1eras Jornadas de CIENCIA y TECNOLOGÍA

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Universidad Tecnológica Nacional
30 y 31 de octubre de 2024
Libro de Actas de Resúmenes DOI: https://doi.org/10.33414/ajea.1846.2025
Enlace: https://rtyc.utn.edu.ar/index.php/ajea/article/view/1874

nano-datasets: Facilitando la investigación eficiente de la comprensión de video con subconjuntos personalizables de datos a gran escala

nano-datasets: Enabling Efficient Video Understanding Research with Customizable Subsets of Large-Scale Datasets

Presentación: 30/10/2024

Joel ERMANTRAUT
Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca
joelermantraut@gmail.com

Lucas TOBIO
Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca
lucasltobio@gmail.com

Segundo FOISSAC
Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca
segufoissac@gmail.com

Javier IPARRAGUIRRE
Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca
jiparraguirre@frbb.utn.edu.ar

Resumen

El avance del aprendizaje auto-supervisado en la comprensión de video ha sido facilitado por conjuntos de datos a gran escala, pero su tamaño plantea desafíos para los investigadores con recursos computacionales limitados. Para abordar esto, presentamos nano-datasets, un repositorio de scripts diseñados para generar subconjuntos personalizables de conjuntos de datos de video establecidos como Kinetics, Something-Something-v2 e ImageNet-1K. Estos scripts mantienen la integridad semántica y la estructura de los conjuntos de datos originales, al tiempo que permiten a los usuarios crear versiones más pequeñas y manejables adaptadas a sus necesidades de investigación específicas. Al permitir que los investigadores experimenten con diversas arquitecturas y modelos de ajuste fino en conjuntos de datos accesibles, nano-datasets tiene como objetivo democratizar la investigación de comprensión de video y fomentar la reproducibilidad y la colaboración dentro del campo.

Palabras clave: nano-datasets, aprendizaje autosupervisado, representaciones de vídeo, visión artificial, aprendizaje automático

Libro de Actas de Resúmenes DOI: https://doi.org/10.33414/ajea.1846.2025
Enlace: https://rtyc.utn.edu.ar/index.php/ajea/article/view/1874

Abstract

The advancement of self-supervised learning in video understanding has been facilitated by large-scale datasets, yet their size poses challenges for researchers with limited computational resources. To address this, we introduce nano-datasets, a repository of scripts designed to generate customizable subsets from established video datasets like Kinetics, Something-Something-v2, and ImageNet-1K. These scripts maintain the semantic integrity and structure of the original datasets while allowing users to create smaller, more manageable versions tailored to their specific research needs. By enabling researchers to experiment with diverse architectures and fine-tune models on accessible datasets, nano-datasets aims to democratize video understanding research and foster reproducibility and collaboration within the field.

Keywords: nano-datasets, self-supervised learning, video representations, computer vision, machine learning

Introducción

El aprendizaje autosupervisado ha emergido como una técnica prometedora para extraer representaciones útiles de datos sin la necesidad de anotaciones extensas (Rani et al., 2023). En particular, el entendimiento de video sin anotaciones externas ha progresado notablemente en los últimos 5 años. Unas de las arquitecturas que han cobrado popularidad son los video transformers (ViT) (Dosovitskiy et al., 2020). Otra arquitectura es el ViT jerárquico (HIERA) (Ryali et al., 2023). Recientemente, V-JEPA (video joint-embedding predictive architecture), una arquitectura diseñada para capturar relaciones espaciales y temporales en videos, se ha destacado en la literatura (Bardes et al., 2024). Aunque el avance ha sido notable en las métricas reportadas, estos modelos necesitan de grandes cantidades de datos para poder entrenar.

Los autores de los trabajos mencionados utilizan conjuntos de datos (datasets) públicos para poder comparar los resultados reportados. Generalmente, los conjuntos de datos públicos demandan gran espacio de almacenamiento lo cual resulta restrictivo para utilizarlos con una computadora personal contemporánea. Además algunos trabajos usan más de un conjunto de datos para entrenar modelos. Otro factor limitante es la incompatibilidad entre distintos datasets públicos. Generalmente, las anotaciones no siguen la misma convención y no resulta trivial trabajar con más de un repositorio de datos para entrenar un modelo de video o imágenes.

Con el fin de facilitar la investigación y desarrollo de modelos de aprendizaje automático en imágenes y vídeo, hemos creado nano-datasets. El proyecto consiste en un repositorio abierto de software que permite crear un conjunto de datos personalizado a partir de datos públicos. Este enfoque no solo permite realizar investigaciones con recursos más limitados, sino que también facilita la réplica y extensión de experimentos en diversos contextos, incluyendo la exploración de nuevas arquitecturas y ajustes finos de modelos. Adicionalmente, el código fuente se publica como un repositorio de abierto en Internet [1].

Conjuntos de Datos Utilizados

Para facilitar la investigación en aprendizaje de representaciones de video, el repositorio nano-datasets soporta la creación de subconjuntos personalizados a partir de tres conjuntos de datos públicos ampliamente reconocidos. Hasta el momento los conjuntos disponibles son Kinetics-400/600/700 (Kay et al., 2017), Something-Something-v2 (Goyal et al., 2017), e ImageNet-1K (Russakovsky, 2015). A continuación, se proporciona una breve descripción de cada uno de estos datasets y su relevancia en el campo.

Kinetics-400/600/700

Kinetics es una serie de conjuntos de datos de videos que han sido fundamentales en la investigación de reconocimiento de acciones. Kinetics-400 fue la primera versión, que incluye aproximadamente 400 clases de acciones humanas, con 400-1150 videos por clase, capturados de videos de YouTube. Las versiones posteriores,

[1] https://github.com/BHI-Research/nano-datasets

Libro de Actas de Resúmenes DOI: https://doi.org/10.33414/ajea.1846.2025
Enlace: https://rtyc.utn.edu.ar/index.php/ajea/article/view/1874

Kinetics-600 y Kinetics-700, ampliaron el número de clases y videos, llegando hasta 700 clases en la última versión. Estos datasets son utilizados extensamente para entrenar y evaluar modelos de aprendizaje profundo en tareas de clasificación de acciones y han sido una piedra angular para la evaluación de arquitecturas como V-JEPA.

El tamaño y la diversidad de Kinetics lo convierten en un benchmark estándar para modelos de reconocimiento de acciones en videos. Sin embargo, debido a su tamaño, trabajar con este dataset completo requiere recursos computacionales significativos, lo que justifica la necesidad de generar subconjuntos más pequeños y manejables a través del repositorio nano-datasets.

Something-Something-v2

Something-Something-v2 es un conjunto de datos diseñado específicamente para el reconocimiento de acciones interactivas, donde se enfatizan las interacciones humanas con objetos cotidianos. Contiene más de 220,000 videos cortos divididos en 174 clases de acciones, como "Empujar algo hacia arriba" o "Mover algo hacia la derecha". Este dataset es particularmente valioso para la investigación en modelado de interacciones complejas y ha sido utilizado para explorar enfoques de aprendizaje más allá de la clasificación de acciones simples.

La complejidad de las acciones en Something-Something-v2 lo hace un recurso clave para investigar la comprensión de las dinámicas de interacción en video. La creación de subconjuntos específicos permite a los investigadores centrarse en tareas más concretas o reducir la carga computacional sin perder representatividad en el entrenamiento de los modelos.

ImageNet-1K

ImageNet-1K es uno de los conjuntos de datos más influyentes en el desarrollo del aprendizaje profundo. Contiene 1.28 millones de imágenes organizadas en 1,000 clases diferentes, y ha sido utilizado históricamente para entrenar y evaluar modelos de clasificación de imágenes. Aunque ImageNet es principalmente un dataset de imágenes estáticas, su influencia se extiende al aprendizaje de representaciones en video, ya que muchos modelos pre entrenados en ImageNet han sido adaptados y utilizados como base para tareas en video.

Aunque ImageNet-1K es un dataset de imágenes, su importancia en la creación de modelos de visión por computadora es indiscutible. En el contexto de nano-datasets, la capacidad de generar versiones reducidas de ImageNet-1K permite a los investigadores experimentar con modelos ligeros que pueden ser aplicados en tareas de video o combinados con otros datasets para estudios comparativos.

Implementación de Subconjuntos Personalizados

El repositorio nano-datasets se diseñó con el objetivo de simplificar y hacer más accesible la creación de subconjuntos de grandes conjuntos de datos de vídeo e imágenes. Para cada dataset soportado, se proporciona un script dedicado que permite generar una versión reducida del conjunto de datos, manteniendo la lógica en cuanto al contenido, las clases asociadas y la estructura de directorios o nombramiento de archivos. Estos scripts se han desarrollado con un enfoque en la facilidad de uso y la intuición, asegurando que los investigadores puedan adaptar rápidamente los conjuntos de datos a sus necesidades específicas sin necesidad de configuraciones complejas.

Todos los scripts comparten una estructura de argumentos coherente, lo que facilita la familiarización y el uso eficiente por parte de los investigadores. A continuación se detalla la estructura básica de argumentos que se utiliza:

Libro de Actas de Resúmenes DOI: https://doi.org/10.33414/ajea.1846.2025
Enlace: https://rtyc.utn.edu.ar/index.php/ajea/article/view/1874

```
# Ejemplo de ejecución sin argumentos
(nano-datasets) python .\gen_nano_imagenet.py
  gen_nano_imagenet.py: error: the following arguments
are required: --src, --n

# Ejecución típica del script
usage: gen_nano_imagenet.py [-h] --src SRC [--dest DEST]
--n N [--class_fp CLASS_FP]
Copy random files and generate class files.

options:
  -h, --help            show this help message and exit
  --src SRC             Source directory
  --dest DEST           Destination directory
  --n N                 Number of files to copy per folder
  --class_fp CLASS_FP   Path to the class associated file
```

Ejemplo de Implementación: mini-ImageNet

Para crear una versión reducida del dataset ImageNet-1K, el script gen_nano_imagenet.py permite al usuario especificar el directorio fuente (--src) donde se encuentran las imágenes originales, el número de archivos que se desea copiar por carpeta (--n), y opcionalmente, un directorio de destino (--dest) donde se almacenará el subconjunto generado. Adicionalmente, el parámetro --class_fp permite definir la ruta hacia un archivo que asocia las imágenes con sus respectivas clases, asegurando que la integridad semántica del conjunto de datos se mantenga en el subconjunto creado.

Adaptación a Otros Datasets

De manera similar, los scripts create_nano.py y refer_by_classes.py están diseñados para trabajar con los datasets de video Kinetics-400/600/700 y Something-Something-v2, respectivamente. Estos scripts permiten seleccionar un número específico de videos por clase, lo que facilita la creación de conjuntos de datos de video más pequeños y manejables, ajustados a las limitaciones de hardware o a los objetivos específicos de una investigación.

Por ejemplo, create_nano.py permite seleccionar el directorio fuente de los videos, definir un archivo .csv que lista los videos y sus clases, y elegir cuántos videos se incluirán en el subconjunto para cada clase. Este enfoque asegura que el investigador tenga un control total sobre la composición del dataset, adaptándolo fácilmente para experimentos en entornos de recursos limitados.

Resultados Experimentales

Los primeros adoptantes de nano-datasets son los autores del presente trabajo. Actualmente estamos trabajando en entrenar una versión de V-JEPA que pueda entrenarse en computadoras personales sin depender de grandes centros de datos. La tarea resultaría extremadamente compleja sin nano-datasets. Aunque los resultados del modelo de comprensión de video en desarrollado excede el propósito de la presente publicación, nano-datasets cumple con el propósito de diseño.

Aunque esta es la primera publicación de avances del proyecto, esperamos que la herramienta sea adoptada por otros grupos que trabajan en compresión de imágenes y videos. La herramienta es útil para facilitar la investigación y la colaboración de investigadores. Al mantener compatibilidad con datasets existentes, permite reproducir resultados de terceros y comparar cuantitativamente modelos de aprendizaje automático.

AJEA – Actas de Jornadas y Eventos Académicos de UTN
Universidad Tecnológica Nacional
30 y 31 de octubre de 2024
Libro de Actas de Resúmenes DOI: https://doi.org/10.33414/ajea.1846.2025
Enlace: https://rtyc.utn.edu.ar/index.php/ajea/article/view/1874

Conclusión y trabajo futuro

Hemos presentado nano-datasets, una herramienta que permite la creación de conjuntos de datos de vídeo e imágenes personalizados a partir de grandes conjuntos de datos públicos. La herramienta se publica como un repositorio de código abierto disponible libremente en Internet. Hemos usado nano-datasets en el desarrollo de nuevos modelos de aprendizaje supervisado.

Como trabajo futuro esperamos ampliar la cantidad de datasets públicos soportados. Esta característica le permitirá a la herramienta ganar popularidad en la comunidad científica. Además esperamos producir nuevas herramientas que permitan la creación y publicación de nuevos conjuntos de datos que faciliten evaluar tareas de aprendizajes no conocidas a la fecha.

Agradecimientos

Se agradece la participación en este proyecto de Santiago Aggio, Guillermo Friedrich y Adrián Rostagno, que por las limitaciones en la cantidad de autores no pudieron ser incluidos como tales.

Referencias

Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised learning: A succinct review. Archives of Computational Methods in Engineering, 30(4), 2761-2775.

Sharir, G., Noy, A., & Zelnik-Manor, L. (2021). An image is worth 16x16 words, what is a video worth?. arXiv preprint arXiv:2103.13915.

Ryali, C., Hu, Y. T., Bolya, D., Wei, C., Fan, H., Huang, P. Y., ... & Feichtenhofer, C. (2023, July). Hiera: A hierarchical vision transformer without the bells-and-whistles. In International Conference on Machine Learning (pp. 29441-29454). PMLR.

Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., ... & Ballas, N. (2024). Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision, 115, 211-252.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950.

Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., & Mei, T. (2020). Learning to localize actions from moments. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16 (pp. 137-154). Springer International Publishing.

Libro de Actas de Resúmenes DOI: https://doi.org/10.33414/ajea.1846.2025
Enlace: https://rtyc.utn.edu.ar/index.php/ajea/article/view/1874

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., ... & Memisevic, R. (2017). The" something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE international conference on computer vision (pp. 5842-5850).

Carreira, J., Noland, E., Hillier, C., & Zisserman, A. (2019). A short note on the kinetics-700 human action dataset. arXiv preprint arXiv:1907.06987.