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 Diego Coca

Universidad Tecnológica Nacional, Facultad Regional Bahía Blanca
{arostag,jiparraguirre,gfried,slaggio}@frbb.utn.edu.ar
{roberbriatore11,lucasltobio,dcoca28}@gmail.com
https://www.frbb.utn.edu.ar/
https://github.com/BHI-Research/pico-jepa

El Problema: La Barrera Computacional

- **Contexto:** Los modelos de comprensión de video como V-JEPA son muy potentes y aprenden de forma auto-supervisada.
- **El Desafío:** Su entrenamiento requiere enormes recursos computacionales (clusters de GPUs, grandes datasets), limitando el acceso para la mayoría de los investigadores.
- **La Pregunta Clave:** ¿Podemos democratizar la investigación en esta área y lograr resultados competitivos sin acceso a supercomputadoras?

Nuestra Propuesta: La Sabiduría Colectiva

- **Idea Principal:** En lugar de un modelo gigante, proponemos usar un ensamble de modelos ultra-ligeros.
- **Presentamos pico-JEPA:** Una arquitectura diseñada para ser una "unidad base" para ensambles, lo suficientemente pequeña para permitir arquitecturas colaborativas.
- **Hipótesis Central:** Un conjunto de modelos pico-JEPA, entrenados con porciones de datos, puede superar a un único modelo más grande entrenado con todos los datos.
- **Código abierto:** `https://github.com/BHI-Research/pico-jepa`

utn bhi

Referencia

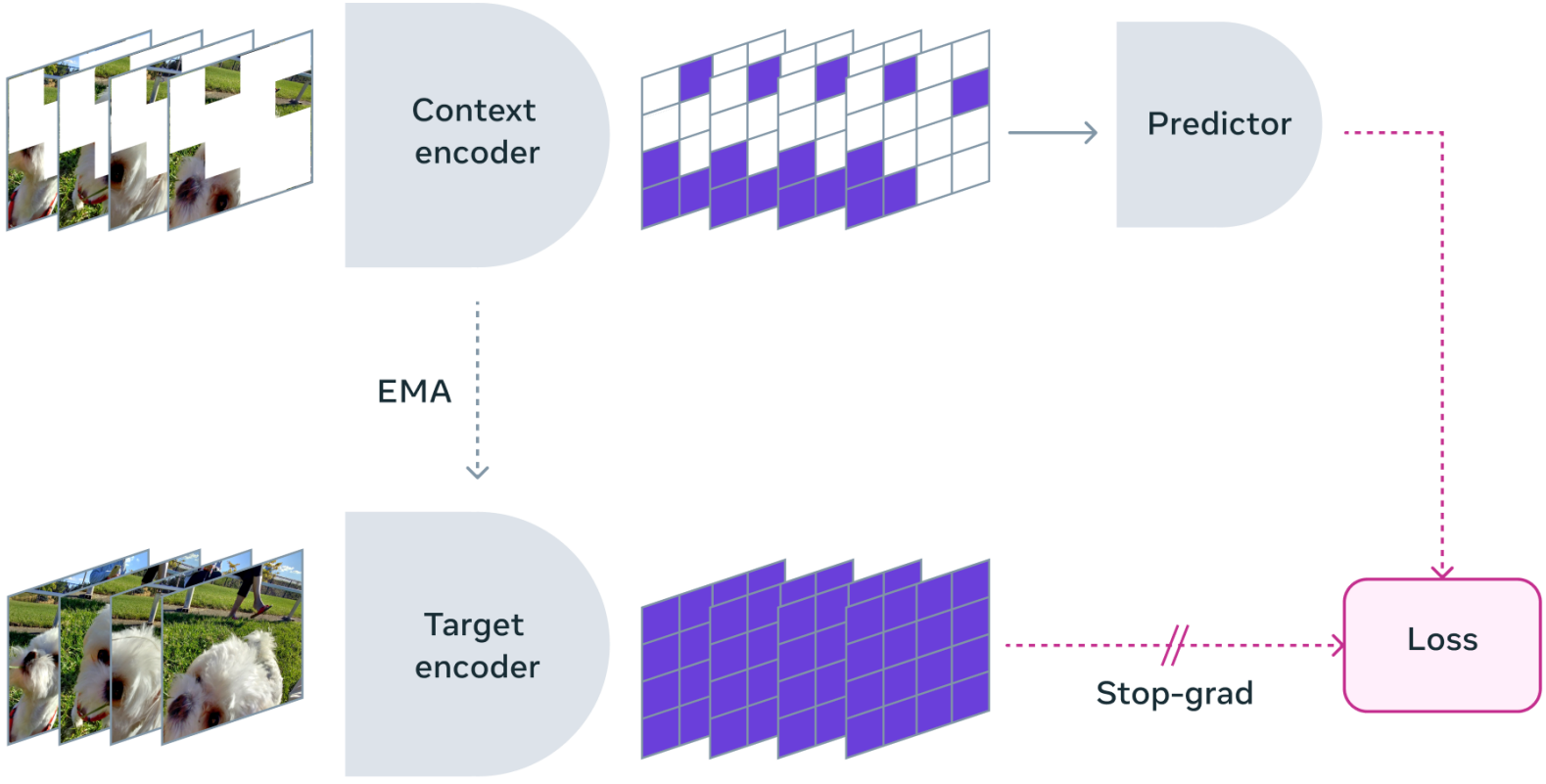

Figura: Arquitectura v-JEPA, imagen cortesía de meta.com

Arquitectura de pico-JEPA

Base Teórica:

- La arquitectura JEPA, predice representaciones en un espacio latente, no a nivel de píxel (conceptos más abstractos y semánticos).

Componentes Clave:

- **Predictor:** Predice el embedding de la región enmascarada, basándose únicamente en el embedding del contexto.
- **Context Encoder:** Genera embeddings del contexto.
- **Target Encoder:** Genera el objetivo del predictor de forma estable.

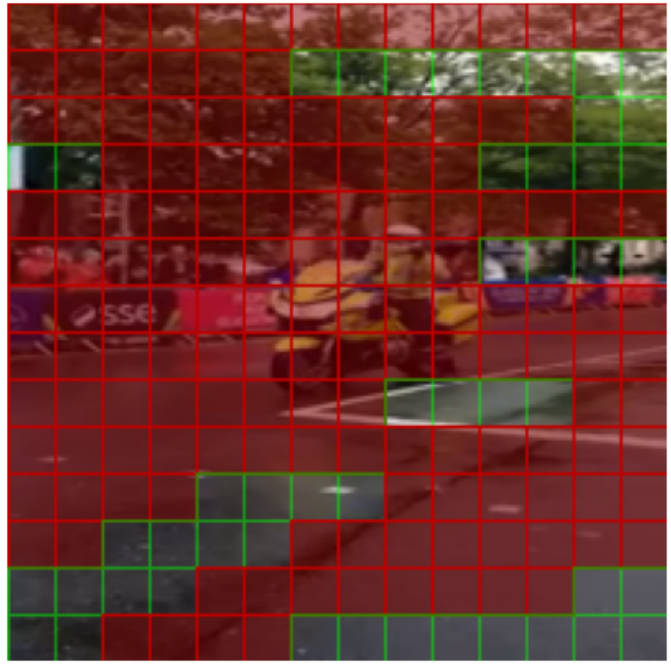

Figura: Visualización del enmascaramiento espacio-temporal en pico-JEPA

El Enfoque de Ensamble

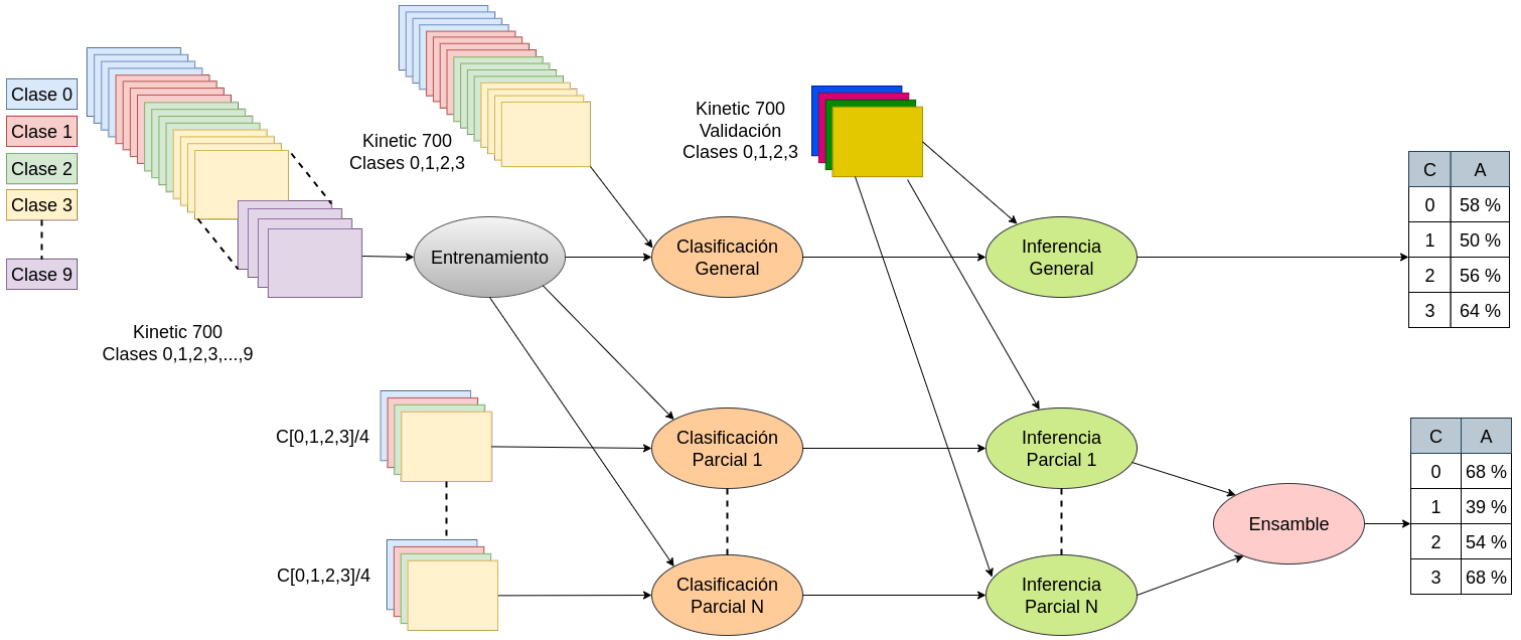

Figura: Diagrama del ensamble. Arriba: modelo general. Abajo: ensamble de N individuos.

- **Inferencia del Ensamble:** La clasificación final se decide por votación de mayoría entre los N modelos parciales.

Diseño del Experimento

- **Objetivo:** Comparar el rendimiento de un modelo único "General" frente a un ensamble de 4 modelos "Parciales".
- **Dataset:**
  - Pre-entrenamiento: 10 primeras clases del dataset Kinetics-700.
  - Clasificación final: Evaluada sobre 4 de esas clases.
- **Hardware:** Entrenamiento realizado en un equipo estándar (Intel i7, 64GB RAM, 2x NVIDIA 1060 6GB), demostrando su accesibilidad.
- **Modelos a Evaluar:**
  - **Modelo "General":** Entrenado con el 100 % de los datos de entrenamiento.
  - **Modelos "Parciales" (1-4):** Cada uno entrenado con un subconjunto (25 %) de los datos.

utn bhi

Instancias de pico-JEPA

Tabla: 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	

Resultados: El Ensamble Supera al Modelo Único

Tabla: Resultados de precisión promedio sobre 4 clases de Kinetics-700.

Clase (K700)	General	Parcial 1	Parcial 2	Parcial 3	Parcial 4	Votación Mayoría
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 %**

- El ensamble (Votación Mayoría) alcanza una precisión de **57.63 %**, superando al modelo General (57.15 %).

utn bhi

Certeza de los modelos

Tabla: 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 %

Análisis: Robustez y Consistencia

- **El poder de la diversidad:** Cada modelo parcial comete errores diferentes. El voto colectivo compensa las debilidades individuales.
- **Menor certeza, mejor decisión:**
  - Los modelos parciales muestran menor confianza individualmente en sus predicciones.
  - Sin embargo, la decisión del ensamble es más acertada y robusta en promedio.

Conclusión Clave

El ensamble demuestra ser más robusto en promedio que cualquiera de sus componentes individuales.

utn bhi

Visualización del Espacio de Características (t-SNE)

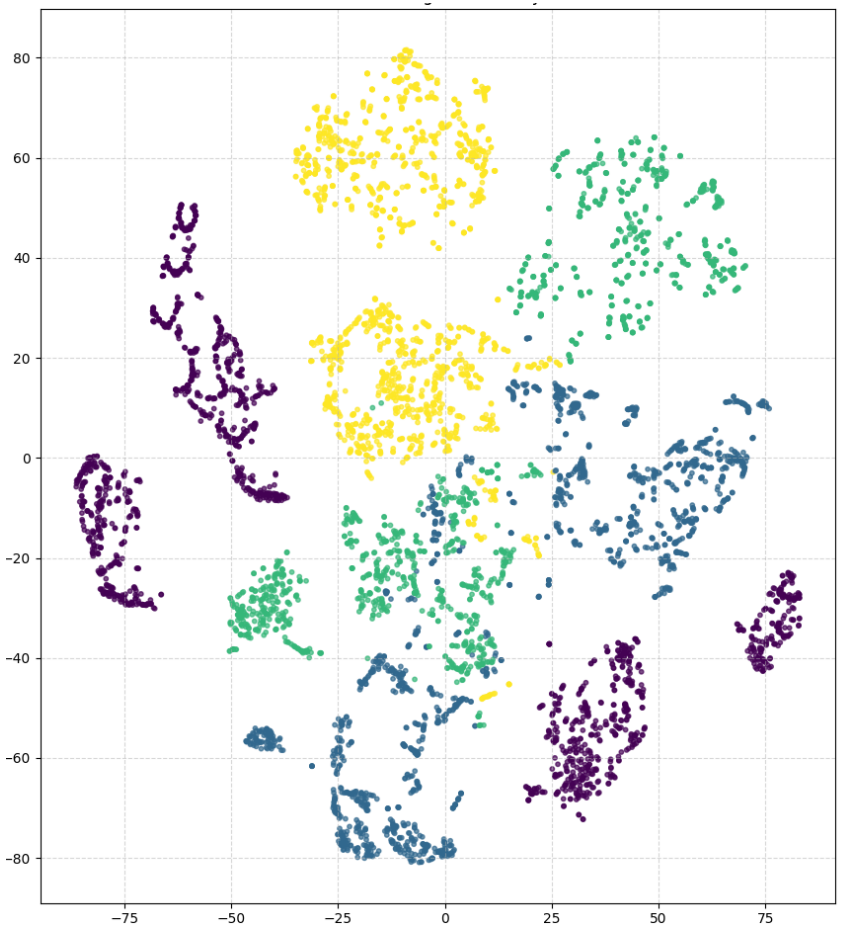

Figura: Mapa de embeddings del modelo Parcial 1

Interpretación:

- Es una "radiografía" de cómo el modelo organiza internamente las clases de video.
- **Clústeres bien definidos** para la mayoría de las clases (violeta, verde, amarillo), indicando una buena separación.
- La clase "adjusting glasses" (azul) muestra **menor cohesión**, lo que se correlaciona con su menor precisión en los resultados.

Contribuciones Fundamentales

1. **pico-JEPA Framework:** Un sistema ultra-ligero, robusto y autocontenido para la investigación en comprensión de video.
2. **Validación del Ensamble:** Demostramos que la "sabiduría de las multitudes" es aplicable y supera a un modelo monolítico en este contexto.
3. **Democratización:** Reducimos la barrera computacional, abriendo la puerta a más investigadores para experimentar con arquitecturas de vanguardia.

Trabajo Futuro

- **Ensambles más Sofisticados:**
  - Explorar técnicas como Stacking, donde un "meta-modelo" aprende a ponderar los votos.
- **Aumentar la Diversidad:**
  - Crear ensambles heterogéneos con variaciones en la arquitectura o hiperparámetros de cada modelo.
- **Aplicaciones Prácticas:**
  - Desplegar en redes de dispositivos de borde (*edge computing*) para tareas como vigilancia inteligente o robótica.

¡Gracias!

Código fuente disponible en:

`https://github.com/BHI-Research/pico-jepa`

¿Preguntas?

utn bhi