Video en la casa de la palabra

Consideraciones sobre la automatización de montaje de video mediante IA

Dis. Mariano Ramis

mariano.ramis@fadu.uba.ar
Universidad de Buenos Aires. Facultad de Arquitectura, Diseño y Urbanismo.
Instituto de Arte Americano e Investigaciones Estéticas “Mario J. Buschiazzo”.

Este texto, fue presentado en el Seminario de Crítica Instituto de Arte Americano en octubre de 2023 y forma parte de mi proyecto de Tesis de Maestría en Diseño Interactivo FADU, UBA. El mismo fue en gran parte desarrollado gracias a la beca ELAP, otorgada por el Canadian Bureau of International Education en 2022, bajo la tutoría del profesor Ricardo Dal Farra, a quien agradezco particularmente, así como también a la Concordia University de Montreal.

Resumen

El presente texto pretende ser una descripción tecnológicamente accesible del estado del arte del montaje automatizado de video, conocido como "*neural edit*", a través del desglose de sus partes y su funcionamiento. Esta descripción incluye una parte predominantemente técnica, en la que el algoritmo de montaje automatizado es abierto para examinar sus partes y funciones, con la intención de advertir mediante esa disección, la concepción y usos para el montaje que establecen. El contexto de esta forma de montaje automatizada, es el de la exploración de múltiples desarrollos y aplicaciones prácticas para redes neuronales. Estas evidencian la proliferación de la denominada inteligencia artificial, con el consecuente reemplazo de tareas que solían llevar adelante seres humanos exclusivamente, ya que explotan capacidades intelectuales, cognitivas y comunicacionales.

En segunda instancia, esta presentación genera comparaciones conceptuales entre la técnica de montaje *neural edit* y textos y prácticas que estudian el montaje como instrumento dentro del campo de la creación cinematográfica y audiovisual. Fundamentalmente, se siguen las ideas del cineasta y pensador ruso del cine Serguei Eisenstein retomadas en el texto "*Film in the house of the word*" del poeta y cineasta estadounidense Hollis Frampton.[1]

Palabras clave: montaje, I.A., cine estructural, automatización, palabras.

El cine constituía una forma narrativa tan nueva e insólita que la inmensa mayoría del público no acertaba en comprender lo que veía en la pantalla, ni establecer una relación entre los hechos.
Luis Buñuel

[1] Los contenidos de corte técnico relacionados a la inteligencia artificial y el texto de Frampton que se presentan en este texto en español son traducciones del autor.

Apuntes técnicos, abrir un algoritmo

Para poder comprender cómo funciona el montaje automatizado de video es necesario estudiar la tecnología que lo motoriza, más allá del hecho de que el *neural edit* es, básicamente, un algoritmo informático. Dicho de otra manera, el núcleo conceptual y práctico del montaje automatizado emerge de la conjunción de tecnologías en estado de desarrollo que lo componen. Entender sus limitaciones y capacidades es, también, conocer de qué modo definen el sentido y utilidad de la yuxtaposición de imágenes. Se considera, además, que esta técnica puede caracterizarse como un eslabón más en el devenir del montaje como "objeto técnico" (Simondon, 2007), que actualiza, en términos de tecnología, uso y formato, un modelo narrativo y comunicacional preexistente.

Comprender un algoritmo significa ir más allá del hecho de que hay algo que entra y sale modificado, luego de atravesar la maquinaria interna que lo compone. En tanto, la intención del presente texto es generar una descripción técnica accesible, que sume elementos tecnológicos a la discusión filosófica sobre el poder del montaje audiovisual. Para comenzar con el desglose se puede decir que, dentro del algoritmo de *natural edit,* hay piezas llamadas redes neuronales utilizadas para reconocimiento de elementos a gran escala y asignación de orden, estas redes de neuronas artificiales son las que identifican el hecho de que este proceso es, o hace uso de, Inteligencia Artificial (de ahora en más IA). Lo primero es, entonces, comprender lo que implica, en términos informáticos, que un sistema automatizado pueda "ver" lo que hay en una secuencia de fotogramas.

Reconocimiento de elementos en imágenes, visión por computadora

La visión por computadora es una disciplina que desarrolla métodos para procesar, analizar, discriminar y finalmente comprender el contenido de información capturada desde el mundo "real" con el fin de generar información numérica o simbólica que pueda ser comprendida por una computadora (Szeliski, 2011). De manera semejante a como los humanos usamos nuestros sentidos para interpretar e interactuar con el mundo que nos rodea, la visión por computadora intenta emular estos procesos para que una computadora

pueda percibir y comprender el contenido de una fotografía, o de acciones desarrolladas en una secuencia de fotogramas, o en una señal de video.

El origen de la visión por computadora se remonta a hace alrededor de 70 años y está vinculado al desarrollo de redes neuronales artificiales primitivas "cuyo objetivos era comprender si modelos computacionales eran capaces de aprender relaciones lógicas" (Kelleher, 2019, s/n). Su utilización a gran escala es reciente, gracias a las capacidades de cómputo disponibles en la actualidad, y a un ecosistema de utilización masivo de datos que demanda soluciones de moderación, discriminación y distribución automatizada natural en los repositorios audiovisuales en la red. Principalmente, en plataformas como *YouTube,* fundada en 2005 y adquirida por *Google* en 2007 (Burgess y Green, 2018).

En experiencias primitivas como la del *Perceptron*, un prototípico sistema de visión artificial desarrollado por el psicólogo norteamericano Frank Rosenblatt, se le asignaban tareas de reconocimiento de carácter binario. Es decir, recibía ejemplos desde los cuales extraer información para su entrenamiento y su resultado tenía sólo dos alternativas: "correcto" o "incorrecto" (Alpaydin, 2019). Las tareas asignadas al *Perceptron*, debido a sus características básicas, nunca pasaron del estado experimental en laboratorio. Es, incluso, difícil imaginar una función o tarea sofisticada que este sistema pudiera resolver con eficiencia. Sin embargo, el principio de entrenamiento de esta tecnología tiene puntos en común que permiten imaginarlo como el origen de los sistemas de reconocimiento actuales, que incorporan redes neuronales profundas de tipo convolucional.

En la actualidad, las redes neuronales convolucionales (CNN)[2] están detrás de procesos como el reconocimiento óptico de caracteres; la inspección de procesos industriales; el estudio de imaginería médica; la seguridad y conducción automotriz; la captura de movimiento; la vigilancia; el estudio de huellas dactilares; y, desde luego, la extracción de etiquetas que enumeran el contenido a partir del estudio de imágenes o secuencias de video (Szeliski, 2011).

---

2 Sigla por su denominación en inglés *Convolutional Neural Networks.*

Redes neuronales convolucionales

Existen diversas aplicaciones para la IA, y también muchos tipos de redes neuronales. Su diseño estructural suele revelar la función para la que fueron creadas. Las CNN fueron originalmente diseñadas para tareas de reconocimiento de imágenes (Kelleher, 2019).

Estas redes tienen la capacidad de contrastar algo que "ven" o se les presenta, con el resultado abstracto de toda la información visual acumulada con la que fueron entrenadas para luego, a medida que avanza el proceso de análisis, devolver porcentajes de coincidencia para clasificación de esos ítems detectados. Por ejemplo, al ver un fotograma en el que hay un sofá responden con grados de aproximación a elementos que conocen con ejemplos fotográficos, como ser 79% sofá, 45% silla y 23% cama.

Las redes convolucionales profundas cuentan con distintos layers o capas, cuyas capacidades están asociadas a su previo entrenamiento. Las neuronas en las capas de la red más superficiales son las más "abstractas" y se encargan de "contrastar" las texturas, mientras que las más profundas lidian con detalles visuales más específicos de una clase y categoría de objetos o cosas. La profundidad de las redes está, de alguna manera, vinculada con la sofisticación de sus funciones. Sin entrar en detalles que no hacen a la intención de esta introducción, es importante comprender que el reconocimiento de elementos gráficos simples, como letras, números o señales de tránsito, implica un proceso de entrenamiento menos complejo que el necesario para reconocer elementos tan diversos como un automóvil familiar fabricado en Francia, o una liebre dando saltos en un una pradera nevada. Las redes neuronales convolucionales dividen entre las neuronas de sus capas el problema de la matriz pixelar de una imagen, para arribar finalmente a un resultado de clasificación (Kelleher, 2019).

La analogía que comúnmente se utiliza para describir cómo funciona el análisis para reconocer los contenidos de una imagen, es la de una linterna alumbrando una habitación oscura. Si nos paramos en la puerta de la habitación y barremos el espacio con la luz de la linterna, los fragmentos que se van haciendo visibles suman argumentos para imaginar qué hay dentro de esa habitación. En las redes convolucionales, el "haz de luz de la linterna" se llama "máscara convolucional". Esta máscara, que suele tener 3x3 píxeles de

tamaño, estudia la estructura pixelar “barriendo” la imagen y sus gradientes pixelares, para obtener similitudes porcentuales que le permitan saber qué elementos reconocibles visualmente están presentes en la imagen. Las interpretaciones siempre son en términos porcentuales, basadas en grados de coincidencia. Este proceso comienza por las coincidencias generales y culmina por las más particulares, sopesando opciones para determinar con mayor certeza el resultado del análisis.

En el caso de las secuencias de video, previo al barrido de cada fotograma por la máscara convolucional, se genera una extracción y filtrado en el que, por ejemplo, se reduce la resolución de cada cuadro y se especifican qué cantidad de fotogramas por segundo serán analizados. Al finalizar esta operación el resultado es una serie de palabras sueltas, sustantivos que representan los ítems que fueron reconocidos durante el análisis, esta instancia es conocida como *video tagging* o video etiquetado y, básicamente, consiste en convertir una secuencia de video en una serie de palabras (**Figura 1**). A modo de ejemplo, si el video analizado es de un partido de fútbol, un probable resultado del proceso de video tagging podría ser una lista de palabras tal como: “pelota”, “césped”, “ropa deportiva”, “personas”, “cielo” o “nubes”.

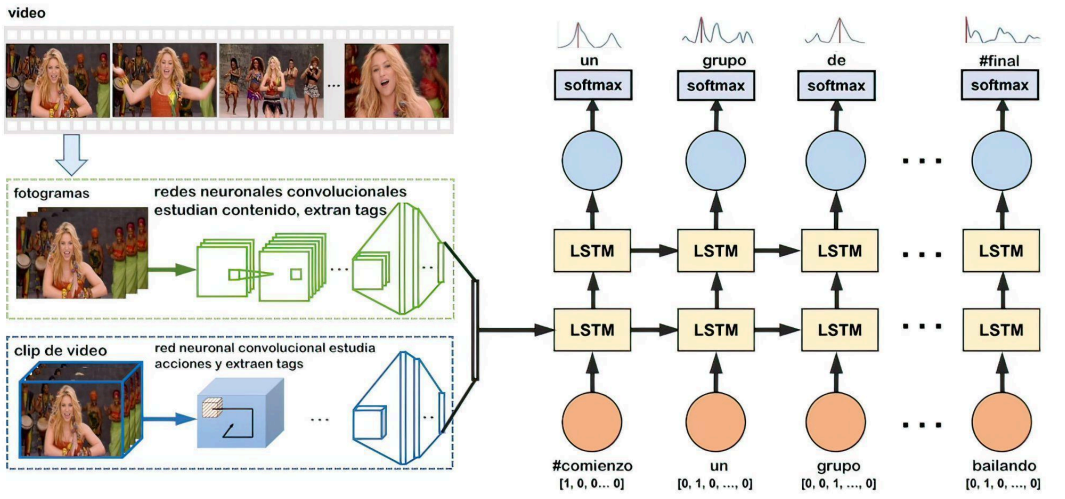


Figura 1: *Esquema de sistema de descripción de contenidos de video. Fuente: Jun Xu, Tao Mei, Ting Yao y Yong Rui (2019).*

Redes neuronales recurrentes o recursivas

Se ha mencionado hasta aquí, brevemente, el contexto de la detección de elementos en fotogramas de un video, pero no aún la descripción y comprensión del contenido. Es necesario, para ello, tratar otros tipos de redes con las que estamos, tal vez sin saberlo, muy familiarizados. Aquellas están cada vez más presentes en nuestra vida cotidiana: sugieren modificaciones en nuestro textos, ayudan a traducir oraciones o hacen más veloz la redacción de un mensaje vía *Whattsapp*.

Estas son las redes neuronales recurrentes o recursivas (RNN)[3] y la subcategoría de la redes de memoria de largo corto plazo (LSTM)[4]. A diferencia de las redes neuronales de una sola dirección, de entrada y salida estándar, las LSTM tienen conexiones de retroalimentación. Pueden procesar no solo puntos de datos individuales (como imágenes), sino también secuencias completas de datos, como por ejemplo series de palabras en una oración o secuencias de frames de video.

Aquí un ejemplo para dejar más en claro qué significan las conexiones de retroalimentación, sin entrar en detalles técnicos. Imaginemos que estamos escribiendo un mensaje y, como suele suceder, el modo predictivo de muchos procesadores de texto se anticipa a la siguiente palabra y la sugiere antes de que sea tipeada. Pero si modificamos el texto previo a esa palabra latente, muy probablemente la palabra que nos sugiere el procesador de texto también cambie. Esto revela que el sistema de recomendaciones está reorganizando y releyendo continuamente en la secuencia, iterando y analizando nuevamente, tratando de comprender el sentido del orden. Esto es posible gracias a las conexiones de retroalimentación (recurrentes) en este tipo de redes neuronales, que son eficaces para estos tipos de problemas temporales y de secuencia.

Las RNN también son originalmente entrenadas, en este caso, mediante secuencias válidas de texto. En ese entrenamiento se basa su capacidad para "aprender" a colaborar con la escritura en tiempo real, a partir de un modelo surgido del entrenamiento con

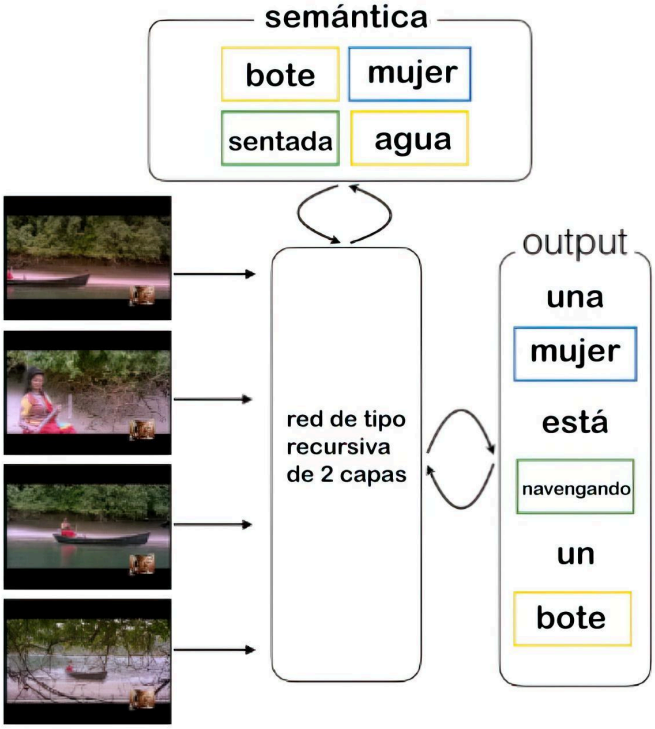
---

3 Sigla por su denominación en inglés *Recurrent Neural Network.*

4 Sigla por su denominación en inglés *Long-Short Term Memory.*

millones de ejemplos de textos validados como correctos. Como fue detallado anteriormente, la capacidad de estas redes es la de ser “conscientes” de una secuencia temporal o de orden. A diferencia de las redes convolucionales, que son utilizadas para escrutar una imagen y determinar los elementos dentro de la misma, las redes recursivas son capaces de ir ordenando datos uno tras otro, y pueden verificar a cada paso el “sentido general” de la secuencia que se conforma.

En este sentido, operan dentro de un algoritmo de edición automatizada, o natural edit,



donde toman las palabras sueltas para organizarlas secuencialmente en oraciones que tengan sentido, dentro de las que estadísticamente son más utilizadas. En comparación al ejemplo anterior, si las etiquetas resultantes de la inspección de la red convolucional son: “pelota”, “césped”, “ropa deportiva” o “personas”, es estadísticamente probable que el video corresponda a un partido de fútbol. Por tanto, la oración resultante podría ser: "Un grupo de personas con ropa deportiva juega al fútbol en una cancha de césped durante el

día". Esta instancia es conocida en el campo de la visión por computadora como ""video descripción". Al sumar ambas redes, la que reconoce elementos (CNN) y la que ordena palabras y crea conexiones para que tengan validez sintáctica (RNN), es que es posible, para un sistema automatizado, comprender con relativa certeza lo que está sucediendo en el video. Cabe destacar que cuanto más habitual sea la acción que acontece en los fotogramas, más posible será que la descripción automatizada se ajuste al contenido.

Edición neuronal de video, *neural edit*.

Como fue descrito hasta aquí, el uso de estos dos tipos de redes configuran lo que se conoce como video descripción. Esta descripción en forma de oración compuesta de palabras es lo que tradicionalmente entendemos como el "guión", llamativamente ubicado aquí como proceso intermedio en la creación, cuando naturalmente suele ser la pieza inicial de una producción audiovisual (**Figura 2**).

Figura 2: *Esquema de algoritmo de video descripción automatizada. Fuente: Laokulra, Okazaki, y Nakayama (2018).*

Una vez que existe este texto que enuncia y describe lo que ocurre en el material de video al que refiere, resta organizar todas esas descripciones en una secuencia narrativa lógica, y asignar luego los videos que cada descripción representa respetando el orden secuencial asignado por ese "guión". La automatización de la edición del video a partir de ese orden cierra la tarea del algoritmo de montaje automatizado o *neural edit*. El resultado final es un video cuyo montaje fue generado en función de parámetros de orden surgidos de un texto previo, creado por redes neuronales artificiales. Conociendo sintéticamente los principios de funcionamiento técnico, vemos cómo se describe la herramienta a sí misma, por ejemplo el texto promocional de una herramienta que hace uso de la técnica edición neuronal llamada FLO (**Figura 3**).

Figura 3: *Imagen publicitaria de Nexgear FLO. Fuente: https://insidebigdata.com/ consultado el 10/07/2021.*

Editar videos es difícil, requiere mucho tiempo y habilidad. Para eso, *NexGear Technology* ha creado *FLO*, una cámara inteligente y una aplicación de creación de películas. Esta edita automáticamente secuencias de video sin procesar y las convierte en películas cinematográficas cortas, por medio del uso de aprendizaje profundo e inteligencia artificial. *FLO* combina el aprendizaje automático, la visión por computadora y el procesamiento del lenguaje natural para reunir los momentos interesantes de sus videos en una película cinematográfica corta. *FLO* reconoce objetos y entiende escenas en el video como un ser humano.

FLO usa redes neuronales convolucionales (CNN) para comprender un video y luego memoria a corto plazo (LSTM) para describir en forma de texto legible por humanos. El modelo de IA ha sido entrenado con aproximadamente 100.000 imágenes descritas por humanos. Actualmente, el sistema es bueno para describir las escenas correctamente, pero puede mejorar aún más si se entrena con más imágenes *(texto publicitario de Nexgear FLO. https://insidebigdata.com/ consultado el 10/07/2021).*

Esta secuencia de explicaciones elementales sobre el funcionamiento de la aplicación *FLO*, además de destacar el hecho de que la tarea de edición es realizada mediante el uso de IA, intenta atraer público promocionando las virtudes de la aplicación y da a entender que la extracción de elementos de un video, su desglose y la posterior traducción a palabras, son sinónimo de comprensión. Al mismo tiempo, detalla el hecho de que editar videos es difícil y toma mucho tiempo pero, aún así, promete reunir mediante el montaje los momentos más "interesantes" de los videos a editar, para generar resultados "cinematográficos".

En síntesis, en términos técnicos y teóricos, el montaje automatizado de video mediante inteligencia artificial se basa en una forma de equivalencia entre imágenes y palabras. Equivalencia que en la sabiduría popular se difunde con la habitual expresión "una imagen vale más que mil palabras", que en este caso pierde su vaguedad e incorpora un número finito y certero de términos que serán representantes de lo que anteriormente era una secuencia de fotogramas de video.

Estas palabras adquirirán luego órdenes coherentes que terminarán de generar descripciones de cosas y situaciones, descripciones válidas, pero arbitrariamente relacionadas con los videos que representan, ya que la manera de conformar esas oraciones descriptivas no es la comprensión del contenido de los videos, sino la asociación de las palabras extraídas a oraciones descriptivas genéricas. Esto, al margen de significar que hay un importante grado de error posible, puesto que la oración resultante puede coincidir en palabras pero describir cosas distintas, implica que comprender un video en términos de inteligencia artificial y montaje neuronal, es convertirlo en palabras. Y editar, convertir el orden de esas palabras en el orden de secuencias de video.

Contexto

Luego de explicar, espero que con cierto éxito, lo elemental del intrincado sistema técnico de edición de video mediante inteligencia artificial, quiero detenerme en el hecho de que el alcance de las tecnologías involucradas allí, están asociadas a otras múltiples tareas en el contexto masivo de moderación, y distribución de videos en la web. Fundamentalmente la plataforma *YouTube* ha sido epicentro de desarrollo de muchas de las habilidades de

redes neuronales, para lidiar de manera automatizada con organización de millones de videos a diario. Comprender el funcionamiento de sistemas de automatización de edición de video es, además de una tarea interesante en términos de ingenio técnico, una manera de estudiar lo que sucede a gran escala. Revela un punto nodal de los sistemas de distribución masivos, ya que la etapa de comprensión y descripción de contenidos es realizada por redes neuronales y sistemas semejantes a las que se vieron en detalle en este texto. Esas mismas redes neuronales y sistemas son los responsables de las sugerencias de un contenido luego de haber visto otro similar, o de "bajar" un video por contener imágenes que infrinjan derechos de copyright o exhiban contenido inapropiado (Burgess y Green, 2018). Pero, al margen de reconocer las ramificaciones e implicancias de estas tecnologías, lo que interesa en este caso es hablar de concepción de montaje que promueven, contrastandolas con otras previas, para comprender el sentido del orden en el montaje, las equivalencias entre imágenes y palabras, y el alcance de la comprensión artificial de las secuencias audiovisuales.

Eisenstein y Frampton: teorías del montaje cinematográfico que estudian el vínculo entre secuencias de palabras y secuencias de imágenes. Film in the house of the word (El cine en la casa de la palabra)

Eisenstein, fallecido a finales de la década de 1940, muy probablemente nunca haya tenido contacto con la idea de que un sistema informático pudiese ser capaz de montar una película. Disciplinas como la cibernética, que estudia la comunicación e intercambio de tareas entre máquinas y personas, apenas se estaban gestando en ese entonces. Vale decir también que, contando con las capacidades de la tecnología de ese momento, hubiese sido sumamente complejo generar un sistema automático capaz de editar una película, por medio del uso de recursos artificiales de comprensión de imágenes. No obstante esa imposibilidad, si se retoman la teorías de montaje de Eisenstein, es posible suponer que los métodos que convierten imágenes en palabras están en conflicto con sus expectativas respecto al poder del cine como forma de arte y comunicación paralela a la palabra, como describe el poeta y cineasta estructural Frampton en su breve texto "*Film in the house of*

the word" de 1980, en el que estudia las ideas de Eisenstein. Según Eisenstein, no era simplemente el sonido, entonces, lo que amenazaba con destruir todos los logros formales del montaje, sino el dudoso don del habla, la decodificación lineal del terreno del pensamiento en una corriente de enunciación (Frampton, 1980).

Entre las razones de Eisenstein para proyectar este revolucionario mecanismo narrativo y comunicacional, muy probablemente estaba el hecho de que el público de sus películas pertenecía a un flamante y gigantesco estado federal en el que se hablaban 15 idiomas distintos, la Unión de Repúblicas Socialistas Soviéticas (URSS) (Fitzpatrick, 2018). Construir películas cuyas estructuras utilizan enunciaciones en texto escrito, diálogos hablados, o descripciones mediante palabras, implicaba algunos problemas. En principio si los diálogos o descripciones estaban escritos en placas entre las escenas, un alto porcentaje de los espectadores no las podrían comprender, ya que en la década de 1920, el porcentaje de analfabetismo en la URSS era cercano al 60% (Fitzpatrick, 2018). Aún más complejo era el hecho de la diversidad de idiomas que existían en ese vasto territorio, lo que trasladaba este inconveniente también a la posibilidad de que el público comprendiera los diálogos o descripciones sincronizadas en las cintas cinematográficas. No obstante estas barreras idiomáticas y técnicas, el objetivo de Eisenstein era que el cine trascendiera esas dificultades con sus propias capacidades audiovisuales, convirtiéndose en un medio autónomo y revolucionario, donde la palabra en cualquiera de sus formas, hablada, escrita, estructurando diálogos o dando forma a una enunciación, sería considerada una suerte de intromisión. Retomando el texto de Frampton, Eisenstein pretendía, más allá del montaje intelectual:

> la construcción de una máquina, muy parecida al cine, más eficiente que el lenguaje, que podría, entrando en competencia directa con el lenguaje, trascender su velocidad, abstracción, compacidad, democracia, ambigüedad, y poder, un proyecto, además, cuya última promesa era la constitución de una crítica externa del lenguaje mismo (Jenkins, 2009, p. 169).

Este ambicioso plan contaba con fundamentos surgidos de un estudio detallado realizado por Eisenstein sobre el funcionamiento enunciativo, expresivo y comunicacional de

jeroglíficos, pictogramas e ideogramas, como antecedentes del montaje basado en imágenes. La síntesis argumental de estas capacidades ancestrales humanas para comunicarse mediante dibujos y sus relaciones condujo a Eisenstein a hablar de un "pensamiento sensible" (Eisenstein, 2020b), aptitud humana que el cineasta exploraba cuando montaba una imagen con la siguiente, complementando ideas que pretendía transmitir a la mente del espectador. Además, aunque cada imagen entre en la conciencia, y en la percepción por agregación, cada detalle se conserva en las sensaciones, y en la memoria como parte del todo (Eisenstein, 2018).

Vale aclarar que la presente investigación se limita a la relación de Eisenstein con las palabras, incluída principalmente en su concepción de montaje de atracciones, pero se obvian sus métodos de montaje, que incluyen elementos rítmicos y formales que profundizan aún más el estudio de las capacidades de la yuxtaposición de imágenes (Eisenstein, 2018).

Comparativamente, la expectativa de Eisenstein respecto al poder de las imágenes cinematográficas como instrumento mediante el cual establecer vínculo con la sensibilidad humana, deja a las herramientas de montaje mediante redes neuronales en evidencia de su artificialidad, o más bien superficialidad, en cuanto a lo que significan en términos de comprensión del contenido de una imagen, cosa que como ya hemos revisado se limita a herramientas que extraen de imágenes un par de palabras, para luego convertidas estadísticamente en descripciones sintácticamente coherentes.

Adentrándose en la complejidad sensible de las imágenes, Frampton cita la inquietud de su colega, el cineasta Stan Brakhage “¿Cuántos colores hay en un campo de hierba, para un bebé gateando que nunca ha oído hablar de verde?” (Jenkins, 2009, p. 166). Da vértigo el solo pensar cuántas cosas hay en el universo que se pueden reconocer e identificar dentro de una imagen, convertir esas cosas en palabras, es de algún modo quitar a lo visual su poder evocativo singular y generalizar una interpretación. Esto, es posible decir, que forma parte del fetiche novedoso de las herramientas IA que pretenden traducir lo visual en texto, el texto en visual, como si nada mermara en ese cruce de fronteras.

En cuanto a Frampton, en su perfil como realizador parece tomar un camino contrario a Eisenstein, en la que es tal su obra más difundida, *Zorn´s lemma* (1970). En este film experimental de casi una hora de duración, Frampton libera palabras encontradas en cartelería urbana a la deriva del orden alfabético, oraciones inconexas se van conformando cuando una palabra sigue a la otra, la estructura enunciativa inconclusa que va surgiendo de las mismas toma el control del montaje (**Figura** 4).

Figura 4: *Secuencia de fotogramas de Zorn´s Lemma (1970). Hollis Frampton.*

Surgida de un proyecto previo llamado Word pictures, *Zorn's lemma*, es una de las películas del denominado cine estructural más representativas. Queda en evidencia en este film el cruce entre palabras e imágenes que tanto interesaba a Frampton, como poeta devenido cineasta. El montaje de *Zorn´s lemma* está basado en una proposición teórica que habla de la lógica de relación entre conjuntos y sus subestructuras, dentro de la que Frampton integra un poema infantil que se utiliza para enseñar las letras del abecedario a niños, pero en esta versión las palabras del mismo son recitadas en orden alfabético en lugar del original. Las palabras registradas por la cámara no parecen por sí solas conducir el sentido, sino poner en evidencia la naturaleza mecánica de la generación del mismo. Las palabras juegan a ser imágenes, y como tales, parecen herir los cimientos de la construcción audiovisual, corte a corte. En la notas y guión de proyecto describe:

> Palabras: La película tuvo sus inicios con una preocupación por la tensión entre elementos gráficos y plásticos/planos, versus elementos ilusionistas[5] en el mismo espacio. En la película, tal como fue realizada, todas las tomas de palabras escritas son capturadas con la cámara en mano y la mayor cantidad posible también contienen movimiento dentro del encuadre. Busqué la máxima variedad de espacio y superficie. Hay referencias conscientes a cada pintura, dibujo y estilo fotográfico que pude manejar, aunque sin duda son sutiles en términos de tiempo de visualización de veinticuatro cuadros (Jenkins, 2009. p 193).

Tanto en la obra de Frampton, como en la de Eisenstein, se detectan similitudes enunciativas y estructurales entre secuencias de palabras y secuencias de imágenes, que han sido insumo de profundo análisis, expresado como legado en obra y enfáticamente en teoría. Tal vez en sentidos opuestos, complementan las expectativas con respecto a lo que la maquinaria narrativa del cine puede lograr en términos expresivos y comunicacionales, haciendo hincapié en el sistema de montaje de la misma.

Es singular que la preocupación de ambos pensadores coincida, al menos en lo formal, con la lógica de procedimientos de montaje automatizados mediante IA, que sitúan a la palabra como referencia técnica de la que distanciarse para comprender y estructurar sentido mediante el montaje de una pieza audiovisual. Como si el audiovisual en su etapa más informatizada no pudiera escapar del dilema de vivir en la casa de las palabras, como antecedente ineludible para comunicar ideas y describir cosas y emociones.

---

[5] Frampton reemplaza rigurosamente en sus textos la idea de *representación*, por la de *ilusión* cuando se refiere a capturas de objetos del mundo vistos a través de la cámara, por considerar que la distancia entre el mundo y su versión mediada por el cine es excesivamente arbitraria y distante como para considerarla una representación (Jenkins, 2009).

A modo de cierre

Esta presentación ha intentado describir técnicamente el montaje automatizado de video mediante el uso de IA, para luego contrastar este método con la obra y conceptos en torno al montaje de dos cineastas que considero trascendentales y que están emparentados por el estudio de la equivalencia entre palabras e imágenes, en sentido secuencial y comunicativo. El resultado de esta comparación da cuenta de la limitada atención que prestan las herramientas artificiales a los dilemas filosóficos que ha despertado el montaje como instrumento distintivo del cine a lo largo de su historia, dejando en evidencia el perfil eminentemente práctico con el que son desarrolladas como "herramientas de productividad". No obstante este resultado, tal vez algo obvio, el estudio de la herramientas informáticas de este tipo, habilita la comprensión contextual del escenario de la web de donde son nativas.

La masividad de contenido en plataformas como *YouTube*, que incorporan más de ***70 años de video por día***, requieren de formas de moderación y distribución de contenido audiovisual en las redes de inédita cantidad y escala para los que éstas herramientas de comprensión artificial fueron originalmente desarrolladas, también la extracción de momentos principales de contenido audiovisual, y la línea de "montaje" de los sistemas de recomendación de video, son aspectos dentro de los que podría avanzar encontrando contextos que hacen uso de las mismas herramientas incorporadas en el montaje automatizado, como son el video-etiquetado y la video-descripción (Zappina, Malika, Dampiera, y Shakshukib, 2022).

Esto tal vez permita entender que la generalidad del modo en el que funciona el *neural edit* y la superficialidad de sus capacidades, tiene que ver con que es una tecnología extraída de su entorno de utilidad original, para ser explotada como aplicación comercial de uso privado, en donde sus limitaciones se vuelven mas evidentes. En *YouTube*, la video descripción sirve para que recibamos videos parecidos a los que ya hemos visto, en los que el montaje que generan los sistemas de recomendación entre video y video, representan la continuidad entre cosas que se parecen, a partir de la detección automatizada de elementos dentro de videos que comparten etiquetas. Esta misma lógica,

puesta en práctica con nuestros archivos de videos personales, termina en la ejecución por generar videos de secuencias similares, donde se repite un elemento o sujeto, devaluando la utilidad expresiva del montaje.

Creo que es importante, entonces, incluir estas observaciones sobre las técnicas que tienen consecuencias estéticas dentro de los contenidos de nuestras materias, que reciben a estudiantes formados por el uso inadvertido de algunas de éstas tecnologías subyacentes en la web. Desde luego, no para prohibirlas o estigmatizarlas, sino para incluirlas y hacerlas parte del diseño de formas narrativas inéditas de las que tal vez sean capaces, por ejemplo reorientando su dirección nuevamente a lo masivo, como instrumentos de "minería de datos audiovisuales", al mismo tiempo resguardando la subjetividad y diversidad sensible y reflexiva, imprescindible en la formación de cada estudiante de diseño.

Bibliografía

Libros:

Alpaydin, E. (2019). *Machine learning.* MIT press.

Buñuel, L. (1983). *Mi último suspiro (memorias).* Plaza y Janés.

Burgess, J. y Green, J. (2018). *YouTube.* Polity Books, Cambridge.

Eisenstein, S. (2020a). *La forma del cine*. Siglo XXI.

Eisenstein, S. (2020b). *El sentido del cine*. Siglo XXI.

Fitzpatrick, S. (2018) *La revolución rusa.* Siglo XXI.

Gombrich, E. H. (2004). *El sentido del orden, estudio sobre la psicología de las artes decorativas.* Phaidon.

Jenkins, B. (2009). *On camera arts and consecutive matters: the writings of Hollis Frampton.* The MIT press.

Kelleher, J. D. (2019). *Deep learning.* MIT press.

Kelleher J. D. y Tierney B. (2018). *Data Science.* MIT press.

Ong, W. J. (2021). *Oralidad y escritura: Tecnologías de la palabra.* Fondo de Cultura económica

Schrage, M. (2020). *Recommendation engines.* MIT Press.

Simondon, G (2007). *El modo de existencia de los objetos técnicos.* Prometeo.

Szeliski, R. (2011). *Computer vision, algorithms and applications.* Springer.

Artículos:

Laokulra, N., Okazaki, N. y Nakayama, H. (2018). *Incorporating Semantic Attention in Video Description Generation, Artificial Intelligence Research Center.* Universidad de Tokio.

Jun Xu, Tao Mei, Ting Yao y Yong Rui (2019). *Large Video Description Dataset for Bridging Video and Language.* Microsoft Research.

Addair, T, (2017). *Deep Learning: YouTube Video Tags.* Stanford University.

Guadarrama, S., Krishnamoorthy, N, Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T. y Saenko, K. (2014) *YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-shot Recognition.* The Computer Vision Foundation.

Knwal, Y., y Tabassam, N. (2021). *A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos.* Department of Software Engineering, University of Engineering and Technology, Taxila, Pakistan.

Zappina, A., Malika, H., Dampiera, D. A. y Shakshukib, E. N. (2022). *YouTube Monetization and Censorship by Proxy: A Machine Learning Perspective.*

Gutierrez, D. (2017) *Introducing Flo – Bringing Deep Learning to Video Editing. Inside Big Data*. Recuperado el 10/07/2021 de: https://insidebigdata.com/2017/06/19/introducing-flo-bringing-deep-learning-video-editing/