to customise its use in television archives.

Paraules clau: Arxius de mitjans de comunicació, Arxius audiovisuals, Radiotelevisión Española, Programes informàtics, Anàlisi de contingut, Intel·ligència artificial, Documents audiovisuals

Keywords: Mass media archives, Audiovisual archives, Radiotelevisión Española, Computer programs, Subject cataloguing, Artificial intelligence, Audiovisual documents

Rebut: 19/04/2018. Acceptat: 28/06/2018.

1 Introducció

Des de la dècada dels anys vuitanta del segle xx els arxius audiovisuals han experimentat una evolució tecnològica constant que afecta no només els formats en què els continguts s'emmagatzemen, sinó també a la pròpia essència. Els arxius han deixat de ser "sitges protegides" per convertir-se en un espai per a la interacció amb la producció i els usuaris (Teruggi, 2018). La digitalització no només ha diluït la frontera entre la producció i l'arxiu, sinó que també ha provocat la transformació tant dels fluxos de treball com dels perfils professionals, qüestions que han estat profundament analitzades (Aguilar; López de Solís, 2010; Giménez Rayo, 2012; Caldera Serrano; Arranz, 2013; Giménez Rayo; Guallar, 2014). Ahora, i com a conseqüència de la multiplicació dels canals de distribució i d'emissió, s'està produint un augment exponencial dels continguts audiovisuals i, per tant, un augment proporcional de la necessitat de metadades per garantir-ne la recuperació i preservació digital (López de Quintana, 2014; Sound and Vision *et al.*, 2017; Pastor; Escribano, 2018).

La sobreabundància de continguts empeny els arxius a la incorporació de tecnologies que converteixen l'etiquetatge en temps real en una solució viable (Teruggi, 2018). Aquest interès es posa de manifest a la *MAM Survey*, una enquesta que la Federació Internacional d'Arxius de Televisió (FIAT/IFTA) distribueix anualment entre els seus membres, a fi de conèixer-ne les expectatives i el grau de familiaritat pel que fa a tecnologies com les de la parla o la descripció automàtica d'imatges. Dels resultats de l'última *MAM Survey*, presentada a Lugano el juny de 2017 (FIAT/IFTA, 2017), es desprèn que tant la descripció d'escenes (dia-nit, interior-exterior, etc.) com el reconeixement d'imatges (logos, edificis, cares) es consideren eines potencialment útils per al 69 % dels arxius, si bé només tenen experiència en la seva aplicació un percentatge molt petit d'organitzacions, com ara la finlandesa YLE (Selkälä, 2017), la italiana RAI (Pandolfi; Desirello, 2017), l'holandesa Sound and Vision (Manders, 2018) o l'espanyola RTVE (Bazán *et al.*, 2018).

Si ens centrem en els arxius que han tingut algun contacte amb aquest tipus d'eines, comprovem que només el 2 % ha incorporat el reconeixement d'escenes a la feina diària, el 6 % pretén incorporar-lo a mitjà o curt termini i el 9 % considera que no està prou madur com per implementar-lo en producció. El reconeixement d'objectes (logos, edificis, cares) no s'empra de moment en cap dels arxius que van participar en l'enquesta, si bé el 4 % dels que l'han provat el vol incorporar als seus processos en el futur, davant d'un 15 % que considera que no és una tecnologia prou madura.

Com és lògic, els proveïdors tecnològics no són aliens als interessos del sector dels mitjans de comunicació. Actualment hi ha al mercat diferents eines que incorporen moltes de les tecnologies abans esmentades. Algunes d'aquestes solucions estan desenvolupades específicament per a l'entorn d'emissió de continguts, com és el cas de l'Etiquemida (2018) o de VSNExplorer (VSN, 2018). La primera és una eina específicament orientada a la catalogació automatitzada supervisada, mentre que VSNExplorer és un sistema de gestió i arxiu de continguts audiovisuals (*MAM—media asset management—*) capaç d'integrar les metadades generades per motors d'indexació com Azure de Microsoft (2018), Watson d'IBM (2018) o Google Cloud Video Intelligence (Google, 2018). Aquests són només alguns exemples orientats als arxius de televisió. No obstant això, el creixement exponencial de continguts audiovisuals, a què abans fèiem referència, ha fet que grans companyies com IBM, Microsoft o Google hagin entrat amb pas ferm de ple en aquest negoci, aportant la seva experiència al sector dels mitjans de comunicació i, fins i tot, establint aliances estratègiques amb organismes de radiodifusió, com és el cas de Google, que participa juntament amb Telefónica, VSN i Watchity en el Journalism Innovation Hub de RTVE, un projecte que pretén explorar noves formes de comunicació innovant en tot el procés de producció i en el qual Google aporta la seva tecnologia en núvol (RTVE, 2018; 24h, 2018).

Des del punt de vista tecnològic els tres pilars sobre els quals se sustenten aquestes eines són: visió artificial, tecnologies de la parla i processament del llenguatge natural (Lleida, 2018). Vegem-los amb més deteniment.

La visió artificial proporciona el reconeixement d'imatges, cares, objectes, edificis i logos. Permet l'agrupació d'imatges per escenes i la seva segmentació, alhora que facilita fer un seguiment de les persones i els objectes que apareixen al vídeo. Per la seva banda, les tecnologies de la parla s'ocupen de la informació que procedeix de l'àudio. El reconeixement automàtic de la parla (*automatic speech recognition*) és el primer pas per a la transcripció de la locució a text (*speech to text*), mentre que la diarització de parlants (*speaker diarization*) permet la identificació i indexació de les persones que parlen i dels seus discursos, l'estructuració del contingut dels documents, com també l'anàlisi de sentiments per a la descripció enriquida de continguts (Anguera *et al.*, 2012). Finalment, el processament del llenguatge natural, i més concretament la semàntica distribucional, és la tecnologia que permet unificar imatge, text i àudio de manera que sigui possible la detecció d'entitats, la descripció d'imatges en llenguatge natural, la generació de resums i la detecció i segmentació per temes.

Per tant, la intel·ligència artificial té un camp d'aplicació important en els arxius audiovisuals, ja sigui de forma continuïsta, és a dir, mantenint els mateixos processos, però ajudats per algorismes que permetin automatitzar la feina, ja sigui de forma disruptiva, buscant dur a terme treballs que s'havien descartat per no ser rendibles (Carrillo; González, 2018). En el model continuïsta, els conjunts de dades, és a dir, els arxius, són fonamentals, així com el paper dels documentalistes que validen i corregeixen els algorismes. L'enfocament rupturista, per la seva banda, permet, a partir de l'experiència adquirida, construir models que abans no es preveïen ja que tenen en compte els interessos o gustos dels usuaris, independentment de si coincideixen amb l'interès general o no. D'aquesta manera es rendibilitzen continguts que abans no es consideraven econòmicament rendibles. Segui quin sigui l'enfocament, l'aprenentatge automàtic (*machine learning*) és una peça clau, ja que permet, utilitzant grans conjunts de dades, que els algorismes aprenguin de manera autònoma i siguin capaços de predir el futur sobre la base d'experiències passades, simulant així l'aprenentatge del cervell (Carrillo; González, 2018; García, 2018).

Des del punt de vista de la intel·ligència artificial, la descripció automàtica d'imatges es considera un problema fonamental que connecta la visió artificial amb el processament del llenguatge natural. En els últims anys, la comunitat investigadora s'ha centrat en la classificació d'imatges (Guerrero, 2017) o en el reconeixement d'objectes, però, generar descripcions automàtiques d'imatges, emprant frases ben formades amb la sintaxi i ortografia adequada, és una tasca molt més complexa. La descripció no només es refereix a un objecte contingut en una imatge, sinó que ha d'expressar, també, com es relacionen els objectes entre ells, com també els seus atributs i les activitats en què estan implicats. Aquest coneixement semàntic s'ha d'expressar en un idioma determinat, generalment l'anglès, i per aquest motiu és necessari un model lingüístic a més de comprensió visual (Vinyals *et al.*, 2015; Karpathy; Fei-Fei, 2016).

Una de les primeres aproximacions existents per resoldre aquest repte ha estat NeuralTalk, un projecte dut a terme pel científic Andrej Karpathy (2015) en un intent d'aplicar els últims avenços de recerca en aprenentatge profund (*deep learning*) i especialment en xarxes neuronals multimodals, amb l'objectiu de desenvolupar un prototip funcional i obert a tot el públic. Aquest projecte consisteix en un programari que rep com a entrada una imatge i que a la sortida dona una descripció en llenguatge natural del seu contingut, de manera similar a la que podria haver fet una persona que veiés la mateixa imatge. Alliberar el codi i els models entrenats per a aquesta tasca ha donat lloc que artistes i enginyers de tot el món, partint d'aquest treball, creessin altres projectes o vídeos demostratius dels avenços de la tecnologia (McDonald, 2015). Després de l'impacte generat per NeuralTalk, per tant, han anat sorgint altres projectes que el continuen i milloren com ara NeuralTalk2 (Karpathy, 2016) o Show and Tell (Vinyals *et al.*, 2015; Shalloe, 2016). Des del punt de vista de la recerca, el nombre de publicacions relacionades amb aquesta temàtica ha tingut un creixement exponencial (Vinyals; Toshev; Bengio, 2016) i aquesta tecnologia s'ha adaptat fins i tot a altres camps més diversos, com el de les imatges mèdiques (Wang, 2018).

2 Metodologia

En aquest context BEEVA i el Fons Documental de RTVE van acordar dur a terme una prova de concepte (POC), entenent com a tal la implementació resumida d'un sistema de *video-to-text* per valorar la viabilitat d'aquesta tecnologia com una solució per a la descripció automàtica de bruts en un entorn de producció. La prova de concepte es va articular en tres iteracions o fases diferenciades que es van dur a terme entre els mesos de juliol de 2016 i gener de 2017.

La primera iteració es va fer el juliol de 2016 sobre un total de deu continguts (entre rotatges originals i peces del *Telediario*) sobre el Saló del Manga de Barcelona. La temàtica escollida va respondre al propòsit d'establir una certa connexió entre aquesta prova de concepte i l'experiència de catalogació automatitzada que es duia a terme amb altres empreses tecnològiques en el context de l'Observatori per a la Innovació dels Informatius a la Societat de la Informació Digital de RTVE.

Després d'analitzar els resultats obtinguts en aquesta primera fase, es va decidir continuar treballant amb els rotatges originals generats pels programes diaris no informatius que integren la redacció digital de Prado del Rey: *La mañana*, *Corazón*, *España Directo*, *Aquí la Tierra* i *Aquí hay trabajo*. Aquests bruts constitueixen la base de l'arxiu de producció, estan metadats de forma exhaustiva i no presenten cap limitació des del punt de vista dels drets d'explotació. La segona iteració es va desenvolupar entre els mesos d'agost i setembre de 2016 i es van processar vint-i-dos rotatges amb els corresponents minuts.

Durant la tercera i última iteració, feta entre octubre de 2016 i gener de 2017, es van processar tres-cents seixanta rotatges originals i els minuts respectius. Amb l'objectiu de millorar els resultats en la fase de processament, es va considerar rellevant establir una tipologia temàtica bàsica del material produït per cada programa per entrenar el sistema amb una mostra prou representativa de cada tipus (quant a quantitat i temàtica) i, finalment, poder determinar la utilitat de la descripció automàtica d'imatges per a cadascuna de les temàtiques identificades.

Tots els rotatges analitzats s'havien fet entre gener de 2015 i juliol de 2016. La taula 1 mostra la distribució per temes, la durada mitjana i el tipus de material característic de cada programa.

Programa	Durada mitjana	Tipus de material	Temàtiques
<i>Aquí hay trabajo</i>	50 minuts	Entrevistes Recursos	Activitat laboral Oficines Indústria Comerços Laboratoris Treballs agrícoles Ramaderia Aules de formació
<i>La mañana</i>	30 minuts	Entrevistes Recursos	Proves mèdiques Assajos clínics Laboratoris Gimnasos
<i>Corazón</i>	30 minuts	Entrevistes /declaracions <i>Photocall</i>	Persones Esdeveniments
<i>España Directo</i>	41 minuts	Recursos	Gastronomia Paratges naturals Urbanisme
<i>Aquí la Tierra</i>	15 minuts	Recursos	Paratges naturals Camps de cultiu
<i>Comando actualidad</i>	30 minuts	Recursos	Paratges naturals Urbanisme Gastronomia
Informatius	12 minuts	Recursos	Societat

Taula 1. Distribució de temàtiques per programes, durada mitjana, tipus de material i temàtiques incloses en els rotatges analitzats

Una vegada establerta la temàtica del material produït per cada programa, es va determinar que la descripció automàtica era potencialment útil per als rotatges d'*Aquí hay trabajo*, *La mañana*, *España Directo* i *Aquí la Tierra*, en els quals predominaven els plans de recurs tant en interiors com en exteriors i en què les declaracions eren menys rellevants. Es van descartar els rotatges del programa *Corazón*, la forma habitual del qual

sol ser l'entrevista i en què és especialment important tant la identificació del personatge entrevistat com el contingut de les seves declaracions. Per a aquest tipus de rodages es consideren potencialment útils les eines de reconeixement de la parla i de reconeixement facial.

Es va partir d'una mostra inicial de mil cinc-cents vint-i-quatre continguts, entre els quals es va començar seleccionant els que incloïen un major nombre de descripcions útils per entrenar el model. S'entén per *útils* frases curtes i ajustades a la descripció de la imatge sense elements inferibles a partir del context del programa o de l'àudio.

Finalment, i per completar la mostra, es van seleccionar bruts procedents de la Redacció Digital d'Informatius (Torrespaña), rodages d'entre febrer i octubre de 2016, arxivats a la categoria de "societat", realitzats majoritàriament en exteriors, que contenien fonamentalment paratges naturals i ciutats, els bruts dels quals tenien una durada mitjana de dotze minuts.

Des del punt de vista tècnic, en les dues primeres iteracions es va utilitzar únicament el material visual. Es va mesurar si la descripció proporcionada pel model era una descripció plausible que podria haver donat un ésser humà.



Figura 1. Descripcions subministrades pel sistema en la primera iteració

En la tercera iteració el procediment va ser diferent, ja que el que es buscava era entrenar un model nou usant els materials generats per RTVE.

Es va realitzar una adaptació dels minutatges a les imatges per obtenir una descripció potencialment útil. Habitualment els minutatges solen agrupar plànols de temàtica similar, de manera que l'equip de desenvolupament va considerar necessari extreure la primera imatge de cada interval inclòs en el minutatge, després de comprovar experimentalment que es tracta de la imatge que té una correlació més gran amb aquesta descripció.

A més, i de forma prèvia al processament de les imatges, va ser necessari preprocessar els continguts, eliminant tecnicismes habitualment presents en els minutatges dels continguts, com ara les abreviatures emprades per descriure els tipus de plans (PG, PM, PP) o expressions recurrents com "recursos de". La figura 2 mostra un minutatge un cop eliminats els tecnicismes.



Figura 2. Contingut preprocessat per a l'eliminació de tecnicismes

Durant el preprocessament es va decidir també identificar i substituir els noms propis més habituals pels noms comuns "home" o "dona", ja que el sistema ha de ser capaç d'identificar homes i dones, però no qui són aquestes persones a títol individual. Per dur a terme aquest procés es va prendre com a referència la llista de noms propis més freqüents publicada per l'Institut Nacional d'Estadística. Aquesta substitució afectava directament els noms d'organismes, institucions o llocs que incloïen en la seva denominació un nom propi, tal com mostra la figura 3, si bé l'equip de desenvolupament va considerar que tenia poca incidència en la taxa d'encert.



Figura 3. Substitució de noms propis per noms comuns durant el preprocessament dels continguts

Durant aquesta fase es van detectar alguns errors ortogràfics en els textos. Després d'avaluar solucions, tipus Aspell, que substitueixen la paraula errònia per la paraula correcta més probable i atesa la necessitat de procedir després a una 00000 per context, es va decidir eliminar les paraules errònies del conjunt de dades per a l'entrenament.

Finalment, i atès que el model que s'estava emprant per desenvolupar el sistema de *video-to-text* era natiu en anglès, va ser necessari traduir tots els minutatges preprocessats. Per a aquesta tasca es va optar pel traductor de Yandex via API per ser gratuït i fàcil d'utilitzar. Tot el procés

descriu es mostra a la figura 4.



Figura 4. Procediment emprat per entrenar el model final amb el material cedit per RTVE

Així doncs, partint del minutatge de la taula 2 com a exemple demostratiu, es va obtenir el text preprocessat de la taula 3, que posteriorment va ser introduït en un servei en línia de traducció automàtica de textos i es va obtenir la cadena de text de la taula 4, que es va usar després per alimentar el programari NeuralTalk

00:08:38–00:08:39: PL DE PAULA GONZÁLEZ REALIZANDO DECLARACIONES SOBRE CUÁL ES EL PRODUCTO QUE MÁS VENDE Y CUÁL ES SU SECRETO PARA EL ÉXITO

Taula 2. Minutatge original

mujer gonzález realizando declaraciones sobre cuál es el producto que más vende y cuál es su secreto para el éxito

Taula 3. Minutatge després del preprocessament

woman gonzalez making statements about which is the product most sold, and what is your secret to success

Taula 4. Minutatge traduït

Un cop adaptades les dades als formats requerits per NeuralTalk, es va fer el procés anomenat *entrenament*, consistent a generar un model matemàtic capaç d'aprendre i aparellar característiques visuals de les imatges amb característiques del text. Per a això es va utilitzar una funció de cost, és a dir, una mètrica que indica com de bé està aprenent el model. D'aquesta manera, un cop assolit un mínim satisfactori, es pot obtenir la versió final del model de descripció.

Finalitzada la fase d'entrenament, podem començar la fase de *test* o *predicció*, consistent a utilitzar el model prèviament calculat per usar-lo sobre noves imatges i així obtenir-ne descripcions al més semblants a com ho faria un ésser humà.

3 Resultats

Els primers resultats obtinguts per a la descripció dels continguts generats durant la cobertura del 22 Saló del Manga de Barcelona llançaven una taxa d'encert propera al 21 %. És important tenir en compte que, ja en el conjunt de dades de partida, no totes les descripcions proporcionades eren útils: aproximadament el 30 % contenien informació que no es podia inferir a partir de la informació visual continguda únicament en les imatges. Es van identificar errors repetitius derivats de les mancances del model utilitzat, concretament els micròfons i les pissarres es van confondre amb telèfons mòbils i portàtils o tauletes, tal com mostra la figura 5.

	
<i>a woman holding a cell phone in her hand</i>	<i>a woman sitting at a table with a laptop</i>

Figura 5. Errors recurrents derivats de la manca d'exemples en el model usat

En la segona iteració es van obtenir resultats dispars, amb una taxa d'encert d'entre el 0 % i el 23 %. Un examen detallat dels resultats demostra que les descripcions menys útils es produeixen en escenes de paisatges en què no hi ha ni objectes ni presència humana, tal com mostra la figura 6. En canvi, les millors descripcions s'obtenen en escenes en què hi ha grups de persones interactuant, tal com mostra la figura 7.

		
<i>a view of a large clock tower in the sky</i>	<i>a large body of water with a bridge in the background</i>	<i>a close up of an orange and a banana</i>

Figura 6. Mostra de els resultats negatius obtinguts en la segona iteració de la prova de concepte

		
<i>a man in a suit and tie standing in front of a store</i>	<i>a group of people sitting on top of a hill</i>	<i>a man is standing in a field with a mountain in the background</i>

Figura 7. Mostra dels resultats positius obtinguts en la segona iteració de la prova de concepte

En la tercera i última iteració els resultats no només no van millorar, sinó que sorprenentment, i de forma recurrent, el sistema va ser incapaç de descriure correctament qualsevol imatge, i la taxa d'encert va caure a percentatges propers al 0 %. Una mostra dels resultats es pot veure a la figura 8.

		
---	---	--

<i>statements of man UNK director of the UNK</i>	<i>UNK of UNK</i>	<i>statements of man UNK director of the UNK of the UNK</i>
--	-------------------	---

Figura 8. Mostra dels resultats de la tercera iteració

L'equip de desenvolupament va identificar tres factors per a aquest descens tan brusc en la taxa d'encert:

1. **Imatges fallides:** són aquells parells (imatge més descripció) que no es corresponen per no estar correctament sincronitzats. És a dir, la primera imatge de l'interval definit al minutatge no té cap correlació amb la descripció de l'interval, tal com mostra la figura 9.

	
<i>Pl recurso de barbero atendiendo a chico joven con barba. El barbero aconseja sobre el tratamiento y comienza el tratamiento</i>	<i>Pls de la tienda. Paula en la tienda oliendo un recipiente. Trabajadores colocando jabones</i>

Figura 9. Exemple d'imatge fallida per falta de sincronia

2. **Informació addicional:** sovint, els documentalistes fan servir en la descripció dels continguts audiovisuals informació que no s'infereix directament de la imatge, per exemple, els noms propis de les persones que apareixen a la imatge. Les descripcions recollides en la figura 10 en són un bon exemple.

	
<i>Recurso de Angelines una cerda que fue abandonada. Esta cerda tiene muchos fans en redes sociales y se han hecho camisetas y tazas con su imagen</i>	<i>Recursos ambiente en la localidad de Pioz (Guadalajara) y la urbanización "La Arboleda", días después de que la Guardia Civil hallara muertos a los cuatro miembros de una familia brasileña en el interior de un chalet</i>

Figura 10. Exemple de l'ús d'informació addicional en la descripció d'imatges

3. **Paraules fora del vocabulari:** es consideren paraules fora del vocabulari les que apareixen menys de cinc vegades en el conjunt d'entrenament. Durant el processament el sistema substitueix aquestes paraules desconegudes per l'ocurrència UNK (*unknown*). L'equip de desenvolupament va detectar que més del 70 % del vocabulari dels minuts no superava aquest llindar, i es va identificar com a factors rellevants la presència de noms propis, abreviatures i errades als textos.

4 Limitacions

Les limitacions principals d'aquesta prova de concepte estan relacionades amb les dades utilitzades per a l'entrenament del model i molt especialment amb els minuts disponibles a l'arxiu. Aquestes descripcions cronològiques són adequades per gestionar els continguts en l'entorn de producció de RTVE, però han demostrat no ser útils per entrenar un model de *video-to-text*. Com a factors que fan que aquestes descripcions no siguin adequades per a aquest propòsit podem identificar els següents:

- No són descripcions fotograma a fotograma i tendeixen a agrupar plans o escenes similars.

- Incorporen informació que no es dedueix únicament de la imatge, generalment procedent de l'àudio.
- En entrevistes o declaracions, l'àudio és l'element protagonista i la descripció de la imatge es converteix en un element secundari.
- El context de la imatge, entès com el programa o reportatge concret per al qual s'ha fet un rodatge, és fonamental per dotar de significat la descripció dels plans i aportar detalls significatius a aquesta descripció.

Vegem alguns exemples que il·lustren les limitacions dels minutatges empleats per a l'entrenament del model. La figura 11 representa un grup de persones assegudes en un banc i així ho ha descrit el sistema, però, la descripció manual feta a l'arxiu incloïa aquest fotograma concret en un agrupament més gran d'imatges que corresponia a persones caminant per un carrer de vianants.



Figura 11. Comparativa de descripció obtinguda pel sistema i descripció manual amb agrupació de plans

La figura 12 recull el moment en què un grup de models es disposa a desfil·lar per una passarel·la. L'escena és descrita pel sistema de manera correcta com un grup de persones en una sala, però el sistema no és capaç d'inferir que es tracta d'un grup de models perquè no té informació de context.



Figura 12. Comparativa de la descripció obtinguda pel sistema sense informació de context i descripció manual

De nou, la figura 13 és un clar exemple de la informació addicional que contenen les descripcions fetes a l'arxiu. En el primer cas el sistema és capaç d'identificar un grup de persones caminant pel carrer, però no pot inferir que és un carrer de Guadalajara; la segona imatge correspon a un televisor, però no és possible determinar, mitjançant un automatisme, que aquest televisor és d'una marca concreta i que es troba ubicat en un hipermercat.

	
<i>a group of people walking down a street</i>	<i>a picture of a tv on a table</i>
02:17:48–02:19:58: Pl recurs de trànsit en un carrer de Guadalajara	00:09:02–00:11:07: Pls recurs hipermercat, secció de tèxtil i de llibreria

Figura 13. Comparativa de descripció automàtica i descripció manual amb informació que no es pugui inferir directament de la imatge

La figura 14 és un exemple encara més evident. La descripció manual no només és capaç d'identificar les persones que hi intervenen, amb noms i cognoms, sinó que molt probablement inclou un resum del contingut de les seves paraules. En aquest sentit, els sistemes de reconeixement facial i de reconeixement d'imatges haurien de complementar-se amb informació obtinguda a partir de solucions de *speech-to-text*.

	
<i>a group of people standing around a building</i>	<i>a man in a suit and tie standing in front of a group of people</i>
00:00:20–00:01:01: DP dels assistents al Saló del Manga fent cua per entrar al recinte de La Farga. Alguns d'ells disfressats de personatges de còmic	00:06:51–00:07:42: es repeteix part de l'entrevista a Carles Santamaria en català

Figura 14. Comparativa de descripció automàtica i descripció manual amb informació relativa a l'àudio

El context, entès com el programa o reportatge per al qual es roden les imatges, és fonamental per dotar les escenes d'un significat complet. La figura 15 mostra una descripció automàtica més o menys correcta de la qual no es pot deduir, excepte per la informació de context, que es tracta d'un paratge determinat. En concret, aquestes imatges van ser rodades per a un reportatge sobre senderisme a Peñalara per al programa *España Directo* emès el 16 de juny de 2016 per La 1 de TVE.

	
<i>a man standing in a field with a mountain in the background</i>	<i>a group of people sitting on top of a hill</i>
00:16:42–00:29:50: PLS muntanyencs al Parc Natural de Peñalara	00:40:13–00:51:21: llacuna de Peñalara

Figura 15. Comparativa de descripció automàtica i descripció manual. Exemple d'imatge que incorpora informació de context obtinguda del programa

Finalment, s'ha de destacar que l'ambigüitat del llenguatge natural és un repte important per a la intel·ligència artificial. Aquesta prova de concepte ens ha demostrat la dificultat d'associar paraules polisèmiques el significat de les quals no és unívoc a una descripció visual determinada. Fixem-nos en l'exemple de minutatge següent: "entrevista a Neus, una jove disfressada que va al Saló del Manga de Barcelona". Si durant la fase de preprocessat el nom propi "Neus" no s'hagués substituït pel nom comú "dona", s'hagués associat la imatge d'una dona al concepte *neu* al model d'entrenament. Això hagués generat un conflicte en aplicar el mateix nom a la descripció de paisatges amb neu. Per aquest motiu les primeres aplicacions de la tecnologia *video-to-text* utilitzen models especialitzats per a cada temàtica, de manera que es limiten les distorsions i biaixos i es millora la precisió obtinguda enfront de models generalistes.

5 Conclusions

Els resultats obtinguts en aquesta prova de concepte són prometedors i demostren que la tecnologia ha avançat prou com perquè, en un futur proper, sigui viable integrar un sistema de *video-to-text* en un entorn de producció en un arxiu de televisió, sempre amb supervisió humana. Una estimació basada en el nombre de publicacions sobre descripció automàtica d'imatges en congressos científics (Computer Vision and Pattern Recognition, International Conference on Computer Vision, European Conference on Computer Vision, etc.) ens fa suposar que, en un termini aproximat de tres anys, tindrem sistemes capaços de descriure automàticament de manera més fiable per a col·leccions temàtiques acotades. D'aquesta manera, podríem emprar diferents models adaptats a les temàtiques més comunes: un model especialitzat en esdeveniments multitudinaris com el Saló del Manga, un altre model especialitzat en paisatges a l'aire lliure per a continguts similars als produïts pel programa *Aquí la Tierra*, etc.

Els resultats obtinguts posen de manifest que les solucions de *video-to-text* no són adequades per a continguts audiovisuals en què la informació rellevant es concentra en l'àudio, com és el cas d'entrevistes i declaracions i en què la identificació del personatge és un factor clau, si bé podrien aplicar-se com a tecnologia complementària a sistemes *speech-to-text* o de reconeixement facial per enriquir els resultats.

Finalment, podem concloure que aquestes solucions estan condicionades per dos factors fonamentals: la qualitat de les dades que es fan servir en la fase d'entrenament i l'ambigüitat del llenguatge natural. Aquest segon factor reforça la idea de la necessitat de construir models especialitzats per temàtiques concretes que permetin eliminar distorsions i biaixos en el model, per millorar la precisió en les descripcions.

Bibliografia

- 24h (2018). *Presentación del Journalism Innovation HUB de @rtve en el @BIT_audiovisual* [vídeo]. <https://www.periscope.tv/24h_tve/1dRKZeZBAjoxB>. [Consulta: 08/06/2018].
- Agirreazaldegi, T. (2007). "Claves y retos de la documentación digital en televisión". *El profesional de la información*, v. 16, n.o 5, p. 433–442. <<http://dx.doi.org/10.3145/epi.2007.sep.05> (<http://dx.doi.org/10.3145/epi.2007.sep.05>) >. [Consulta: 10/06/2018].
- Aguilar, M.; López de Solís, I. (2010). "Nuevos modos de trabajo de una redacción digital integrada: el caso de los servicios informativos de TVE". *El profesional de la información*, vol. 19, n.o 4, p. 395–403. <<http://dx.doi.org/10.3145/epi.2010.jul.09> (<http://dx.doi.org/10.3145/epi.2010.jul.09>) >. [Consulta: 10/06/2018].
- Anguera, X. et al. (2012). "Speaker diarization: A review of recent research". *IEEE Transactions on audio, speech, and language processing*, vol. 20, no. 2, p. 356–370.
- Bazán, V. et al. (2018). *Semantics, automatic metadata and audiovisual contents. A case of study: Barcelona International Manga Fair*. <<https://es.slideshare.net/fiatifta/semantics-automatic-metadata-and-audiovisual-contents-a-case-of-study-the-barcelona-international-manga-fair-bazn-gil-virginia-matas-pascual-roberto-gmez-zotomanuel-pastor-sanchez-rtve-spain>>. [Consulta: 06/06/2018].
- Caldera Serrano, J.; Arranz, P. (2013). *Documentación audiovisual en televisión*. Barcelona: Editorial UOC.
- Carrillo, J.; González, A. (2018). *La inteligencia artificial aplicada a los archivos de televisión*. <http://www.rtve.es/contenidos/documentos/instituto/5_Jornada_Archivos_tv.pdf (http://www.rtve.es/contenidos/documentos/instituto/5_Jornada_Archivos_tv.pdf) >. [Consulta: 04/06/2018].
- Etiquedia (2018). *Catalogación automática en entornos audiovisuales* [vídeo]. <<https://www.youtube.com/watch?v=f9j2OhOe5tA>>. [Consulta: 08/06/2018].
- FIAT/IFTA Media Management Commission (2017). *FIAT/IFTA MAM Survey 2017: Highlights from the results analysis*. <<https://es.slideshare.net/fiatifta/2nd-mam-survey-declercq>>. [Consulta: 04/06/2018].
- García, J. (2018). *Machine Learning aplicado en el sector media*. <http://www.rtve.es/contenidos/documentos/instituto/7_Jornada_Archivos_tv.pdf (http://www.rtve.es/contenidos/documentos/instituto/7_Jornada_Archivos_tv.pdf) >. [Consulta: 04/06/2018].
- Jiménez Rayo, M. (2012). "La documentación audiovisual en televisión en el mundo 2.0: retos y oportunidades". *Tripodos*, n.o 31, p. 79–97.
- Jiménez Rayo, M.; Guallar, J. (2014). "Centros de documentación en televisión y productos documentales". *El profesional de la información*, vol. 23, n.o 1, p. 13–25. <<https://doi.org/10.3145/epi.2014.ene.02>>. [Consulta: 10/06/2018].
- Google (2018). *Cloud Video Intelligence*. <<https://cloud.google.com/video-intelligence/>>. [Consulta: 08/06/2018].
- Guerrero, R. (2017). *Introducción al Deep Learning y su uso en clasificación de imágenes* [vídeo]. <<https://www.youtube.com/watch?v=r6PsfQh05xA>>. [Consulta: 04/06/2017].
- IBM (2018). *Watson*. <<https://www.ibm.com/watson/>>. [Consulta: 08/06/2018].
- Karpathy, A. (2015). "NeuralTalk is a Python+numpy project for learning Multimodal Recurrent Neural Networks that describe images with

sentences". <<https://github.com/karpathy/neuraltalk>>. [Consulta: 10/06/2018].

— (2016). "Efficient Image Captioning code in Torch, runs on GPU". <<https://github.com/karpathy/neuraltalk2>>. [Consulta: 10/06/2018].

Karpathy, A.; Fei-Fei, L. (2016). "Deep Visual-Semantic Alignments for Generating Image Descriptions". *IEEE Transactions on pattern analysis and machine intelligence*, vol. 39, no. 4 (April 1 2017). <<https://doi.org/10.1109/TPAMI.2016.2598339>>. [Consulta: 10/06/2018].

Lleida, E. (2018). *Tecnologías para el análisis y metadato de contenidos audiovisuales*. <<http://www.rtve.es/rtve/20180202/jornadas-archivos-television/1672420.shtml> (<http://www.rtve.es/rtve/20180202/jornadas-archivos-television/1672420.shtml>) >. [Consulta: 04/06/2018].

López de Quintana, E. (2014). "Rasgos y trayectorias de la documentación audiovisual: logros, retos y quimeras". *El profesional de la información*, vol. 23, n.o 1, p. 5–12. <<https://doi.org/10.3145/epi.2014.ene.01>>. [Consulta: 10/06/2018].

Manders, T. (2018). *It's all about data*. <<https://es.slideshare.net/fiatifta/its-all-about-data-media-management-information-management-at-nisv-manders-and-van-arkel>>. [Consulta: 06/06/2018].

McDonald, K. (2015). *NeuralTalk and walk*. [Video]. <<https://vimeo.com/146492001>>. [Consulta: 07/06/2018].

Microsoft (2018). *Video Indexer*. <<https://azure.microsoft.com/es-es/services/cognitive-services/video-indexer/>>. [Consulta: 08/06/2018].

Pandolfi, F. M.; Desirello, D. (2017). *A framework for visual search in broadcast archives*. <<https://es.slideshare.net/fiatifta/a-framework-for-visual-search-in-broadcasting-companies-multimedia-archives>>. [Consulta: 06/06/2018].

Pastor, J.; Escribano, M. (2018). *Descripción semántica de escenas ¿El esperanto de las búsquedas?* <http://www.rtve.es/contenidos/documentos/instituto/6_Jornada_Archivos_tv.pdf (http://www.rtve.es/contenidos/documentos/instituto/6_Jornada_Archivos_tv.pdf) >. [Consulta: 08/06/2018].

RTVE (2018). *Google se suma al proyecto "Journalism Innovation HUB" de RTVE*. <<http://www.rtve.es/rtve/20180531/google-se-suma-proyecto-journalism-innovation-hub-rtve/1742823.shtml> (<http://www.rtve.es/rtve/20180531/google-se-suma-proyecto-journalism-innovation-hub-rtve/1742823.shtml>) >. [Consulta: 08/06/2018].

Selkälä, E. (2017). *Automated metadata generation. Projects at YLE*. <<https://es.slideshare.net/fiatifta/automated-metadata-generation-projects-at-yle-2017-selkala-elina>>. [Consulta: 06/06/2018].

Shallue, C. (2016). *Show and tell: image captioning open sourced in TensorFlow*. <<https://ai.googleblog.com/2016/09/show-and-tell-image-captioning-open.html>>. [Consulta: 08/06/2018].

Sound and Vision; Waag Society; Amsterdam Museum; University of Amsterdam (2017). *Freeze! A manifesto for safeguarding and preserving born-digital heritage*. <<https://waag.org/sites/waag/files/public/freeze-manifesto.pdf>>. [Consulta: 01/02/2017].

Teruggi, D. (2018). *Preservación audiovisual: ¿Qué queda por hacer y qué podemos llegar a hacer?* <http://www.rtve.es/contenidos/documentos/instituto/1_Jornada_Archivos_tv.pdf (http://www.rtve.es/contenidos/documentos/instituto/1_Jornada_Archivos_tv.pdf) >. [Consulta: 08/06/2018].

Vinyals, O.; Toshev, A.; Bengio, S. (2016). "Show and Tell: Lessons Learned from the 2015 MSCOCO Image Captioning Challenge". *IEEE Transactions on pattern analysis and machine intelligence*, p. 625–663.

Vinyals, O. *et al.* (2015). "Show and tell: A neural image caption generator". *IEEE Conference on computer vision and pattern recognition (CVPR)*, p. 3.156–3.164.

VSN (2018). *VSNExplorer Artificial Intelligence*. <[https://vsinternal.cloud4tv.com/Storage/Downloads/Online/Videos%20Demo/DEMO%20-%20VSNEXPLORER%20A_7928/7607/DEMO%20-%20ARTIFICIAL%20INTELLIGENCE%20\(voiceover\).mp4?file=5781](https://vsinternal.cloud4tv.com/Storage/Downloads/Online/Videos%20Demo/DEMO%20-%20VSNEXPLORER%20A_7928/7607/DEMO%20-%20ARTIFICIAL%20INTELLIGENCE%20(voiceover).mp4?file=5781)>. [Consulta: 08/06/2018].

Wang, X. (2018). "TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-rays". <<https://arxiv.org/abs/1801.04334>>. [Consulta: 10/06/2018].
