pico-JEPA: Understanding Video with Ultra-Light Models and Collective Wisdom

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

utn bhi

The Problem: The Computational Barrier

- **Context:** Video understanding models like V-JEPA are very powerful and learn in a self-supervised manner.
- **The Challenge:** Their training requires enormous computational resources (GPU clusters, large datasets), limiting access for most researchers.
- **The Key Question:** Can we democratize research in this area and achieve competitive results without access to supercomputers?

Our Proposal: Collective Wisdom

- **Main Idea:** Instead of one giant model, we propose using an ensemble of ultra-light models.
- **Introducing pico-JEPA:** An architecture designed to be a "base unit" for ensembles, small enough to allow for collaborative architectures.
- **Central Hypothesis:** An ensemble of pico-JEPA models, trained on data subsets, can outperform a single, larger model trained on all the data.
- **Open source:** `https://github.com/BHI-Research/pico-jepa`

utn bhi

Reference

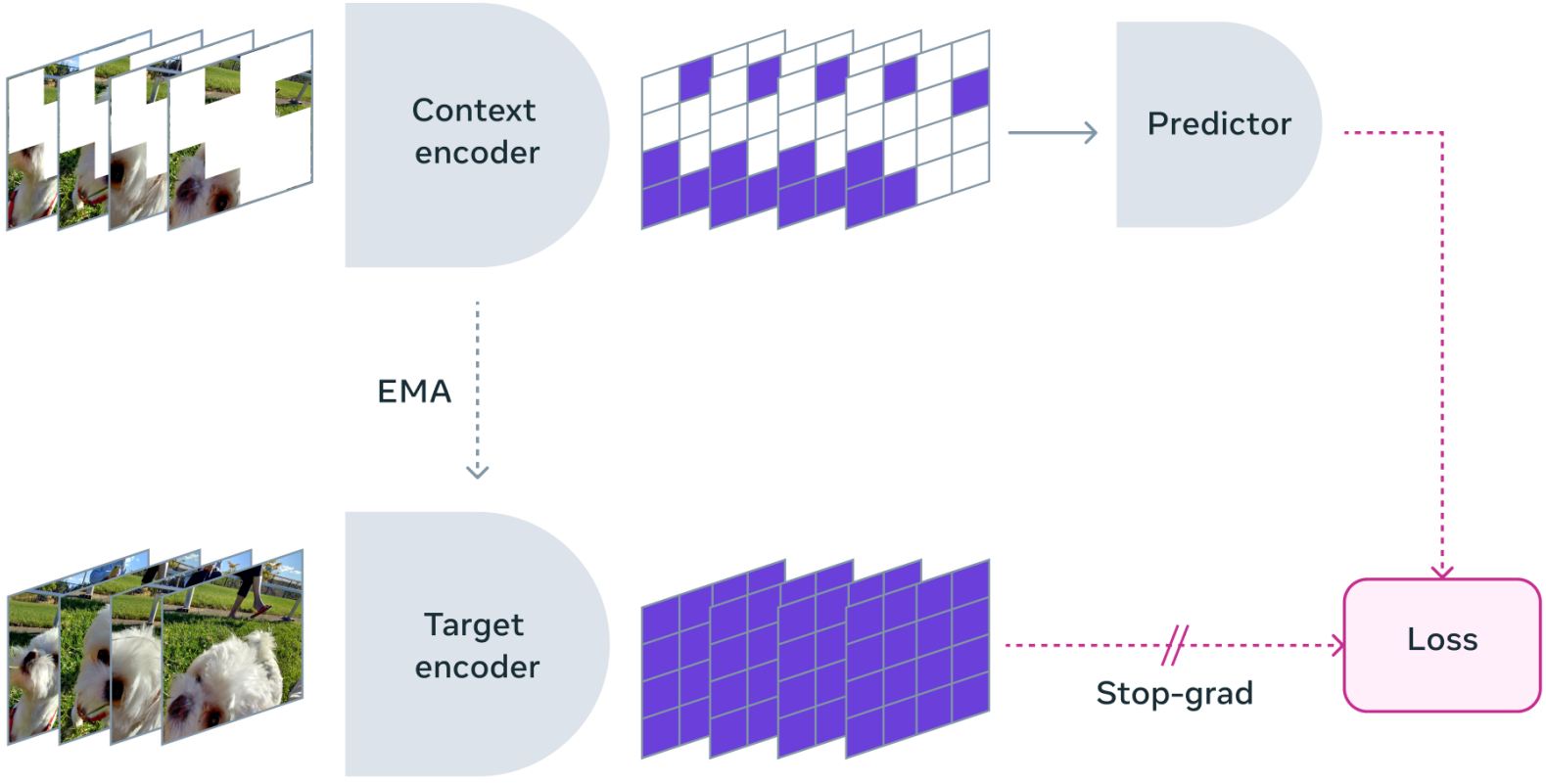

Figura: V-JEPA architecture, image courtesy of meta.com

pico-JEPA Architecture

Theoretical Basis:

- The JEPA architecture predicts representations in a latent space, not at the pixel level (more abstract and semantic concepts).

Key Components:

- **Predictor:** Predicts the embedding of the masked region, based only on the context embedding.
- **Context Encoder:** Generates context embeddings.
- **Target Encoder:** Generates a stable target for the predictor.

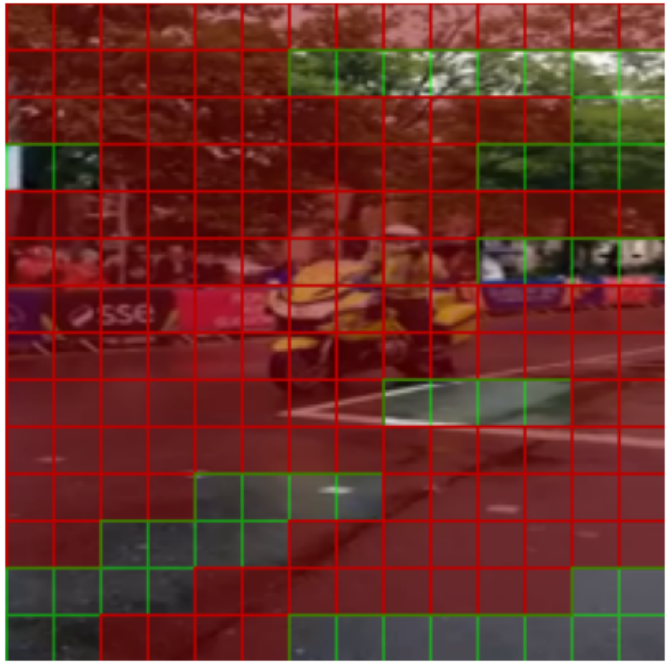

Figura: Visualization of spatio-temporal masking in pico-JEPA

The Ensemble Approach

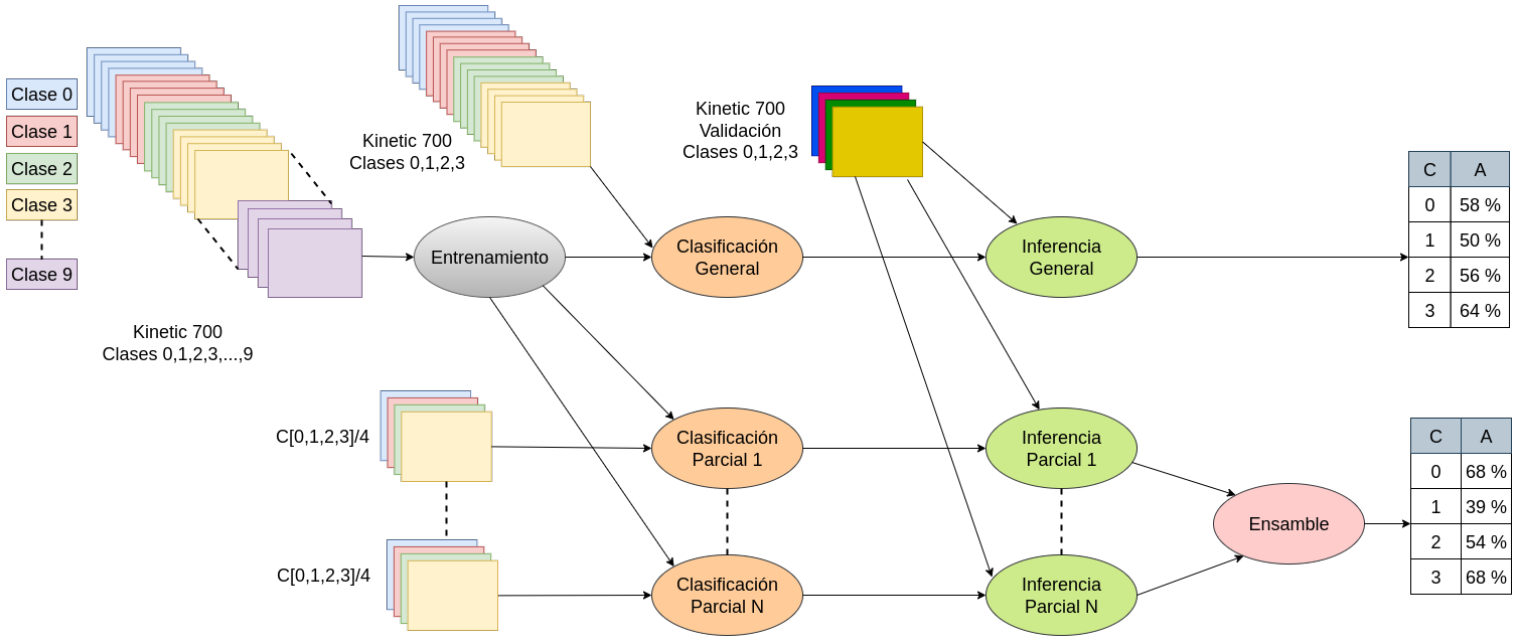

Figura: Ensemble diagram. Top: general model. Bottom: ensemble of N individuals.

- **Ensemble Inference:** The final classification is decided by majority vote among the N partial models.

Experiment Design

- **Objective:** Compare the performance of a single "General" model against an ensemble of 4 "Partial" models.
- **Dataset:**
  - Pre-training: First 10 classes of the Kinetics-700 dataset.
  - Final classification: Evaluated on 4 of those classes.
- **Hardware:** Training performed on standard equipment (Intel i7, 64GB RAM, 2x NVIDIA 1060 6GB), demonstrating its accessibility.
- **Models to Evaluate:**
  - **"General" Model:** Trained with 100 % of the training data.
  - **"Partial" Models (1-4):** Each trained with a subset (25 %) of the data.

utn bhi

pico-JEPA Instances

Tabla: pico-JEPA instances to be evaluated.

Models Evaluated				
Model	**Embedding**	**Pre-training Videos**	**Classification Videos**	**Validation Videos**
General	192	7203	2817	194
Partial 1			706	
Partial 2			707	
Partial 3			707	
Partial 4			697	

Results: The Ensemble Outperforms the Single Model

Tabla: Average accuracy results on 4 classes from Kinetics-700.

Class (K700)	General	Partial 1	Partial 2	Partial 3	Partial 4	Majority Vote
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 %
Average 4 classes	**57.15 %**	53.44 %	50.42 %	55.08 %	53.54 %	**57.63 %**

- The ensemble (Majority Vote) achieves an accuracy of **57.63 %**, outperforming the General model (57.15 %).

Model Certainty

Tabla: Certainty of the classification models.

Certainty Percentage					
Class	**General**	**Partial 1**	**Partial 2**	**Partial 3**	**Partial 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 %

Analysis: Robustness and Consistency

- **The power of diversity:** Each partial model makes different errors. The collective vote compensates for individual weaknesses.
- **Lower certainty, better decision:**
  - The partial models show lower individual confidence in their predictions.
  - However, the ensemble's decision is more accurate and robust on average.

Key Conclusion

The ensemble proves to be more robust on average than any of its individual components.

Feature Space Visualization (t-SNE)

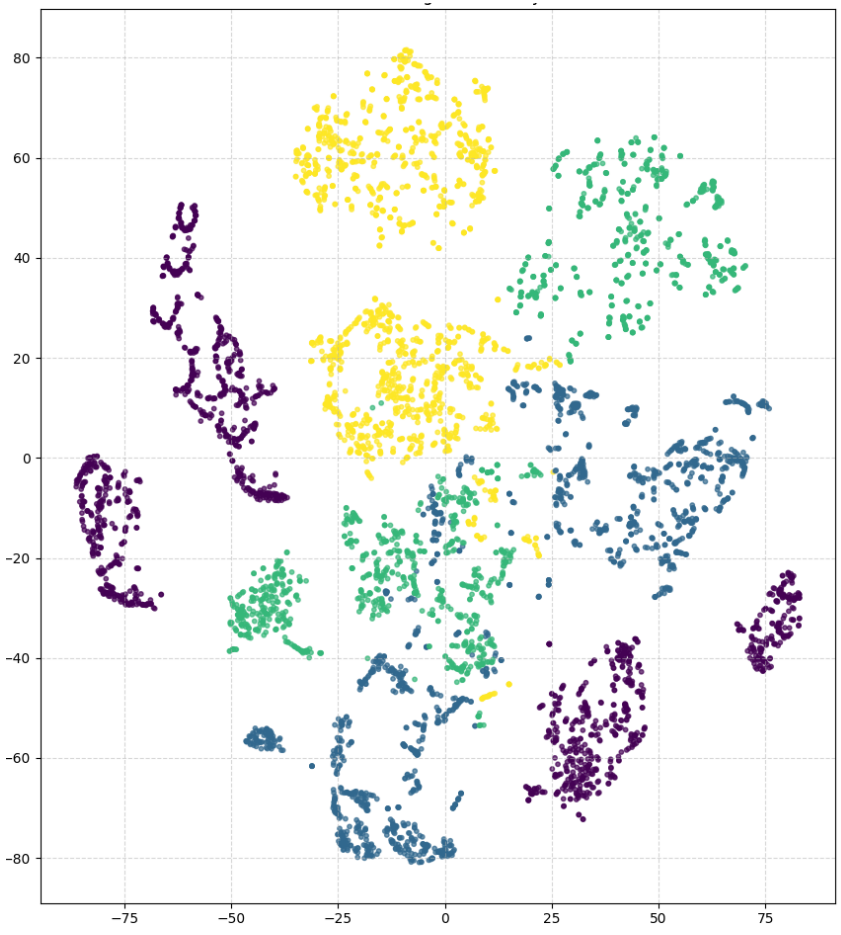

Figura: Embedding map of the Partial 1 model

Interpretation:

- It is an "x-ray.ºf how the model internally organizes the video classes.
- **Well-defined clusters** for most classes (purple, green, yellow), indicating good separation.
- The .ªdjusting glassesçlass (blue) shows **less cohesion**, which correlates with its lower accuracy in the results.

Fundamental Contributions

1. **pico-JEPA Framework:** An ultra-light, robust, and self-contained system for research in video understanding.
2. **Ensemble Validation:** We demonstrate that the "wisdom of crowdsïs applicable and outperforms a monolithic model in this context.
3. **Democratization:** We lower the computational barrier, opening the door for more researchers to experiment with state-of-the-art architectures.

Future Work

- **More Sophisticated Ensembles:**
  - Explore techniques like Stacking, where a "meta-model" learns to weigh the votes.
- **Increase Diversity:**
  - Create heterogeneous ensembles with variations in the architecture or hyperparameters of each model.
- **Practical Applications:**
  - Deploy on networks of edge devices (*edge computing*) for tasks like smart surveillance or robotics.

Thank You!

Source code available at:

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

Questions?

utn bhi