TP2: Clustering de imágenes

Sebastián A Romano, Hernán Varela, Santiago Bezchinsky, Julián Devouassoux
Data Mining en Ciencia y Tecnología

3 de junio de 2025

1. Introducción

Dentro de la minería de datos el procesamiento de imágenes es un campo en constante evolución. Hoy en día son muchas las aplicaciones que funcionan a partir de cámaras, sensores, equipamiento médico, o satélites. Estos sistemas permiten extraer una gran cantidad de información, lo que dificulta la inspección humana. Para poder extraer atributos y poder agrupar las imágenes en base de ellas es sumamente importante la utilización de algoritmos de aprendizaje automático.

2. Objetivos

Realizar un agrupamiento de un conjunto de imágenes. Comparar las segmentaciones generadas por diferentes algoritmos de clustering en base a métodos de validación interna y externa.

En función del dataset elegido se podrá definir una pregunta específica, ésta puede estar guiada por una pregunta básica o por una aplicación. Les proponemos pensar esta preguntar e incluirla como guía del trabajo. Deberá estar motivada en la sección *Introducción* (idealmente con bibliografía) y discutida en la sección *Discusión* o *Conclusiones*.

3. Estructura de los datos

Deberán elegir un dataset, donde haya al menos 5 categorías. Dejamos ejemplos de algunos conjuntos de datos que pueden usar, pero pueden elegir otro que sea de su interés tomando estos como parámetro.

- **Imágenes naturales**: 6.899 imágenes de 8 clases diferentes (aviones, autos, gatos, flores, perros, frutas, motos y personas): `https://www.kaggle.com/datasets/prasunroy/natural-images` [1].
- **Imágenes de tipos de arroz**: 75.000 imágenes de 5 clases de arroz diferentes (Arborio, Basmati, Ipsala, Jasmine, Karacadag,): `https://www.kaggle.com/datasets/muratkokludataset/rice-image-dataset` [2].

- **Imágenes de prendas de ropa**: 44.000 imágenes de prendas de vestir, provenientes de 7 grandes categorías y 45 sub-categorías: `https://www.kaggle.com/datasets/paramaggarwal/fashion-product-images-small`
- **Imágenes de galaxias**: 270.000 imágenes galaxias de 10 categorías diferentes (anotadas por voluntarios online). Dataset: `https://astronn.readthedocs.io/en/latest/galaxy10.html`, Plataforma para anotar: `https://www.zooniverse.org/projects/zookeeper/galaxy-zoo/classify`

Dependiendo del número total de muestras y categorías del dataset elegido, podría ser recomendable hacer un subsampleo de los datasets originales (o sea, hacer un remuestreo más reducido).

4. Preparación de los datos

La preparación de los datos incluye los siguientes pasos:

(a) Levantar las imágenes y sus etiquetas.

(b) Documentar las propiedades del dataset a utilizar.

(c) Describir sus atributos, de forma semejante a lo realizado en el pre-TP.

(d) Recuerde que las imágenes deben ser comparables en color, valor, rango y tamaño.

5. Extracción de atributos con modelo VGG16

Se propone el uso de un modelo de redes neuronales convolucionales (CNNs) que es muy utilizado en el campo de la visión [3]. **VGG16**, es un modelo de *Transfer Learning* de CNN de 16 capas, con pesos pre entrenados para la clasificación de imágenes. Dicho modelo está implementado en *Keras* `https://keras.io/api/applications/vgg/` y por defecto trabaja con imágenes de $224x224$.

Este modelo es entrenado de forma supervisada para la clasificación y aprende en las capas intermedias atributos que son relevante para dicha tarea. En el presente trabajo les proponemos utilizar estos atributos utilizando la salida de las capas intermedias. En este tutorial se ilustra como obtener y trabajar con la salida de atributos intermedios del modelo VGG16.

Siguiendo los pasos descritos en las clases, explorar los atributos intermedios obtenidos al procesar el dataset elegido con el modelo VGG16 y documentar dicha exploración.

6. Clustering

(a) Aplicar *KMeans* sobre el conjunto de atributos obtenidos previamente. Determinar la cantidad óptima de *clusters* utilizando *silhuette* y *SSE*.

(b) Evaluar si el agrupamiento para el k óptimo se condice con las etiquetas de categorías de las imágenes, utilizando la matriz de confusión y los índices de *Rand* y *van Dongen* en los casos que correspondan.

(c) Visualizar los datos coloréandolos de acuerdo a los *clusters* obtenidos y a las etiquetas de categorías, usando una representación de baja dimensión con alguna técnica de reducción (PCA, TSNE, MDS, etc).

(d) Discuta brevemente los resultados obtenidos.

Repetir los pasos con al menos otros dos algoritmos de *clustering*. En el paso *a)* considerar los parámetros y la métrica de validación interna que corresponda. Si el algoritmo lo permite, probar con diferentes métricas de distancia, normalizaciones de los datos, etc. Comparar y discutir brevemente los resultados obtenidos.

7. Detección automática de objetos en una imagen

Se propone detectar objetos dentro de una imagen que ustedes hayan elegido en la que se observen diferentes objetos, siguiendo los siguientes pasos.

(a) Seleccionar una imagen.

(b) Realizar el pre-procesamiento que consideren necesario.

(c) Aplicar los algoritmos de *Connected-component labelling* y *clustering espectral* sobre los píxeles.

(d) Describir el proceso y comparar los resultados obtenidos.

Formato

Les proponemos seguir el formato de publicación en una revista científica, pueden encontrar muchos formatos directamente en *Overleaf*, por ejemplo *NeurIPS*[1] o *IEEE Conference Template for ANCS 2019*. No es obligatorio seguir ese formato, pero si elegir el formato de alguna revista.

Las revistas suelen tener además instrucciones respecto al formato online, desde restricciones en el tamaño de cada sección, en el número de figuras/tablas, las secciones que debe contener, hasta formato de los números, referencias, etc.

Aquí ponemos nuestras restricciones, pero si quieren adaptarlo a alguna revista en particular también vale (explicítenlo en el informe).

Secciones.

1. Título (máx. 100 caracteres), tiene que ser expresivo (no vale TP1).

[1] https://www.overleaf.com/read/mzbjfyxsfxqn

2. Resumen (máx. 200 palabras), tiene que contener una descripción de todo el trabajo: Motivación, Antecedentes, Objetivos, Métodos, Resultados y alguna Conclusión.

3. Introducción Comienza con la motivación, sigue con los antecedentes, y termina siempre con un párrafo de objetivos (no es necesario que este dividido en sub-secciones). Típicamente, una vez que motivaron el trabajo y mostraron lo que hay hecho, viene una frase del estilo "Por ende, nos proponemos..." o "Aquí nos proponemos...".

4. Métodos Detalle de los métodos a utilizar, en este caso no es necesario profundizar mucho pero pueden enumerarlos y sobre todo es el lugar para incluir cualquier método fuera de lo común que hayan utilizado.

5. Resultados y discusión Aquí se enumeran y discuten los resultados. Es muy importante que no sea una seguidilla de figuras y tablas. Como regla pueden considerar: *"Si una figura no se describe/comenta en el texto es que: O bien está de más y no hace a la historia, o bien se olvidaron de incluirla."*.

6. Conclusiones Comienza generalmente con un resumen muy breve de los principales resultados obtenidos (uniendo distintas secciones), y luego se pasa a conclusiones generales, detallando problemas detectados, posibles explicaciones y trabajo a futuro.

7. Referencias Citas bibliográficas utilizadas durante el reporte. Si son sitios web o repositorios se incluyen generalmente al pie de la página que corresponde y no como cita bibliográfica.

Considerando Introducción, Métodos, Resultados y Conclusiones no deben superar las 5000 palabras (aprox.). Finalmente, considerando el formato de este trabajo pueden dividirlo en

1. Título
2. Resumen
3. Introducción
4. Métodos generales (si los hubiese)
5. Experiencia 1: KMeans por ejemplo,

   Métodos específicos

   Resultados y discusión
6. Experiencia 2: Otro algoritmo,

   Métodos específicos

   Resultados y discusión
7. Experiencia N:
8. Conclusiones
9. Referencias

Figuras y tablas

Es muy importante pensar y tomar la decisión de qué mostrar y cómo mostrarlo. A veces no se le da importancia cuando el espacio no está acotado pero igualmente afecta seriamente a la comunicación de los resultados, ya que una mala visualización puede esconder lo relevante y lo que se quiere comunicar. Les dejamos algunas referencias para explorar al respecto especificamente de las decisiones, en cuanto a cómo implementarlo podemos discutirlo en clase [4, 5, 6]

Nota final

Se puede usar cualquier herramienta de análisis o combinación de herramientas, debiendo indicarlas en el informe. Si usan una función ya armada dentro de una librería, detallen los parámetros con la que la corrieron. El lenguaje (Python o R) en el que se desarrolle el TP no es excluyente.

Referencias

[1] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of degradations on deep neural network architectures. *arXiv preprint arXiv:1807.10108*, 2018.

[2] Ilkay Cinar and Murat Koklu. Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7(3):188–194, 2019.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Seán I O'Donoghue, Benedetta Frida Baldi, Susan J Clark, Aaron E Darling, James M Hogan, Sandeep Kaur, Lena Maier-Hein, Davis J McCarthy, William J Moore, Esther Stenau, et al. Visualization of biomedical data. *Annual Review of Biomedical Data Science*, 1(1):275–304, 2018.

[5] Stephen R Midway. Principles of effective data visualization. *Patterns*, 1(9):100141, 2020.

[6] Elena A Allen, Erik B Erhardt, and Vince D Calhoun. Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron*, 74(4):603–608, 2012.