Conectando la visión y el lenguaje

JESÚS PÉREZ-MARTÍN Estudiante de Doctorado del Departamento de Ciencias de la Computación de la Universidad de Chile e Investigador del Instituto Milenio Fundamentos de los Datos.

BENJAMÍN BUSTOS Profesor Titular del Departamento de Ciencias de la Computación de la Universidad de Chile e Investigador Asociado del Instituto Milenio Fundamentos de los Datos.

JORGE PÉREZ Profesor Asociado del Departamento de Ciencias de la Computación de la Universidad de Chile e Investigador Asociado del Instituto Milenio Fundamentos de los Datos.

En este minuto más de 500 horas de video se están publicando en YouTube.[1] Además, el último *Digital Global Overview Report* estima que diariamente se visualizan mil millones de horas de video en la misma plataforma. Con los videos ganando tanta popularidad, YouTube Creator Academy[2] recomienda que las descripciones transmitan información valiosa para ayudar a los espectadores a encontrar videos en los resultados de búsquedas y comprender lo que mirarán.[3] En este sentido detalla: *"Las descripciones bien redactadas con las palabras clave correctas pueden ayudar a mejorar las visualizaciones y el tiempo de reproducción, ya que ayudan a que el video tenga una mayor visibilidad en los resultados de la búsqueda"*.

La forma de comunicación que más usamos los humanos es el lenguaje natural. Es entonces esencial que sistemas interactivos de Inteligencia Artificial (IA) y robots auxiliares sean capaces de generar texto automáticamente a partir de datos no lingüísticos. Reiter y Dale [1] caracterizan *Natural Language Generation* (NLG) como la producción de textos comprensibles a partir de una representación no lingüística subyacente de la información. Esta definición de NLG generalmente se asocia con la de *data-to-text generation*, asumiendo que la entrada exacta puede variar sustancialmente.

Hoy en día, la generación de texto a partir de una entrada perceptiva no estructurada —como una imagen sin

1 | Estadísticas de YouTube 2021 [infografía] - 10 datos fascinantes de YouTube: https://cl.oberlo.com/blog/estadisticas-youtube.

2 | Academia de creadores de YouTube, educación y cursos: https://creatoracademy.youtube.com.

3 | Consejos de YouTube para crear descripciones inteligentes: https://creatoracademy.youtube.com/page/lesson/descriptions?hl=es-419#strategies-zippy-link-1.

procesar o un video— se ha convertido en un desafío importante en el campo de investigación reciente que combina Visión y Lenguaje (V+L). Específicamente, obtener texto a partir de un video (*video-to-text*) puede efectuarse, principalmente, recuperando las descripciones más significativas de un corpus o generando una nueva descripción dado el video de contexto. Estas dos formas representan tareas esenciales para las comunidades de procesamiento de lenguaje natural y visión computacional, y son ampliamente conocidas como *video-to-text retrieval* y *video captioning/description*, respectivamente. Ambas tareas son sustancialmente más complejas que generar o recuperar una oración desde una única imagen. La información espacio-temporal presente en los videos introduce diversidad y complejidad respecto al contenido visual y a la estructura de las descripciones de lenguaje asociadas.

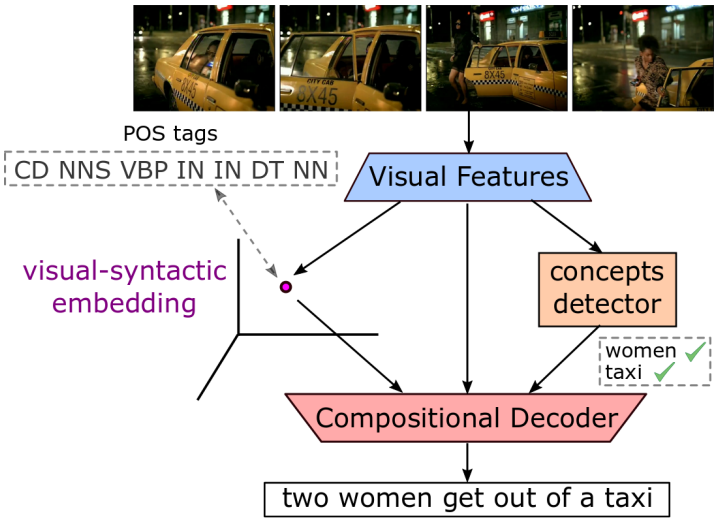

Figura 1. *Video captioning* usando un *embedding visual-sintáctico*. El método obtiene representaciones semánticas y sintácticas de alto nivel a partir de la representación visual del video. A continuación, el decodificador genera una oración a partir de ellos.

Con gran atención de ambas comunidades, V+L incluye otras tareas desafiantes que conectan o combinan las modalidades de la visión y el lenguaje, como *visual question-answering* (responder preguntas basadas en texto sobre imágenes), *caption-based image/video retrieval* (dados un texto y un grupo de imágenes, debemos recuperar la imagen que mejor se describe con el texto), *video generation from text* (generar un video plausible y diverso a partir de un texto de entrada) y *multimodal verification* (dada una o más imágenes y un texto, debemos predecir alguna relación semántica).

Sintaxis y semántica de un video

Es impresionante el progreso que los investigadores han logrado en conjuntos de datos específicos, pero a pesar de este progreso, la conversión de video a texto sigue siendo un problema abierto. Las técnicas del estado del arte aún están lejos de lograr un desempeño similar al humano. No obstante, las técnicas basadas en *deep learning* han logrado resultados prometedores, tanto para la generación de descripciones como para los métodos basados en la recuperación.

Como una tarea de generación de texto, el proceso de describir videos requiere predecir una secuencia de palabras semántica y sintácticamente correcta dado el contexto presente en el video. Los primeros trabajos en esta área siguieron la estrategia de, primero, detectar sujeto, verbo y objeto, formando un *triplete SVO*; y luego, generar una oración usando un conjunto reducido de plantillas que aseguran la correctitud gramatical. Este enfoque requiere que los modelos reconozcan a los sujetos y objetos que participan en la acción que debemos describir, logrando sus mejores resultados en videos cortos de entornos específicos, como deporte o cocina. En este tipo de videos, la cantidad de objetos y acciones que se debe detectar es limitada.

A partir de esta idea, podemos notar que para los modelos de *video captioning* dos aspectos esenciales son la identificación de contenidos visuales de forma explícita y la intención de producir oraciones correctas. Desarrollar técnicas que aborden alguno de estos aspectos ha guiado la investigación en los últimos años. Por un lado tenemos métodos que intentan conectar las palabras generadas a regiones específicas dentro del video (*visual grounding*) [2] y modelar las relaciones entre ellas [3, 4]. Mientras que por el otro tenemos métodos que consideran el aprendizaje de una representación sintáctica como un componente esencial de los enfoques de *video captioning* [5, 6, 7].

En el Departamento de Ciencias de la Computación (DCC) de la Universidad de Chile nos encontramos desarrollando métodos de *video captioning* que extraen información valiosa sobre las posibles descripciones a partir de dimensiones implícitas en la información visual. Nuestros resultados recientes muestran que los videos contienen, además de la apariencia y el movimiento, información semántica y sintáctica que podemos extraer directamente de la información visual para guiar el proceso de generación de

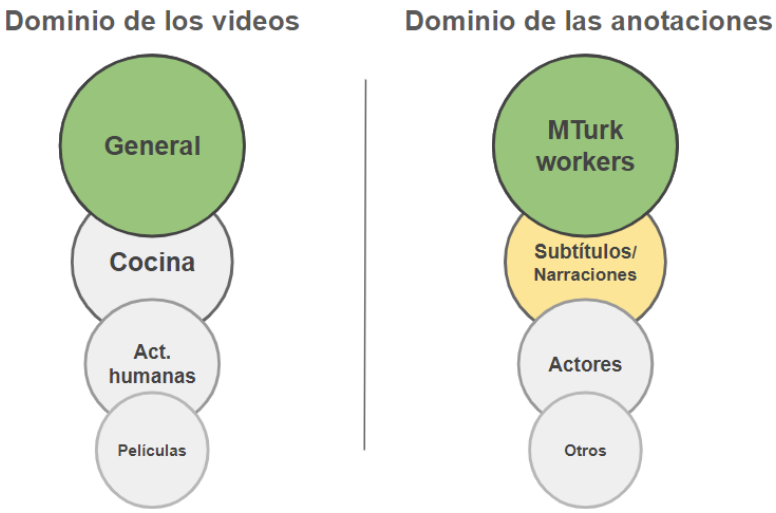

Figura 2. Para entrenar estos métodos, existen más de veinticinco conjuntos de datos anotados que podemos agrupar según el dominio de video y de diferentes formas se obtienen las descripciones.

texto. Sin embargo, tener una fuerte dependencia de sólo una de ellas puede perjudicar el rendimiento de los modelos, produciendo brechas semánticas u oraciones sintácticamente incorrectas. Por eso, para nosotros es fundamental determinar cómo fusionar estos canales de información de forma adaptativa. En dos artículos que presentamos recientemente en las conferencias internacionales ICPR 2020 [8] y WACV 2021 [7], proponemos estrategias efectivas que combinan técnicas de recuperación y generación para evitar estas brechas y aprender representaciones de forma multimodal.

Específicamente, en nuestro trabajo propusimos un modelo llamado *Visual-Semantic-Syntactic Aligned Network* (SemSynAN) [7]. Este modelo basado en el esquema *encoder-decoder* es capaz de generar oraciones con semántica y sintaxis más precisas. Una de las innovaciones más importante fue proponer una técnica de recuperación de secuencias de etiquetado gramatical (POS por sus siglas en inglés)[4] provenientes de las descripciones de video, para generar representaciones sintáctica de alto nivel directamente desde la información visual (ver Figura 1). Con este trabajo mostramos que prestar atención especial a la sintaxis puede mejorar sustancialmente la calidad de las descripciones. Además, nuestro método garantiza la relación contextual entre las palabras de la oración, controlando el significado semántico y la estructura sintáctica de las descripciones generadas [7].

Conjuntos de datos de entrenamiento

V+L es un área de investigación recientemente planteada. Aunque ha recibido mucha atención en los últimos años, todavía se necesitan más datos para entrenar y evaluar nuevos modelos. Para distinguir con precisión entre diferentes clases de información visual, los modelos deben entrenarse a escala, con descripciones diversas y de alta calidad que contengan una amplia variedad de videos.

La creación de conjuntos de datos a gran escala requiere un esfuerzo humano significativo y costoso para su anotación, ya que recopilar una gran cantidad de referencias puede llevar mucho tiempo y ser difícil para los idiomas menos comunes. Debido a esto —y a pesar de que la mayor cantidad de *datasets* ha sido creada a partir de videos de dominio general anotados por humanos (ver Figura 2)—, el *dataset* más grande a la fecha ha sido creado a partir de la generación automática de subtítulos y narraciones (*dataset HowTo100M* [9]).

Con trabajos recientes como CLIP [10], el campo se ha movido a nuevas arquitecturas y modelos (*transformers* [11], *pre-training* y *fine-tuning* ahora se han convertido en el enfoque dominante). Básicamente, estos estudios han mostrado los beneficios de preentrenar los modelos para tareas de V+L y luego ajustar el modelo para tareas específicas.

Por ejemplo, podemos aprender previamente representaciones genéricas a partir de tareas de V+L, como *visual question-answering* o *cross-modal retrieval* (recuperación a través de diferentes modalidades, como imagen-texto, video-texto y audio-texto), y luego ajustar su codificación visual en la tarea de *video captioning*. Esta técnica requiere un gran volumen de datos para aprender dicha representación en un espacio común entre la información visual y textual. Por ejemplo, para entrenar CLIP se usaron 400 millones de pares (imagen, texto) obtenidos de Internet.

Los modelos de *video captioning* basados en esta estrategia, como COOT [12],

4 | Categorizar y etiquetar palabras de acuerdo a categorías léxicas: https://www.nltk.org/book/ch05.html.

generalmente son preentrenados sobre datos obtenidos de forma automática de los subtítulos y narraciones (ver Figura 2) que brindan las plataformas de video *online*. Sin embargo, un gran inconveniente de este tipo de corpus es la gran cantidad de *tokens* desconocidos (términos que no se pueden asociar a una palabra del vocabulario) que se producen. Por ejemplo, en *HowTo100M* [9] sólo el 36,64% de las palabras del vocabulario (217.361 de las 593.238 palabras únicas) aparecen en el vocabulario ampliamente utilizado *GloVe-6B*[5] [13], que tiene 400.000 *tokens*. Este alto nivel de "ruido" en los subtítulos es un aspecto interesante del proceso de entrenamiento que debemos aprender a aprovechar.

Conclusiones

Hace diez años pocos hubieran imaginado que sistemas de V+L serían capaces de generar descripciones textuales plausibles como las que se logran hoy. Los investigadores han logrado modelos que extraen, hasta cierto sentido, información espacio-temporal compleja presente en los videos. No obstante, una característica de la que carecen los sistemas actuales es la capacidad de representar el *sentido común*, por lo que aún queda mucho para comprender y representar la diversidad en cuanto a contenido visual de los videos y la estructura de sus descripciones textuales.

Es muy probable que en el futuro la cantidad de videos que los buscadores deberán procesar sea mayor que en la actualidad. Siempre ha sido así y al día de hoy, que la pandemia nos incita a ser más digitales, no hay ningún indicador que señale que esta dinámica cambiará. Al contrario, esta tendencia aumentará la necesidad de transformar la información visual en descripciones textuales que la resuman, verbalicen y simplifiquen de forma precisa. ▪

REFERENCIAS

[1] Reiter, E. & Dale, R. Building natural language generation systems. (Cambridge University Press, 2000).

[2] Pan, B. et al. Spatio-Temporal Graph for Video Captioning with Knowledge Distillation. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 10870–10879 (2020).

[3] Zhou, L., Kalantidis, Y., Chen, X., Corso, J. J. & Rohrbach, M. Grounded Video Description. In Proc. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 6571–6580 (IEEE, 2019).

[4] Zhang, Z. et al. Object Relational Graph with Teacher-Recommended Learning for Video Captioning. In Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 13278–13288 (2020).

[5] Hou, J., Wu, X., Zhao, W., Luo, J. & Jia, Y. Joint Syntax Representation Learning and Visual Cue Translation for Video Captioning. In Proc. IEEE International Conference on Computer Vision (ICCV) (2019).

[6] Wang, B. et al. Controllable Video Captioning with POS Sequence Guidance Based on Gated Fusion Network. In Proc. IEEE International Conference on Computer Vision (ICCV) (2019).

[7] Pérez-Martín, J., Bustos, B. & Pérez, J. Improving Video Captioning with Temporal Composition of a Visual-Syntactic Embedding. In Proc. IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2021).

[8] Pérez-Martín, J., Bustos, B. & Pérez, J. Attentive Visual Semantic Specialized Network for Video Captioning. In Proc. 25th International Conference on Pattern Recognition (2020).

[9] Miech, A. et al. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In Proc. IEEE/CVF International Conference on Computer Vision (ICCV) 2630–2640 (IEEE, 2019).

[10] Radford, A. et al. Learning Transferable Visual Models From Natural Language Supervision. (2021).

[11] Vaswani, A. et al. Attention is all you need. In Proc. 31st International Conference on Neural Information Processing Systems 6000–6010 (Curran Associates Inc., 2017).

[12] Ging, S., Zolfaghari, M., Pirsiavash, H. & Brox, T. COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning. In Proc. Conference on Neural Information Processing Systems (2020).

[13] Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. IN EMNLP (2014).

5 | Proyecto Stanford GloVe (vectores globales) que usa aprendizaje no supervisado para obtener vectores representativos para un gran conjunto de palabras: https://nlp.stanford.edu/projects/glove/.