# pico-JEPA: Comprendiendo el Video con Modelos Ultra-Ligeros y la Sabiduría Colectiva

Adrián Rostagno, Javier Iparraguirre, Guillermo Friedrich, Santiago Aggio, Roberto Briatore, Lucas Tobio, and Diego Coca

Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca,
11 de Abril 461, 8000 Bahía Blanca, Argentina
{arostag,jiparraguirre,gfried,slaggio}@frbb.utn.edu.ar
{roberbriatore11,lucasltobio,dcoca28}@gmail.com
http://www.frbb.utn.edu.ar

**Resumen** Basándose en nuestro trabajo previo con nano-JEPA, presentamos pico-JEPA, una adaptación ultraligera de la Joint Embedding Predictive Architecture (JEPA) diseñada para dispositivos con recursos significativamente limitados. Mientras que nano-JEPA demostró la viabilidad de entrenar modelos tipo V-JEPA en ordenadores personales, pico-JEPA va más allá al crear modelos lo suficientemente pequeños como para permitir arquitecturas de conjuntos a través de mecanismos de votación colaborativa.

Presentamos un marco integral donde múltiples modelos pico-JEPA, cada uno entrenado en diferentes subconjuntos de datos o con configuraciones arquitectónicas variadas, resuelven colectivamente tareas de clasificación de video mediante estrategias de votación inteligentes. Nuestro enfoque aborda el compromiso fundamental entre el tamaño del modelo y el rendimiento, aprovechando el principio de la sabiduría de las multitudes en el aprendizaje profundo.

Los resultados experimentales obtenidos sobre el conjunto de datos Kinetics-700, demuestran que 4 modelos pico-JEPA pueden lograr un rendimiento competitivo, en comparación con modelos únicos, manteniendo la capacidad de ejecutarse en hardware estándar. Este trabajo abre nuevas posibilidades para la comprensión de video distribuida en escenarios de computación de borde y proporciona un camino para el aprendizaje colaborativo en entornos con recursos limitados. Finalmente, el proyecto se publica en un repositorio abierto con el fin de acelerar la colaboración entre investigadores interesados.

**Keywords:** self-supervised learning, ensemble learning, video classification, lightweight models, distributed computing, edge AI

## 1. Introducción

La arquitectura JEPA (Joint Embedding Predictive Architecture) se ha consolidado como un prometedor enfoque autosupervisado que aprende representaciones del mundo en un espacio latente [1]. Este enfoque se ha extendido a los datos de vídeo con la arquitectura V-JEPA (Video Joint Embedding Predictive Architecture) [3]. El modelo

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 eficaz para el aprendizaje de representaciones visuales a partir de vídeo y ha superado a los enfoques previos de aprendizaje de representaciones a nivel de píxeles.

El entrenamiento de V-JEPA requiere una enorme cantidad de recursos computacionales. Por ejemplo, el modelo V-JEPA más grande se entrenó con 2 millones de videos para 90.000 iteraciones, con un tamaño de lote de 2.400 videos. Esto implica contar con una gran cantidad de GPUs (Unidad de Procesamiento Gráfico), procesadores de propósito general (CPUs) y de espacio de memoria asociado. Típicamente estos modelos se entrenan en sistemas distribuidos interconectados por redes de alto desempeño (clusters). Avanzar en esta área de conocimiento requiere disponer de acceso a recursos computacionales costosos y puede ser restrictivo para la mayoría de los investigadores.

Es posible encontrar en la literatura trabajos que implementan el paradigma en sistemas de cómputos con recursos limitados. Un ejemplo es nano-JEPA [7]. Aunque los resultados obtenidos con nano-JEPA fueron prometedores, sigue siendo un modelo único. En este trabajo exploramos la idea de poder utilizar modelos aún más pequeños, inspirados en "ensemble learning" y sabiduría de masas, que nos permita avanzar hacia una democratización más completa de la comprensión de videos.

Presentamos pico-JEPA, una adaptación ultraligera de V-JEPA, diseñada para el aprendizaje conjunto en dispositivos con recursos muy limitados. Si bien nano-JEPA demostró la viabilidad del entrenamiento de modelos similares a V-JEPA en ordenadores personales, pico-JEPA pretende ampliar los límites al crear modelos lo suficientemente pequeños como para permitir arquitecturas de conjunto mediante mecanismos de votación colaborativa [5, 6, 8]. Además, el código fuente del proyecto se publica en un repositorio abierto con el fin de acelerar la colaboración entre investigadores interesados [1].

La siguiente sección del artículo describe los trabajos relacionados. El marco propuesto se describe en la Sección 3. La Sección 4 documenta los resultados. Los pasos futuros se mencionan en la Sección 5. Finalmente, las conclusiones se exponen en la Sección 6.

## 2. Trabajo Relacionado

El campo del aprendizaje auto-supervisado para la comprensión de vídeo ha experimentado un rápido avance, impulsado por arquitecturas capaces de aprender representaciones robustas del mundo sin necesidad de etiquetas explícitas. Nuestro trabajo se sitúa en la intersección de tres áreas de investigación fundamentales: las arquitecturas predictivas en espacios latentes, la democratización de modelos de gran escala y el aprendizaje por ensamblaje.

La base conceptual de nuestro trabajo es la Arquitectura Predictiva de Incrustación Conjunta (JEPA), propuesta inicialmente para imágenes (I-JEPA) y posteriormente extendida con gran éxito al dominio del vídeo (V-JEPA) [3]. A diferencia de los

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

métodos generativos que buscan reconstruir cada píxel de una entrada enmascarada, las arquitecturas JEPA operan en un espacio de representaciones latentes. Su objetivo es predecir la representación de una porción oculta de los datos (por ejemplo, un conjunto de parches de un vídeo) a partir de la porción visible. Este enfoque permite al modelo aprender una comprensión más abstracta y semántica del contenido, ignorando detalles irrelevantes y demostrando una eficiencia superior en el aprendizaje de características. Investigaciones posteriores, como V-JEPA 2, han demostrado la escalabilidad de esta arquitectura para crear modelos del mundo capaces de comprender, predecir e incluso planificar [2]. Sin embargo, un denominador común de estas implementaciones de vanguardia es su enorme costo computacional, que requiere clusters de GPUs a gran escala y vastos conjuntos de datos, limitando su accesibilidad para gran parte de la comunidad investigadora.

Paralelamente, el aprendizaje por ensamblaje (ensemble learning) se ha consolidado como una de las técnicas más efectivas para mejorar el rendimiento y la robustez de los modelos de aprendizaje automático [5, 6, 8]. El principio fundamental, a menudo denominado "sabiduría de las multitudes", consiste en combinar las predicciones de múltiples modelos base para obtener una decisión final más precisa que la de cualquiera de los modelos individuales. Estrategias como Bagging, Boosting y Stacking han demostrado su eficacia en una amplia gama de tareas. No obstante, la aplicación de estas técnicas a modelos de aprendizaje profundo, especialmente en el dominio del vídeo, es a menudo prohibitiva. El costo de entrenar y mantener múltiples redes neuronales profundas de gran tamaño hace que la creación de ensamblajes sea computacionalmente inviable en la mayoría de los escenarios prácticos.

pico-JEPA se diferencia de los trabajos anteriores al abordar directamente la tensión entre la potencia de las arquitecturas JEPA y la viabilidad del aprendizaje por ensamblaje. Mientras que V-JEPA y sus sucesores se centran en escalar hacia modelos cada vez más grandes para maximizar el rendimiento, nuestro trabajo explora la dirección opuesta: la miniaturización estratégica.

Partiendo de nuestro trabajo previo, nano-JEPA [7], que ya demostró la factibilidad de entrenar un modelo de tipo JEPA en un ordenador personal, pico-JEPA va un paso más allá. No se trata simplemente de una versión más pequeña de V-JEPA, sino de un rediseño fundamental con un propósito específico: ser una unidad base para ensamblajes. Su novedad no reside en proponer un nuevo mecanismo de aprendizaje, sino en habilitar la aplicación de un paradigma de aprendizaje bien establecido (el ensamblaje) a una arquitectura auto-supervisada de última generación.

Por lo tanto, pico-JEPA se posiciona como un puente entre dos mundos: el de los modelos fundacionales de vídeo, potentes pero no siempre disponibles, y el de las técnicas de ensamblaje, efectivas pero computacionalmente demandantes. Al crear un modelo "ultra-ligero", pico-JEPA permite, por primera vez, explorar de manera práctica cómo la combinación de múltiples modelos del mundo, entrenados de forma auto-supervisada, puede colaborar para resolver tareas complejas de clasificación de vídeo en entornos con recursos limitados.

## 3. Diseño de pico-JEPA

El diseño de pico-JEPA se concibió con el propósito fundamental de crear un marco de trabajo para el aprendizaje auto-supervisado que fuera a la vez minimalista en su consumo de recursos y completo en su funcionalidad. La motivación principal fue doble: en primer lugar, democratizar el acceso a la investigación con arquitecturas de tipo JEPA, eliminando la barrera que imponen los altos requisitos computacionales de modelos a gran escala. En segundo lugar, desarrollar un modelo base lo suficientemente ligero y eficiente como para poder ser integrado en futuras investigaciones sobre aprendizaje por ensamblaje (ensemble learning), donde múltiples modelos deben ser entrenados y ejecutados de forma concurrente.

La filosofía de diseño se centró en tres pilares clave: modularidad, configurabilidad y fidelidad a los principios teóricos de JEPA. Se buscó un equilibrio entre la simplicidad, para facilitar la comprensión y modificación, y la implementación de técnicas de entrenamiento avanzadas para asegurar la robustez y efectividad del aprendizaje.

La arquitectura del sistema se estructura en torno a una separación lógica de responsabilidades. Por un lado, se encuentra el núcleo del modelo, que encapsula la lógica del aprendizaje predictivo en el espacio latente. Este núcleo está compuesto por un codificador principal basado en la arquitectura Vision Transformer (ViT), adaptado para procesar datos de vídeo mediante la extracción de máscaras espacio-temporales. Este codificador, denominado online encoder, es el componente que se entrena activamente.

Para generar los objetivos de aprendizaje, se utiliza un segundo codificador, el target encoder. Este componente no aprende a través de retropropagación, sino que sus pesos se actualizan lentamente como una media móvil exponencial (EMA) de los pesos del codificador online. Esta técnica proporciona un objetivo de predicción estable, evitando los colapsos que podrían ocurrir en un esquema de auto-predicción directa. Finalmente, un predictor, implementado como un decodificador Transformer más superficial, tiene la tarea de predecir la representación latente del target encoder a partir de la representación del encoder online.

El manejo de datos se diseñó para ser eficiente y desacoplado del modelo. Se implementó un cargador de datos capaz de procesar vídeos de manera acelerada, aplicando un protocolo de preprocesamiento estandarizado que incluye muestreo de fotogramas, redimensionamiento y normalización. Esta estandarización asegura que el modelo reciba datos consistentes, independientemente de la variabilidad de las fuentes de vídeo originales.

Para el entrenamiento, aunque el framework es minimalista, no se escatimó en la implementación de técnicas de optimización modernas. Se utiliza el optimizador AdamW junto con un planificador de tasa de aprendizaje de tipo Cosine Annealing que incluye una fase inicial de calentamiento (o warmup en inglés). Este enfoque ha demostrado ser crucial para entrenar modelos de tipo Transformer de manera estable, evitando divergencias tempranas y permitiendo una convergencia más refinada hacia el final del entrenamiento. Adicionalmente, se emplea el recorte de gradientes (gradient clipping) para prevenir la inestabilidad numérica.

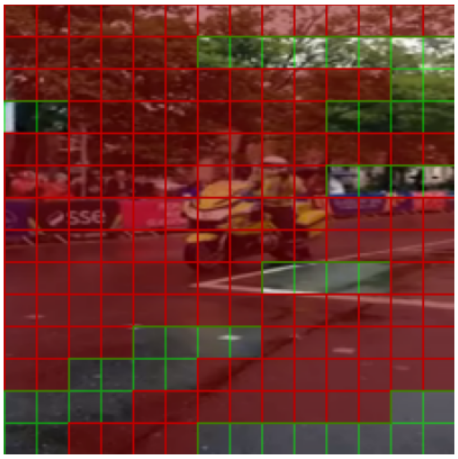

Figura 1: Visualización de oclusiones en pico-JEPA.

En resumen, la construcción de pico-JEPA es un ejercicio de ingeniería enfocado en la eficiencia y la accesibilidad. El resultado es un sistema autocontenido y robusto que, si bien es significativamente menos exigente que sus contrapartes a gran escala, preserva intactos los mecanismos fundamentales que hacen de la arquitectura JEPA un enfoque tan prometedor para el aprendizaje de representaciones de vídeo. La Figura 1 muestra un ejemplo donde las baldosas rojas están ocluidas y las verdes se exponen al modelo. Esta herramienta está disponible para los investigadores interesados.

La Figura 2 muestra un diagrama de un conjunto de modelos pico-JEPA como un ensamble. El resultado de la inferencia se basa en la votación de la mayoría. En el caso de los experimentos reportados en este trabajo, todos los individuos del conjunto comparten el mismo modelo pre-entrenado. Sin embargo, solo se entrena el clasificador de cada individuo con un subconjunto de datos. En este ejemplo, se muestra en la parte superior de la figura un modelo *general*, y en la parte inferior, un ensamble de 2 *individuos*.

## 4. Resultados

Para evaluar la eficacia de nuestro enfoque, se llevó a cabo una serie de experimentos centrados en una tarea de clasificación de acciones de vídeo. El objetivo principal era comparar el rendimiento de un modelo único entrenado con la totalidad de los datos disponibles frente a un ensamblaje de modelos pico-JEPA más pequeños, cada uno entrenado con un subconjunto de los datos.

El procesamiento se realizó en un equipo con procesador Intel i7, 64GB de memoria RAM y dos GPU NVIDIA 1060 de 6 GB de RAM cada una. Aunque se han utilizado aceleradores, es posible ejecutar pico-JEPA en arquitecturas con procesadores de propósito general. Desde el inicio del diseño del modelo, se ha contemplado la posibilidad de no depender de GPUs.

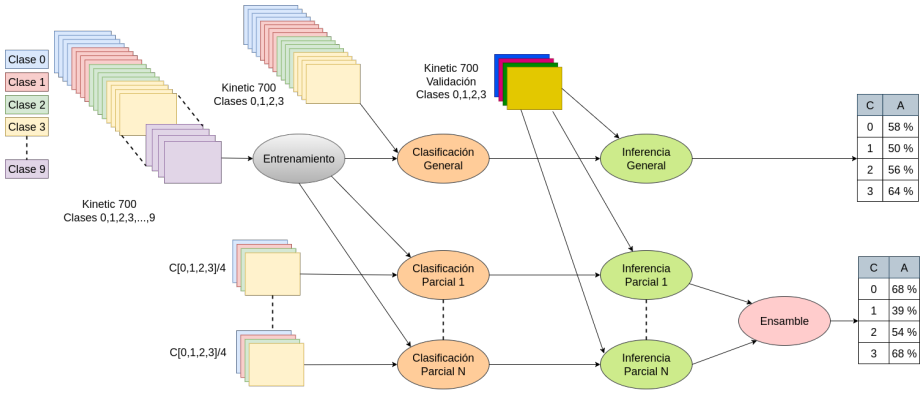


Figura 2: pico-JEPA como ensamble de modelos.

Respecto a los datos, se eligieron las 10 primeras clases del dataset Kinetics-700 [4] para realizar el pre-entrenado. Luego, se eligieron 4 entre las 10 primeras para realizar las comparaciones entre los modelos individuales y el ensamble en general. Estos son los primeros resultados reportados para arquitectura. Entendemos que es necesaria una evaluación más extensiva con más clases y otros conjuntos de datos. Sin embargo, basta con las evaluaciones reportadas para confirmar la hipótesis sobre los beneficios de la arquitectura propuesta.

Se definieron varias instancias de modelos, como se detalla en la Tabla 1. Se entrenó un modelo "General" utilizando la totalidad de los vídeos de pre-entrenamiento y clasificación. Paralelamente, el conjunto de datos se dividió en cuatro subconjuntos para entrenar cuatro modelos "Parciales" (Parcial 1 a 4). Esta configuración nos permite simular un escenario en el que múltiples agentes ligeros aprenden de diferentes porciones de la experiencia total disponible. La evaluación se realizó sobre un conjunto de validación común para todas las instancias.

Tabla 1: Instancias de pico-JEPA a evaluar.

| **Modelos Evaluados** | | | | |
|---|---|---|---|---|
| **Modelo** | **Embedding** | **Videos Pre-entrenamiento** | **Videos Clasificación** | **Videos Validación** |
| General | 192 | 7203 | 2817 | 194 |
| Parcial 1 | | | 706 | |
| Parcial 2 | | | 707 | |
| Parcial 3 | | | 707 | |
| Parcial 4 | | | 697 | |

Los resultados de la clasificación para cuatro clases del dataset Kinetics 700 se presentan en la Tabla 2. El modelo "General", que tuvo acceso a todos los datos de entrenamiento, alcanzó una precisión promedio del 57.15 %. Los modelos parciales,

como era de esperar, mostraron un rendimiento variable, con precisiones promedio que oscilan entre el 50.42 % y el 55.08 %, reflejando la diversidad y la limitación de los datos con los que fueron entrenados.

El hallazgo más significativo de nuestro estudio se observa en la columna "Votación Mayoría". Al combinar las predicciones de los cuatro modelos parciales mediante un esquema de votación simple, el ensamblaje alcanzó una precisión promedio del 57.63 %. Este resultado es notable por dos razones.

En primer lugar, se logra superar el modelo "General". El ensamblaje de modelos no solo fue competitivo, sino que logró un rendimiento ligeramente superior al del modelo único "General", a pesar de que cada componente individual del ensamblaje fue entrenado con solo una fracción de los datos.

Segundo, es posible observar robustez y consistencia. Aunque en clases específicas como "archaeological excavation" un modelo parcial individual ("Parcial 1") superó al ensamblaje, el método de votación demostró ser más robusto en promedio al compensar el rendimiento más bajo de algunos modelos (p. ej., "Parcial 2"). En resumen, el ensamblaje supera al modelo general en términos de desempeño.

Tabla 2: Resultados sobre 4 clases del dataset k700.

| **Porcentaje Aciertos (Accuacy)** | | | | | | |
|---|---|---|---|---|---|---|
| **Clase (K700)** | **General** | **Parcial 1** | **Parcial 2** | **Parcial 3** | **Parcial 4** | **Votación Mayoría (1 a 4)** |
| abseiling | 58.33 % | 52.08 % | 52.08 % | 66.67 % | 68.75 % | 68.75 % |
| adjusting_glasses | 50.00 % | 39.58 % | 37.92 % | 39.58 % | 33.33 % | 39.58 % |
| alligator_wrestling | 56.25 % | 52.08 % | 47.67 % | 52.08 % | 52.08 % | 54.17 % |
| archaeological_excavation | 64.00 % | 70.00 % | 64.00 % | 62.00 % | 60.00 % | 68.00 % |
| Promedio 4 clases | 57.15 % | 53.44 % | 50.42 % | 55.08 % | 53.54 % | 57.63 % |

La Tabla 3 muestra el nivel de certeza reportado por los modelos de clasificación. En el caso del modelo general, se reportan mejores valores. En los casos de los modelos parciales se ven niveles de certeza menores. Sin embargo, la inferencia a partir del ensamble logra mejores resultados de clasificación a pesar de la menor confianza de los individuos.

Tabla 3: certeza de los modelos de clasificación.

| **Porcentaje certeza** | | | | | |
|---|---|---|---|---|---|
| **Clase** | **General** | **Parcial 1** | **Parcial 2** | **Parcial 3** | **Parcial 4** |
| abseiling | 52.43 % | 37.30 % | 39.15 % | 37.76 % | 37.61 % |
| adjusting_glasses | 60.36 % | 40.76 % | 33.95 % | 40.06 % | 42.48 % |
| alligator_wrestling | 63.68 % | 42.36 % | 39.29 % | 41.75 % | 45.74 % |
| archaeological_excavation | 60.60 % | 46.05 % | 38.47 % | 41.43 % | 42.45 % |

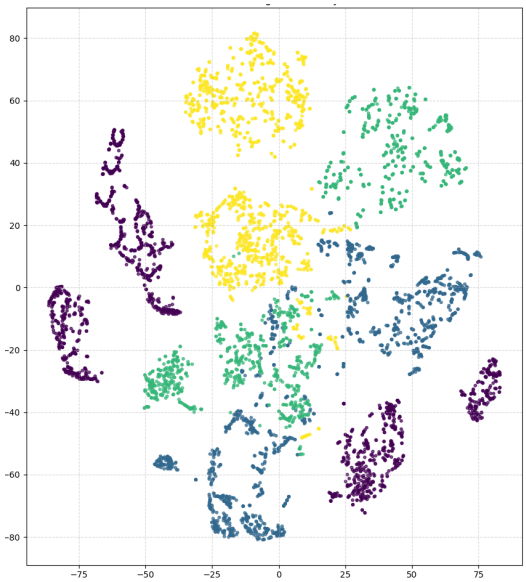


Figura 3: Visualización de características (o features) de pico-JEPA. En este caso se evalúa el individuo 1 del ensamble (Parcial 1). El color violeta representa la clase *abseiling*, el verde *alligator_wrestling*, el amarillo a *archaeological_exavaction*, y el azul a *adjunting_glasses*.

La Figura 3 muestra un mapa de como los puntos latentes (embeddings) se relacionan con las clases aprendidas. La gráfica del extractor de características no solo es una representación visual, sino una herramienta de diagnóstico. El modelo Parcial 1 ha aprendido a crear un espacio de características donde las distancias entre los puntos representan con precisión la facilidad o dificultad de la clasificación. En la gráfica t-SNE, los clusters de las clases “abseiling(violeta)”, “alligator_wrestling(verde)” y “archaeological_exavaction (amarillo)” están extremadamente bien definidos, son compactos y no se solapan. La clase “adjunting_glasses (azul)” es la única que muestra una menor cohesión y se encuentra en una ligera proximidad al cluster de la clase “alligator_wrestling”. Lo que muestra que los videos de la “adjunting_glasses” son, en términos de sus características internas, los más ambiguos y más propensos a ser confundidos con los de la clase “alligator_wrestling”. Además, nos permite validar con los resultados de este modelo en la tabla de aciertos. Esta herramienta de extractor de características está disponible en el código fuente.

## 5. Trabajo futuro

Los resultados obtenidos con el ensamblaje de modelos pico-JEPA mediante votación por mayoría son prometedores y validan la hipótesis de que la colaboración de múltiples agentes ligeros puede igualar o superar el rendimiento de un modelo

monolítico. Sin embargo, este es solo el primer paso. Proponemos varias líneas de investigación futuras para expandir y refinar este marco de trabajo.

Una de las extensiones más directas es la exploración de mecanismos de ensamblaje más sofisticados. Si bien la votación por mayoría es simple y efectiva, técnicas como el Stacking (o generalización apilada) ofrecen un potencial considerable. En lugar de tratar cada voto con igual peso, se podría entrenar un "meta-modelo" que aprenda a combinar de manera óptima las predicciones de los clasificadores pico-JEPA individuales. Este meta-modelo, que podría ser desde una regresión logística hasta una red neuronal superficial, puede utilizar las predicciones de los modelos base como sus características de entrada, permitiéndole aprender a ponderar la *opinión* de cada experto según el contexto de la entrada, potencialmente mejorando la precisión final del ensamblaje.

Otra área crucial para la investigación es el incremento de la diversidad dentro del ensamblaje. Actualmente, la diversidad se logra entrenando cada modelo en un subconjunto diferente de los datos. Para potenciar aún más el principio de la "sabiduría de las multitudes", se pueden introducir fuentes adicionales de variabilidad. Proponemos experimentar con ensamblajes heterogéneos, donde cada modelo pico-JEPA no solo se entrene con datos distintos, sino que también posea ligeras variaciones arquitectónicas (por ejemplo, diferente profundidad, número de cabezales de atención o tamaño de parche) o sea entrenado con distintos hiper parámetros. Esta heterogeneidad podría forzar a los modelos a aprender representaciones complementarias del mundo, enriqueciendo la capacidad colectiva del ensamblaje.

La utilidad práctica de pico-JEPA radica en su capacidad para democratizar la comprensión de video en entornos con recursos limitados, permitiendo el despliegue de modelos de IA en dispositivos de borde. Priorizando aplicaciones que usan votación, pico-JEPA facilita la creación de sistemas de IA colaborativos donde múltiples modelos ultraligeros, entrenados en diferentes subconjuntos de datos o con variaciones arquitectónicas, combinan sus predicciones a través de mecanismos de votación. Esto resulta en un rendimiento competitivo o superior al de modelos monolíticos, incluso en escenarios como la vigilancia inteligente, la robótica, el análisis de video móvil y la investigación de IA accesible, donde la eficiencia y la robustez son cruciales.

Finalmente, el marco de pico-JEPA abre la puerta a la investigación en aprendizaje distribuido y en el borde (edge computing). Se podría explorar la viabilidad de entrenar y ejecutar estos ensamblajes en una red de dispositivos con recursos limitados, donde cada nodo alberga un único modelo pico-JEPA. Esto implicaría investigar no solo los algoritmos de votación, sino también los protocolos de comunicación eficientes para compartir predicciones y, potencialmente, gradientes, abriendo el camino hacia sistemas de inteligencia artificial colaborativa y descentralizada.

## 6. Conclusiones

Este trabajo ha presentado pico-JEPA, un framework ultraligero que demuestra la viabilidad de aplicar técnicas de aprendizaje por ensamblaje a modelos de comprensión de vídeo auto-supervisados en entornos con recursos computacionales

limitados. Nuestros resultados experimentales validan la hipótesis central: un ensamblaje de múltiples modelos pico-JEPA, cada uno entrenado con una fracción de los datos, puede alcanzar e incluso superar el rendimiento de un modelo monolítico único entrenado con la totalidad de los datos. Este hallazgo subraya el poder de la diversidad inducida por los datos en el aprendizaje por ensamblaje y confirma que la "sabiduría de las multitudes" es un principio aplicable y beneficioso incluso en el contexto de arquitecturas de aprendizaje profundo complejas.

La contribución fundamental de pico-JEPA no es solo la miniaturización de una arquitectura de vanguardia, sino la habilitación de un nuevo paradigma de investigación. Al reducir drásticamente la barrera computacional de entrada, nuestro framework abre la puerta a la exploración de sistemas de inteligencia artificial colaborativos y distribuidos, donde múltiples agentes ligeros pueden aprender de manera independiente y cooperar para lograr una comprensión del mundo más robusta y generalizable. pico-JEPA se posiciona, por tanto, como una herramienta valiosa para la democratización de la investigación en el aprendizaje de representaciones de vídeo y como un punto de partida para futuros trabajos en ensamblajes de modelos del mundo.

## Referencias


1. Assran, M., Duval, Q., Misra, I., Bojanowski, P., Vincent, P., Rabbat, M., LeCun, Y., Ballas, N.: 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)
2. Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., Rizvi, A., Roberts, C., Sinha, K., Zholus, A., et al.: V-jepa 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985 (2025)
3. Bardes, A., Garrido, Q., Ponce, J., Chen, X., Rabbat, M., LeCun, Y., Assran, M., Ballas, N.: Revisiting feature prediction for learning visual representations from video. arXiv preprint arXiv:2404.08471 (2024)
4. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
5. Khan, A.A., Chaudhari, O., Chandra, R.: A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. Expert Systems with Applications 244, 122778 (2024)
6. Mohammed, A., Kora, R.: A comprehensive review on ensemble deep learning: Opportunities and challenges. Journal of King Saud University-Computer and Information Sciences 35(2), 757–774 (2023)
7. Rostagno, A., Iparraguirre, J., Ermantraut, J., Tobio, L., Foissac, S., Aggio, S., Friedrich, G.R.: nano-jepa: Democratizing video understanding with personal computers. In: XXX Congreso Argentino de Ciencias de la Computación (CACIC)(La Plata, 7 al 11 de octubre de 2024) (2024)
8. Sagi, O., Rokach, L.: Ensemble learning: A survey. Wiley interdisciplinary reviews: data mining and knowledge discovery 8(4), e1249 (2018)