VOLUMEN 21
XXVIII Verano De la Ciencia
ISSN 2395-9797
www.jóvenesenlaciencia.ugto.mx

Recolección y análisis de datos mediante técnicas de IA aplicadas al sector turístico.

Collection and analysis of data through AI techniques applied to the tourism sector.

Angel Eladio Coronado Guerrero[1] ,Bryan de Jesús Mares Barrientos[1], Jacobo Hernandez Varela[1], Jorge Emiliano Mora Herrera[1], Mauro Pantoja Guitierrez[1], Piotr Enriquevitch Lopez Chernyshov[1] y Angel Díaz Pacheco[1]

[1] División de Ingenierías, Campus Irapuato-Salamanca, Universidad de Guanajuato, México.
angel.diaz@ugto.mx*

Resumen

El impacto social y la relación con la conservación de los recursos de las localidades con vocación turística, hacen del turismo una de las más importantes fuentes de ingresos de nuestro país. La gran disponibilidad de fuentes de información han cambiado de forma radical las reglas de este sector, obligando a una adaptación gradual de los nuevos paradigmas. Pese a la comprobada eficacia de las encuestas para la obtención de información estratégica, el costo económico y logístico que suponen, dificultan la aplicación de dichos instrumentos de forma periódica. Para aprovechar la disponibilidad de grandes cantidades de información relacionadas al turismo se propone la creación de dos conjuntos de datos (texto e imágenes) que permitan construir modelos útiles para el análisis de información relevante para la toma de decisiones de los organismos de gestión de los destinos turísticos.

Palabras clave: turismo, inteligencia artificial, deep learning, NLP.

Introducción

La reciente crisis sanitaria provocada por el COVID-19 ha confirmado que el turismo es una industria clave para la economía global. De acuerdo con datos de la Organización Mundial de Turismo, en el año 2019 la industria turística generó cerca de $1.5 billones de dólares y fue la tercera más grande exportación a nivel mundial (UNWTO, 2021). No hay duda que los progresos tecnológicos y los nuevos paradigmas como las redes sociales han contribuido al crecimiento de este sector en las últimas dos décadas (Feizollah et al.,2021; Khoa et al., 2021). La alta disponibilidad de información en línea ha hecho del internet el punto de partida para los turistas de todo el mundo en su ciclo de viaje, pues les permite conocer la oferta de productos y servicios disponibles como el hospedaje, medios de transporte, condiciones climatológicas, atracciones turísticas, entre otros, así como tomar decisiones sobre qué lugares visitar y qué actividades realizar.

Dicha abundancia de datos nos permite investigarlos y transformarlos en información útil para la toma de decisiones por parte de los organismos de gestión de los destinos turísticos. A pesar de esto, los algoritmos para el análisis de dichas colecciones requieren de conjuntos de datos creados a medida. Con el fin de que un algoritmo de inteligencia artificial sea capaz de aprender, es necesario presentarle ejemplos cuyo valor es conocido de antemano. Esta labor no es trivial ya que requiere de la recopilación y etiquetado de datos, que en algunos casos se hace a mano. Para el caso del texto, esto se refiere a recolectar fragmentos de información y determinar, en base al contenido el tono del mensaje (positivo, negativo o neutro). Con respecto a las fotografías, un analista debe realizar el proceso de recopilación y evaluación de si la foto en cuestión es relativa al tema bajo análisis o no.

Para el presente proyecto, seleccionamos dos problemas relevantes a la industria turística: el análisis de sentimientos y la detección de objetos en fotografías. El primero se refiere a la capacidad de discriminar si una opinión o reseña en línea (sobre una atracción turística) es positiva (expresada en términos cordiales), negativa (quejas sobre el lugar) o neutra cuando no puede clasificarse en las anteriores. Por otra parte, la detección de objetos en fotografías consiste en determinar la presencia o ausencia de un objeto en particular en una imagen, como determinar si existen palmeras en una foto de una playa o no. Dentro del contexto del turismo esto puede ayudar a evaluar de forma automática la existencia de problemas que afectan la imagen de un destino turístico como basura en las calles, baches y graffiti, por medio de fotografías del lugar.

VOLUMEN 16
XXVII Verano De la Ciencia
ISSN 2395-9797
www. jóvenesenlaciencia.ugto.mx

Este documento está organizado de la siguiente forma: en la sección 2 se presenta una revisión de los trabajos relacionados al tema bajo investigación, la sección 3 presenta la metodología seguida para la recopilación y etiquetado de datos, la sección 4 presenta los análisis preliminares en los conjuntos de datos. Finalmente, la sección 5 presenta las conclusiones y el trabajo futuro.

Trabajo relacionado

La interacción entre los enfoques del ámbito de las ciencias computacionales y la investigación turística no es un tópico nuevo. En varios estudios, las técnicas de Procesamiento del Lenguaje Natural (PLN) han sido utilizadas en un amplio rango de investigaciones relacionadas al turismo (Álvarez-Carmona et al., 2022), y unos cuantos estudios se han propuesto investigar el constructo conocido como imagen del destino mediante técnicas de inteligencia artificial (Díaz-Pacheco et al., 2022).

Los datos multimedia de las redes sociales (en particular el texto) han sido frecuentemente utilizados para entender las percepciones de los turistas en referencia a un destino turístico. Un ejemplo de lo anterior es el estudio de Feizollah et al. (2021) el cual exploró el uso de tweets relacionados al turismo halal (aprobado por la ley Islámica). Ellos adquirieron y pre-procesaron los datos, los cuales posteriormente analizaron utilizando una combinación de análisis de frecuencia, grafos de concordancia y análisis de redes semánticas para identificar los principales temas dentro del corpus de documentos. Por otra parte, Olmos-Martínez et al. (2023), realizaron una investigación exhaustiva sobre artículos de noticias publicados en línea. Esto con el objetivo de ganar comprensión de las percepciones de los turistas sobre una atracción turística. Empleando el web scraping se llevó a cabo un análisis extensivo del corpus de noticias. Utilizando diversas técnicas de PLN como LDA (Latent Dirichlet Allocation) y métricas para la coherencia de tópicos, los autores identificaron los temas clave prevalentes en los artículos publicados por medios de E.U.A y Canadá con respecto a Cancún. Algunos de estos fueron la preocupación por las medidas sanitarias post pandemia y la seguridad de la localidad.

Por otra parte, las fotografías juegan un rol crucial para entender como un destino es percibido por los turistas y sobre la información que proveen sobre el lugar. Desde el punto de vista de las organizaciones de gestión de destino, es muy importante el tomar en cuenta la posibilidad de divergencias entre lo que se promueve por medio de la publicidad y lo que experimenta el turista en su viaje.

Un ejemplo de lo anterior, es el trabajo de Nixon (2016), donde probó como las fotos de las redes sociales influencian a los consumidores. Él utilizó un grupo de enfoque con encuestas con fotos antes y después del viaje con respecto a dos destinos: Jordania y Costa Rica. Se encontró que hay una diferencia significativa entre las percepciones anteriores a la prueba y posteriores a ella con respecto a los atributos de los destinos mencionados. Concluyó que los organismos de gestión de destino deben promover diferentes aspectos y no solo características singulares del lugar. Con respecto al uso de técnicas de IA, Arefieva et al. (2021) propusieron el uso de varias técnicas de aprendizaje profundo para analizar fotos en Instagram. Para ello, los autores utilizaron la API de Google Vision para extraer descripciones de los elementos presentes en cada foto y transformaron dichas descripciones en diferentes representaciones vectoriales para realizar segmentado de tópicos. El análisis resultante identificó las principales ciudades de Austria preferidas por los turistas y los atributos más relevantes para los turistas.

Los trabajos analizados, nos permitieron identificar las técnicas más adecuadas para dar tratamiento o caracterizar los conjuntos de datos recolectados. Por otra parte, también han proporcionado una guía sobre los métodos a utilizar para su posterior análisis.

Recolección de datos

Para mayor claridad en la lectura del documento, los procesos llevados a cabo para la recolección de datos se separarán en dos subsecciones, la primera referente a la captura de los datos textuales y la segunda para la obtención de las fotografías. Con el fin de obtener información relevante de cada colección, se realizaron análisis de frecuencia de los términos más empleados en los descriptores de cada instancia. Para la colección de fotografías, se utilizó una herramienta para la descripción de las escenas.

Recolección de Tweets para análisis de texto

Para la obtención de los datos de texto, se seleccionó la plataforma Tweeter por ser una de las más ampliamente utilizadas para compartir experiencias y quejas sobre los destinos más visitados por los turistas. Como primera etapa de este proyecto que ha de prolongarse a otros veranos de investigación, se seleccionó como destino turístico a investigar, el estado de Guanajuato. Debido a que tanto estado como capital comparten el mismo nombre, en esta primera etapa solo se tiene interés en obtener un conjunto de opiniones que puedan separarse en dos clases bien diferenciadas, "Temática" y "Polaridad" (ver Tabla 1). Temática se refiere al tópico principal discutido en el tweet, donde por ejemplo si el tema es promover un destino turístico, el analista lo clasificará como (0) para la clase Turismo, (1) si se habla de actividades culturales y así sucesivamente pues el nombre de cada clase es auto explicativo. La clase "Genérico" se empleó para aquellos tweets que no pueden categorizarse en ninguna de las anteriores.

Para su obtención, se utilizó la bien conocida herramienta "Phantom Buster" con subscripción gratuita para uso estudiantil (Sun, 2020; Egger & Yu, 2022). Dicha herramienta proporciona una interfaz accesible para comunicarse con la API de Twitter y hacer búsqueda y descarga de tweets en base a sus Hashtags. En particular se utilizó la herramienta (dentro de Phantom Buster) "Hashtag Collector", utilizando las variaciones léxicas mostradas en la Tabla 2. Se utilizaron los parámetros permitidos de la herramienta para eliminar tweets duplicados o retuiteados. De lo anterior se obtuvieron 1260 tweets disponibles para ser clasificados.

El proceso de clasificación fue realizado por los 6 miembros del equipo. La colección de tweets se dividió en tres partes. Cada tweet fue evaluado por 2 miembros del equipo y clasificado por consenso en la categoría correspondiente. Si por alguna razón no se alcanzó unanimidad en la clasificación, la opinión de un tercer miembro fue utilizada para el desempate.

Tabla 1. *Etiquetas utilizadas para clasificar los tweets en el corpus.*

Temática	Polaridad
Turismo (0)	Positivo (1)
Cultura (1)	Neutral (0)
Naturaleza (2)	Negativo (-1)
Gastronomía (3)	No Texto (2)
Limpieza (4)	
Promoción (5)	
Seguridad (6)	
Política (7)	
Genérico (8)	

Tabla 2. *Hashtags utilizados para realizar la búsqueda de tweets.*

Destino	Hashtags
Guanajuato	#Guanajuato, #GUANAJUATO, #guanajuato, #GTO, #gto

Recolección de fotos en internet

Con el fin de construir un clasificador capaz de detectar elementos detrimentales de la imagen del destino turístico, se definieron las siguientes clases: *"basura", "baches" y "graffiti"*. Se eligieron dichos factores por ser elementos tangibles y fácilmente identificables, por lo cual se realizó la búsqueda y descarga de imágenes en internet con dichos elementos. Por otra parte, adicional a los factores anteriores, fue necesaria la obtención de imágenes que no presentaran dichas fallas para poder instruir al clasificador como son las condiciones *"normales"* libres de los defectos antes mencionados en diferentes lugares.

Debido a la dificultad de esta tarea, los 6 miembros realizaron la recolección de las imágenes asistidos por la herramienta ***"Bulk Bing Image Downloader"***, la cual permite realizar la descarga automatizada de grandes cantidades de imágenes en base a los parámetros definidos para su búsqueda. Para este caso en particular se utilizaron los nombres de cada clase en diferentes idiomas. En cuanto a la clase *"normal"* se utilizaron las palabras paisaje, ciudad, calle, lugar, etc. A pesar de que la descarga es automatizada, no sucede lo mismo con el análisis de si la imagen contiene el factor buscado o no. Dicho análisis se realizó por inspección visual de cada una de las imágenes.

De lo anterior se obtuvieron 2,937 imágenes para la clase *"baches", 3,*228 para la clase *"basura",* 3,865 para la clase "*graffitti"* y 3,016 para la clase *"normal".*

Análisis preliminares de los conjuntos de datos

Nuevamente, para hacer más fácil la lectura del documento, esta sección se dividió en el análisis de texto y el análisis de las fotos. Para evaluar si los datos cumplían con las características requeridas, se extrajeron las principales características (las más frecuentes) de las colecciones. Para el conjunto de imágenes, se utilizó la popular herramienta ***"Places-CNN"*** de Zhou et al. (2018), la cual obtiene descripciones de las escenas en las fotos. Con respecto al conjunto de datos de texto, se realizó una análisis de la frecuencia de los tokens en los documentos pertenecientes a cada clase.

Análisis preliminar del texto

Como primer paso para el análisis de los datos se obtuvieron las siguientes estadísticas sobre la distribución de las clases (ver Tabla 3). Debido a la disponibilidad de ejemplos de algunas clases, el número de instancias totales varía entre las categorías de polaridad y temática. Por otra parte, dentro de los textos analizados se encontraron instancias escritas en Inglés o defectuosas (sin texto o ininteligibles) por lo cual fueron descartadas. Como puede verse en la Tabla 3, el conjunto formado por las "Temáticas" presenta gran desbalance en las clases, por lo cual es necesario realizar una mayor recopilación y clasificación de instancias antes de poder utilizarlo. En cuanto al conjunto de texto de la categoría "Polaridad", este presenta condiciones ideales para poder realizar algunos análisis preliminares.

Tabla 3. *Clases con su respectivo número de instancias entre paréntesis.*

Temática	Polaridad
Turismo (40)	Positivo (323)
Cultura (44)	Neutral (398)
Naturaleza (16)	Negativo (216)
Gastronomía (1)	No Texto (323)
Limpieza (6)	
Promoción (67)	
Seguridad (221)	
Política (235)	
Genérico (252)	

Temática	Polaridad
Turismo (40)	Positivo (323)
Cultura (44)	Neutral (398)
Total: 882	1260

Como parte de los análisis llevados a cabo, se realizó el pre procesamiento de los datos. Aunque la información se obtuvo de la forma más estructurada posible, es necesario realizar algunas acciones antes de poder utilizarla como entrada de los algoritmos de aprendizaje automático. Las etapas seguidas en el pre-procesamiento son:

- Convertir el texto de mayúsculas a minúsculas.
- Eliminación de palabras que no aportan información (artículos y preposiciones).
- Eliminación de signos de puntuación.
- Reemplazo de dígitos con el carácter “d”.
- Reducción de la variabilidad léxica.
- Eliminación de las palabras que aparecen menos de 10 veces en el corpus.

Con lo anterior se limpió el conjunto de datos y se obtuvo un formato más adecuado para ser analizado con otros algoritmos. Como ejemplo para el lector, dicho pretratamiento permitió pasar de tweets de la forma:

> *“#Celaya #Guanajuato | La tarde de este jueves, sujetos armados atacaron una carpintería en el Mercado Cañitos y asesinaron a un hombre a sangre fría, tras dispararle a quema ropa, esto en el barrio de Tierras Negras, estas son las imágenes https://t.co/Ao1BppWd9r”*

A tweets limpios y sin información irrelevante como se muestra a continuación:

> *“celaya guanajuato tarde jueves sujetos armados atacaron carpinteria mercado canitos asesinaron hombre sangre fria tras dispararle quema ropa barrio tierras negras imagenes https t co ao bppwd”*

Como se mencionó previamente, el conjunto de tweets sobre temáticas es aún un trabajo en curso, por lo cual los análisis se realizaron sobre el conjunto de tweets clasificados por polaridad. Sobre el conjunto preprocesado se realizó un análisis de la frecuencia de cada token en cada clase de la colección. Para facilitar la lectura, se crearon nubes de palabras con los hallazgos. En las figuras siguientes (de Fig. 1 a Fig 3), se presentan los tokens más frecuentes en las clases “positiva”, “negativa” y “neutra”. El tamaño de las palabras representa la frecuencia del token dentro del documento.

Figura 1. Nube de palabras con los atributos principales de la clase "positivo".

Figura 2. Nube de palabras con los atributos principales de la clase "negativo".

Figura 3. Nube de palabras con los atributos principales de la clase "neutro".

De las nubes de palabras, podemos observar que existen tokens que introducen "ruido" a los documentos como: "https", "t" y "co", sin embargo, de los tokens restantes podemos observar que para la clase ***positivo*** existen ciertos marcadores relevantes como "salud", "mejor", "municipio", "descubre gto", etc. En la clase ***negativa*** aparecen términos como "ataque", "masacre" y "seguridad". Finalmente en la clase ***neutra*** podemos encontrar términos similares a los de la clase positiva, pero con otros añadidos como "gobierno", "resultados", "informe", etc. A pesar de lo superficial del análisis, las frecuencias de los términos antes

mencionados nos proporcionan una idea de la idoneidad del proceso de clasificación llevado a cabo debido a la adecuada relación de dichos términos con las clases presentadas.

Análisis preliminar de las imágenes

A partir de las imágenes obtenidas se realizó una caracterización de las escenas mostradas utilizando la popular herramienta "Places-CNNs" para obtener las características más destacadas de las escenas presentadas en cada categoría. Para una mejor comprensión, en la Figura 4, se muestra una imagen para mostrar un ejemplo de cada categoría.

Figura 4. Ejemplo de las fotografías en cada clase.

Las descripciones producidas por Places-CNN se utilizaron para construir histogramas de frecuencias con los cuales crear nubes de palabras que muestran las características más importantes de cada conjunto de fotos. A continuación se muestran las nubes de palabras de los conjuntos "baches" (Fig. 5), "basura" (Fig. 6), "graffitti" (Fig. 7) y "normal" (Fig. 8).

Figura 5. Nube de palabras con los atributos principales de la categoría baches.

VOLUMEN 16
XXVII Verano De la Ciencia
ISSN 2395-9797
www. jóvenesenlaciencia.ugto.mx

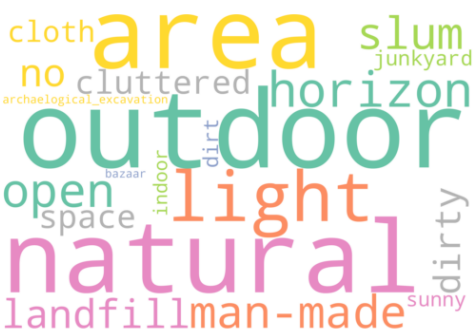

Figura 6. Nube de palabras con los atributos principales de la categoría basura.

Figura 7. Nube de palabras con los atributos principales de la categoría graffiti.

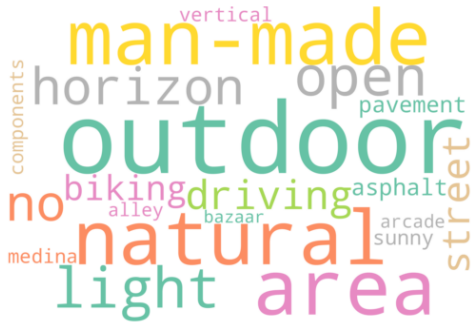

Figura 8. Nube de palabras con los atributos principales de la categoría normal.

Debido a que la herramienta Places-CNN fue desarrollada para funcionar con vocabulario en Inglés, las nubes de palabras resultantes también emplean términos en Inglés. A pesar de esto y que el vocabulario utilizado para describir las escenas es limitado, podemos observar ciertas tendencias similares en todas las fotografías como “outdoor”, “natural area” o “man-made” , ya que dichas imágenes toman lugar en el exterior y con eliminación natural en construcciones artificiales. A pesar de lo anterior, los descriptores para las escenas de la clase ***baches*** contienen términos como “asphalt”, “dirty”, “driveway” and “highway”, los cuales por sentido común podemos fácilmente relacionar con escenas relativas a baches y caminos descuidados. Con respecto a la clase ***basura*** “dirt”, ”dirty”, “junkyard” o “slum” son al parecer términos regulares y muy adecuados para

un conjunto de escenas relativas a basura y basureros. La clase ***graffiti*** contiene términos muy relevantes como “brick”, “art_studio” y “art_school”, mientras la clase ***normal*** contiene varios descriptores comunes a las demás clases pero ninguno de los más relacionados a dichas categorías.

Conclusiones y trabajo futuro

Este trabajo presentó los resultados preliminares de los esfuerzos realizados para obtener los conjunto de datos necesarios para poder construir los clasificadores y herramientas de análisis de datos, que permitan obtener información relevante para la toma de decisiones en la industria turística. La relevancia de los conjuntos de datos recolectados, también estriba en la poca disponibilidad de datos similares como un corpus de texto en español de México, etiquetados sobre temáticas relevantes al sector turístico. Debido a la importancia central de este sector para la economía de nuestro país, contar con herramientas bien calibradas para detectar patrones e información útil será de gran ayuda para la investigación.

Por otra parte, la construcción de un conjunto de datos de imágenes sobre temas desagradables como la basura, baches y el graffiti (no artístico), no ha sido construido hasta el momento. Un clasificador de elementos como los antes mencionados, será de gran utilidad para las organizaciones de gestión de destino para medir de forma automática la percepción de los turistas con respecto a la higiene e infraestructura del destino turístico.

Como se mencionó anteriormente en el documento, esta investigación es aún trabajo en curso, de lo cual se desprenden las siguientes actividades a realizar en subsecuentes ediciones del verano científico.

- Ampliación y revisión de los conjuntos de datos de texto, especialmente para el conjunto de datos sobre temáticas relativas al turismo.
- Ampliación a nuevos defectos de los destinos turísticos detectables visualmente a través del análisis de fotos.
- Construcción de modelos de aprendizaje automático para la clasificación de temáticas y polaridad de texto en español de México.
- Construcción de modelos de reconocimiento de objetos en imágenes sobre los defectos de los destinos turísticos.
- Pruebas y validación en casos de estudio subsecuentes.

Bibliografía/Referencias

Álvarez-Carmona, M. Á., Aranda, R., Rodríguez-Gonzalez, A. Y., Fajardo-Delgado, D., Sánchez, M. G., Pérez-Espinosa, H., ...Díaz-Pacheco, Á. (2022). Natural language processing applied to tourism research: A systematic review and future research directions. Journal of King Saud University - Computer and Information Sciences, 34(10, Part B), 10125–10144. doi: 10.1016/j.jksuci.2022.10.010

Arefieva, V., Egger, R., & Yu, J. (2021). A machine learning approach to cluster destination image on Instagram. Tourism Management, 85, 104318. doi: 10.1016/j.tourman.2021.104318

Diaz-Pacheco, A., Álvarez-Carmona, M. Á., Guerrero-Rodríguez, R., Chávez, L. A. C., Rodríguez-González, A. Y., Ramírez-Silva, J. P., & Aranda, R. (2022). Artificial intelligence methods to support the research of destination image in tourism. A systematic review. J. Exp. Theor. Artif. Intell., 1–31. doi: 10.1080/0952813X.2022.2153276

Egger, R., & Yu, J. (2021). Identifying hidden semantic structures in Instagram data: a topic modelling comparison. Tourism Review, 77(4), 1234–1246. doi: 10.1108/TR-05-2021-0244

Feizollah, A., Mostafa, M. M., Sulaiman, A., Zakaria, Z. and Firdaus, A. (2021), ‘Exploring halal tourism tweets on social media’, Journal of Big Data 8(1), 1–18.

Khoa, B. T., Ly, N. M., Uyen, V. T. T., Oanh, N. T. T. and Long, B. T. (2021), The impact of social media marketing on the travel intention of z travelers, in ‘2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)’, IEEE, pp. 1–6. doi:http://dx.doi.org/10.1300/j073v02n02_12

Nixon, L. (2016, January 01). How Instagram Influences Visual Destination Image – a Case Study of Jordan and Costa Rica. Retrieved from https://www.academia.edu/73603261/How_Instagram_Influences_Visual_Destination_Image_a_Case_Study_of_Jordan_and_Costa_Rica

Olmos-Martínez, E., Álvarez-Carmona, M. Á., Aranda, R., & Díaz-Pacheco, A. (2023). What does the media tell us about a destination? The Cancun case, seen from the USA, Canada, and Mexico. International Journal of Tourism Cities, ahead-of-print(ahead-of-print). doi: 10.1108/IJTC-09-2022-0223

Sun, G. (2020). Symmetry Analysis in Analyzing Cognitive and Emotional Attitudes for Tourism Consumers by Applying Artificial Intelligence Python Technology. Symmetry, 12(4), 606. doi: 10.3390/sym12040606

UNWTO, Panorama del turismo internacional, Edición 2020 (2021), World Tourism Organization (UNWTO). URL: https://doi.org/10.18111/9789284422746

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 Million Image Database for Scene Recognition. IEEE Trans. Pattern Anal. Mach. Intell., 40(6), 1452–1464. doi: 10.1109/TPAMI.2017.2723009