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/1875

nano-JEPA: Una propuesta para posibilitar la interpretación de video usando computadoras personales

nano-JEPA: A Proposal to Enable the Video Understanding Using Personal Computers

Presentación: xx/10/2024

Adrián ROSTAGNO
Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca
arostag@frbb.utn.edu.ar

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

Joel ERMANTRAUT
Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca
arostag@frbb.utn.edu.ar

Guillermo R. FRIEDRICH
Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca
gfried@frbb.utn.edu.ar

Resumen

V-JEPA es un modelo de inteligencia artificial cuyo objetivo es comprender y predecir el contenido de vídeos. Utiliza un enfoque de aprendizaje autosupervisado; se pre-entrena con datos sin etiquetar y luego se adapta a tareas específicas. Aprende a predecir partes perdidas o enmascaradas de un vídeo, obligando al modelo a comprender y desarrollar una visión integral de la escena. Pretende desarrollar una inteligencia artificial que aprenda de manera similar a los humanos, formando modelos internos del mundo que les rodea para adaptarse y completar tareas de manera eficiente. Sin embargo, sus enormes demandas computacionales, que suelen requerir potentes clústeres de GPU, limitan la accesibilidad para muchos investigadores. Por ello se propone nano-JEPA, una adaptación de V-JEPA para ejecutarse en computadoras personales, incluso sin GPU. Se presenta además el repositorio de nano-conjuntos de datos (nano-datasets), que facilita la creación de subconjuntos manejables a partir de grandes conjuntos públicos de datos de vídeo. El objetivo es permitir una mayor participación y experimentación en la investigación con modelos similares a V-JEPA. Se pudo observar un rendimiento razonable de nano-JEPA en tareas posteriores, abriendo puertas para una mayor exploración e innovación.

Palabras clave: Predicción de Características, Aprendizaje No Supervisado, Representación Visual, Video, JEPA.

Abstract

V-JEPA is an artificial intelligence model whose objective is to understand and predict video content. Uses a self-supervised learning approach; It is pretrained on unlabeled data and then tailored to specific tasks. It learns by predicting missing or masked parts of a video, forcing the model to understand and develop a comprehensive view of the scene. It aims to develop artificial intelligence that learns in a similar way to

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/1875
humans, forming internal models of the world around them to adapt and complete tasks efficiently. However, their enormous computational demands, which often require powerful GPU clusters, limit accessibility for many researchers. Therefore, nano-JEPA, an adaptation of V-JEPA, is proposed to run on personal computers, even without GPU. The nano-dataset repository is also presented, which facilitates the creation of manageable subsets from large public video data sets. The goal is to enable greater participation and experimentation in research with models similar to V-JEPA. Reasonable performance of nano-JEPA could be observed in subsequent tasks, opening doors for further exploration and innovation.

Keywords: Feature Prediction, Unsupervised Learning, Visual Representations, Video, JEPA.

Introducción

La Arquitectura Predictiva de Incrustación Conjunta (JEPA, por sus siglas en inglés) ha surgido como un enfoque auto-supervisado prometedor que aprende utilizando un modelo del mundo (Assran y otros, 2023). Este enfoque se ha extendido a datos de video con la Arquitectura V-JEPA (Bardes y otros, 2024). V-JEPA se entrena para predecir la representación de una región enmascarada de un video a partir de la representación de la región no enmascarada. Este enfoque ha demostrado ser efectivo para aprender representaciones visuales a partir de video y ha superado a enfoques anteriores de aprendizaje de representaciones de video en evaluaciones congeladas en tareas de reconocimiento de acciones, detección de acciones espacio-temporales y clasificación de imágenes.

Sin embargo, entrenar V-JEPA requiere una gran cantidad de recursos computacionales. Por ejemplo, el modelo V-JEPA más grande se entrenó con 2 millones de videos durante 90.000 iteraciones con un tamaño de lote de 2.400. Esto requiere una gran cantidad de memoria y poder computacional de GPU (Unidad de Procesamiento Gráfico). Realizar avances en esta área del conocimiento requiere acceso a recursos computacionales costosos y puede ser restrictivo para un porcentaje significativo de investigadores. Además, la manipulación de los conjuntos de datos utilizados en los experimentos puede requerir un tiempo considerable para descargar y configurar.

En este trabajo se presenta nano-JEPA[1], un modelo compatible con V-JEPA que puede entrenarse en una sola computadora personal con recursos limitados. Puede entrenarse e inferir en CPU, así como utilizando GPUs. Todo el código fuente liberado es completamente compatible con el código fuente original de V-JEPA. Finalmente se demuestra que nano-JEPA puede entrenarse con un conjunto de datos de video mucho más pequeño y aún así lograr un rendimiento razonable en tareas posteriores. Tanto el modelo como el conjunto de datos se liberan como proyectos de código abierto.

Las contribuciones de este trabajo se pueden resumir en los siguientes aspectos: primero, es posible duplicar los resultados de la plataforma V-JEPA sin la necesidad de un clúster de computadoras equipadas con GPUs; segundo, es posible entrenar instancias personalizadas del modelo; tercero, entrenar múltiples instancias de modelos de visión que pueden abrir el camino para experimentar con arquitecturas de conjunto.

Trabajos relacionados

Recientemente, diversos grupos de investigación han hecho progresos en el aprendizaje de representaciones visuales a partir de video. Algunos de ellos se basaron en la arquitectura de transformación en visión (Dosovitskiy y otros, 2020). Los trabajos más relevantes utilizan auto-encoders enmascarados (Xie y otros, 2022) o agrupación de características basada en consultas (Oquab y otros, 2023). Aunque los resultados son prometedores, todos los métodos mencionados aprenden en el espacio de píxeles.

En contraste, la idea principal detrás de V-JEPA es más simple que los enfoques anteriores (Bardes y otros, 2024). El modelo aprende representaciones visuales en un espacio abstracto y utiliza enmascaramiento para predecir qué falta en la representación. Este enfoque tiene múltiples beneficios. Primero, la predicción de características es versátil para tareas de imagen y video posteriores. Luego, el esfuerzo de entrenamiento es significativamente menor que los modelos equivalentes que operan en el espacio de píxeles. Finalmente, la

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

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/1875

cantidad de etiquetas requeridas es menor en el caso de las arquitecturas JEPA.

V-JEPA es una versión refinada de la propuesta inicial de incrustación conjunta (LeCun, 2022). La arquitectura predice la salida, y dada una entrada x en un espacio abstracto (transformado por codificadores en un espacio de incrustación). Un predictor aprende la transformación utilizando una variable adicional z, que proporciona información sobre el proceso de cómputo. El entrenamiento auto-supervisado se basa en enmascarar parte de las entradas y minimizar el error de la predicción del sistema. Se utiliza Vision Transformer (ViT) (Arnab y otros, 2021) (Dosovitskiy y otros, 2020) como backbone de video.

Se ha demostrado que la arquitectura V-JEPA entrega resultados acordes al estado del arte utilizando conjuntos de datos públicos (Bardes y otros, 2024). Para lograr el objetivo, el modelo se entrenó en un sistema distribuido, equipado con GPUs NVIDIA A100 de 80 GB. El mejor rendimiento se logra utilizando las instancias más grandes de ViT. Lograr este nivel de resultados exige acceso a un sistema de cómputo similar. Este hecho es una de las motivaciones que inspiraron el presente trabajo.

El modelo nano-JEPA

Dos cuestiones motivaron el desarrollo de nano-JEPA. La primera fue proporcionar un modelo que cualquier investigador pueda entrenar en una computadora personal actual. La otra fue no perder la compatibilidad con V-JEPA. En particular, no se querían perder los parámetros relacionados con el entrenamiento del modelo. Para lograr eso, se modificó la base de código para eliminar los aspectos de cómputo distribuido de la implementación. Esto trajo como consecuencia la necesidad de tiempos de entrenamiento e inferencia más largos. Sin embargo, es posible ejecutar *checkpoint* de V-JEPA en nano-JEPA, si el sistema host tiene al menos una arquitectura de CPU actual y suficiente memoria RAM.

Además, nano-JEPA proporciona un conjunto de herramientas que ayudan al usuario a entrenar, inferir y visualizar fácilmente las características aprendidas por el modelo. Estas características ayudan a iterar rápidamente y a confirmar los resultados obtenidos después del entrenamiento. En términos de arquitectura, nada esencial se ha modificado en la versión actual del proyecto. Dado que esta es una de las primeras publicaciones sobre la iniciativa. Los resultados mencionados en la presente publicación se centran en mostrar que nano-JEPA se comporta como se esperaba a escala de un solo nodo de cómputo.

Hacer inferencia en un conjunto de datos de video puede no ser trivial cuando se utilizan modelos experimentales grandes y puede ser una barrera para un uso posterior de los modelos. Motivados por este hecho, se desarrolló una herramienta que infiere sobre un conjunto dado de videos. Antes de hacer inferencia, es necesario cargar el punto de control pre-entrenado así como el modelo downstream entrenado para la tarea específica. Otra herramienta que se proporciona dentro de la estructura es un script de visualización de características. Esto resulta de utilidad cuando se intenta entender qué está viendo un modelo auto-supervisado cuando se le proporciona una imagen.

Resultados experimentales

Los resultados presentados en esta sección tienen dos motivaciones: la primera es mostrar que nano-JEPA produce los mismos resultados que V-JEPA; la segunda es mostrar que nano-JEPA mantiene la compatibilidad con V-JEPA. Considerando que nano-JEPA fue creado como un entorno de trabajo para el desarrollo rápido en computadoras personales individuales, los resultados obtenidos respaldan ambas motivaciones.

Los ensayos realizados estuvieron enfocados en las etapas de pre-entrenamiento y evaluación; en este último caso se ensayó el modelo para la clasificación de imágenes, por una parte, y de videos por otra parte.

Pre-entrenamiento

Se entrenó un modelo base que usa ViT-T (tiny) cómo *background* visual y consumiendo 400 videos del dataset K400. Este proceso fue realizado en una máquina con 32 GB de RAM y procesador de 8 núcleos. En (Fig. 1) se puede observar que el proceso converge luego de 15 épocas. Se puede concluir que la pérdida (loss) del modelo se decrementa a medida que avanza el pre-entrenamiento. El modelo pre-entrenado fue utilizado para tareas posteriores, que se detallan en la sección de Evaluación.

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/1875

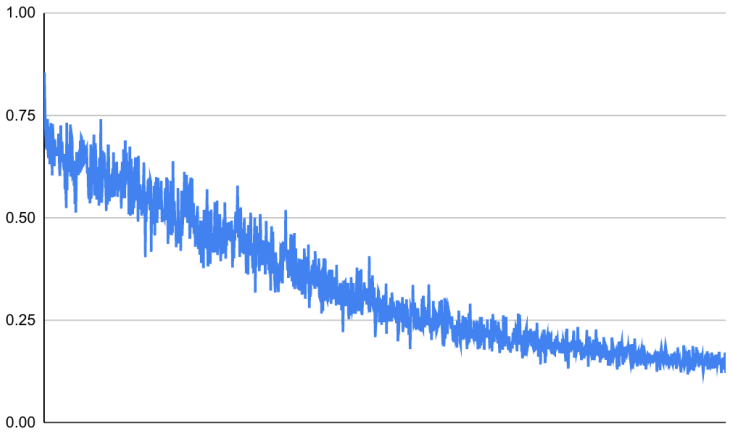

Figura 1: pre-entrenamiento ViT-T (tiny) usando 400 videos del dataset K400 dataset y 15 épocas.

Evaluación: Clasificación de Imágenes

Se utilizó ImageNet-1K como base de referencia para evaluar el rendimiento de nano-JEPA y V-JEPA en la tarea de clasificación de imágenes en 1000 clases. En este ensayo, se utilizaron ViT-T para nano-JEPA (tiny) y ViT-L (large) para V-JEPA. Los resultados se presentan en la Tabla 1, donde se comparan las precisiones en el conjunto de prueba (test-acc) y en el conjunto de entrenamiento (train-acc). Es lógico esperar que la precisión de nano-JEPA sea menor que la de V-JEPA para el mismo número de épocas, dado que nano-JEPA es un modelo más pequeño. Sin embargo, el entrenamiento de nano-JEPA avanza como se esperaba y puede ejecutarse en CPU o GPU.

Época	nano-JEPA ViT-T (tiny)				V-JEPA ViT-L (large)			
	CPU		GPU		CPU		GPU	
	train-acc	test-acc	train-acc	test-acc	train-acc	test-acc	train-acc	test-acc
1	2,678	7,954	2,487	7,954	7,334	18,181	7,397	16,477
2	3,635	3,409	3,125	7,386	19,196	28,977	20,535	30,681
3	5,293	6,818	5,612	6,818	28,252	31,818	26,849	35,795

Tabla 1: Clasificación de imágenes de ImageNet 1k.
ViT-T es usado en nano-JEPA y ViT-L es usado en el caso de V-JEPA.

Evaluación: Clasificación de Videos

Se entrenaron seis modelos usando la plataforma nano-JEPA. La mitad de los modelos usaron el checkpoint obtenido del entrenamiento con ViT-T y 400 videos de K400. La otra mitad usaron el checkpoint provisto por los autores de V-JEPA, usando ViT-L. La Tabla 3 muestra un resumen de los resultados.

Los modelos A, B y C fueron entrenados usando el checkpoint del pre-entrenamiento mediante nano-JEPA. Se presenta el *accuracy* y la época en la cual este fue obtenido. Como era de esperar, el *accuracy* es menor que para los modelos más grandes. Sin embargo, los resultados son consistentes y correlacionados con el uso de ViT-T. En el caso del modelo B, *accuracy* fue alcanzado más pronto. Esto tiene sentido, dado que el dataset de entrenamiento es de menor tamaño. En la Figura 2 se pueden observar más detalles.

Cuando se utilizó un checkpoint pre-entrenado provisto por V-JEPA, se pudo observar que los modelos alcanzan rápidamente alto *accuracy*. Esto es un indicador de que el dataset es pequeño para la cantidad de parámetros del modelo. Se puede decir que los modelos D, E y F están sobre ajustados para los datasets provistos. En la Figura 3 se pueden observar más detalles.

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/1875

Modelo	Clases	Videos de entrenamiento	Videos de evaluación	Ratio	Accuracy	Épocas	Pre-entrenamiento
A	4	100	50	0,5	45,50	20	nano-JEPA ViT-T
B	4	25	12	0,48	35,41	10	
C	8	100	50	0,5	41,50	20	
D	4	100	50	0,5	99,50	6	V-JEPA ViT-L
E	4	25	12	0,48	91,66	6	
F	8	100	50	0,5	94,25	6	

Tabla 2: Una colección de modelos entrenados usando nano-JEPA. Se usaron dos checkpoints: ViT-T con 400 K400 videos (entrenado mediante nano-JEPA) y ViT-L (entrenado mediante V-JEPA de sus autores). Se usó ViT-T en el caso de nano-JEPA y ViT-L en el caso de V-JEPA.

En la Figura 4 se presenta la visualización de características aprendidas por el modelo nano-JEPA pre-entrenado. Se puede observar que los colores están asociados a partes semánticas de la imagen. Esto confirma que el modelo aprende y que es un razonable proveedor de características (features) para tareas más complejas tales como clasificación de acciones en video.

Figura 2: Entrenamiento de los modelos A, B y C (el eje de ordenadas corresponde al accuracy).

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/1875

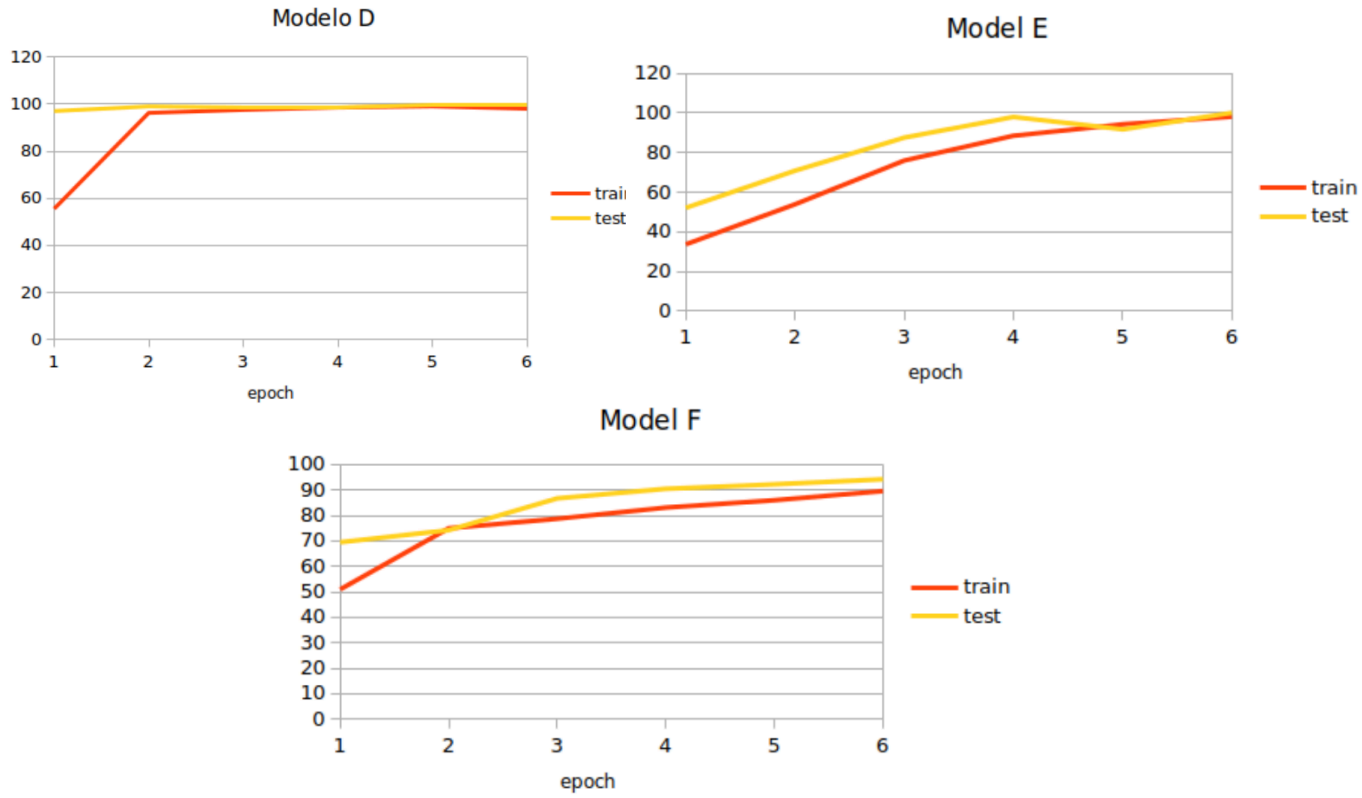

Figura 3: Entrenamiento de los modelos D, E y F (el eje de ordenadas corresponde al accuracy).

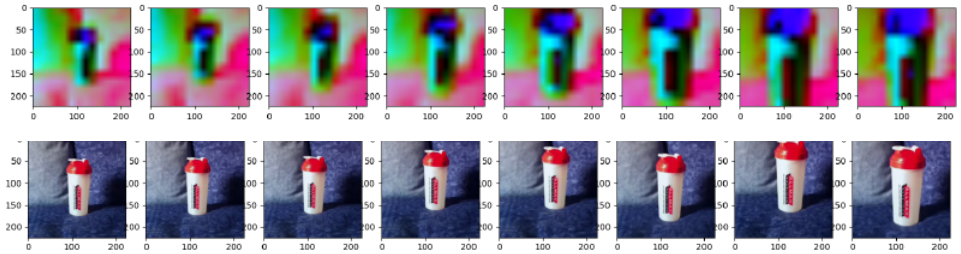

Figura 4: Características de nano-JEPA

Conclusiones y futuros pasos

La plataforma presentada, nano-JEPA, permite el entrenamiento e inferencia sobre V-JEPA sin la necesidad de contar con una infraestructura de computación distribuida. La plataforma es compatible con *checkpoints* pre-entrenados de V-JEPA y pueden producir nuevos modelos para tareas posteriores.

Los resultados presentados muestran cómo funciona el proceso de entrenamiento y cómo se degrada el *accuracy* a medida que se reduce la cantidad de parámetros del modelo. Sin embargo, se considera que teniendo una plataforma que permite la rápida experimentación en una computadora personal permitiría la realización de futuras investigaciones de interés. Adicionalmente se desarrolló un conjunto de herramientas que simplifican el proceso de inferencia y la visualización de características.

Los avances presentados son los primeros pasos en el camino acerca del aprendizaje no supervisado en el espacio embebido. Como trabajo futuro, se pretende explorar cómo podría funcionar un ensamble a gran escala de nano-JEPA realizando votaciones para alcanzar resultados (Sagi y Rokach, 2018). En base a avances recientes en el ensamble de arquitecturas para el procesamiento de lenguajes naturales (Jia y otros, 2023), se supone que esta investigación puede producir resultados relevantes. Otro camino que se pretende explorar es

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/1875

cómo una infraestructura nano-JEPA pre-entrenada puede ser utilizada en una arquitectura de múltiples cabezas. Se pretende evaluar algunos resultados prometedores que hay en esta área (Goldblum y otros, 2024). Por último, se tiene planeado explorar arquitecturas alternativas de *vision transformer*. Un posible candidato es ConnTransformer (Liu y otros, 2020). Se espera obtener beneficios en términos de *accuracy* y tiempos de entrenamiento e inferencia.

Agradecimientos

Se agradece la participación en este proyecto a Santiago Aggio, Lucas Tobio y Segundo Foissac, que por las limitaciones en la cantidad de autores no pudieron ser incluidos como tales.

Referencias

Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C. (2021). Vivit: A video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6836–6846.

Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N. (2023). Self-supervised learning from images with a joint-embedding predictive architecture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15619–15629 (2023).

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

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

Goldblum, M., Souri, H., Ni, R., Shu, M., Prabhu, V., Somepalli, G., Chattopadhyay, P., Ibrahim, M., Bardes, A., Hoffman, J., et al. (2024). Battle of the backbones: A large-scale comparison of pretrained models across computer vision tasks. Advances in Neural Information Processing Systems 36.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al. (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.

Jia, J., Liang, W., Liang, Y. (2023). A review of hybrid and ensemble in deep learning for natural language processing. arXiv preprint arXiv:2312.05589.

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

LeCun, Y. (2022). A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. Open Review 62(1), 1–62.

Liu, Z., Luo, S., Li, W., Lu, J., Wu, Y., Sun, S., Li, C., Yang, L. (2020). Convtransformer: A convolutional transformer network for video frame synthesis. arXiv preprint arXiv:2011.10185.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. (2023). Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. International journal of computer vision 115, 211–252.

Sagi, O., Rokach, L. (2018). Ensemble learning: A survey. Wiley interdisciplinary reviews: data mining and knowledge discovery 8(4), e1249.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., Hu, H. (2022). Simmim: A simple framework for masked image modeling. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9653–9663.