Contrastive Predictive Coding for Audio & Images

Multi-Modal Context Learning with Hybrid Negative Sampling

Anirudh Garg Satyankar Chandra

EE782: Advanced Machine Learning
Instructor: Prof. Amit Sethi
Indian Institute of Technology Bombay

November 24, 2025

Background: Original CPC (van den Oord et al., 2018)

- Goal: learn **unsupervised representations** by predicting the *future* in latent space.
- Architecture:
  - Non-linear encoder g_{enc}: maps raw data (audio, images) to local latents z_t.
  - Autoregressive model g_{ar} (GRU): aggregates $z_{\leq t}$ into context c_t.
- Training objective: **InfoNCE** contrastive loss.
  - Score true future latent z_{t+k} vs. many negative latents z^-.
  - Encourages c_t to keep information useful for predicting the future.
- Demonstrated strong performance on:
  - Speech (phone classification), images (ImageNet), and RL state representation.

Our Motivation

Key questions we wanted to explore:

1. **Multi-modal context:** Can a *single* context model summarise both audio and images?
2. **Negative sampling design:** Beyond in-batch negatives, how do
   - memory queues,
   - synthetic "hard" negatives (noisy / shuffled latents),
   - and cross-modal negatives

   affect representation quality?
3. **Measuring context quality:** How do we decide which CPC variant is better *without* relying only on downstream accuracy?
   - Alignment & uniformity,
   - Linear probes,
   - t-SNE geometry.

Method: Architecture & Training Setup

Encoders

- **Audio encoder:** Mel-spectrogram + 2D CNN $\rightarrow$ sequence of 64-dim latents.
- **Image encoder:** small ConvNet on 128×128 images, flattened spatial grid $\rightarrow$ sequence of 64-dim latents.

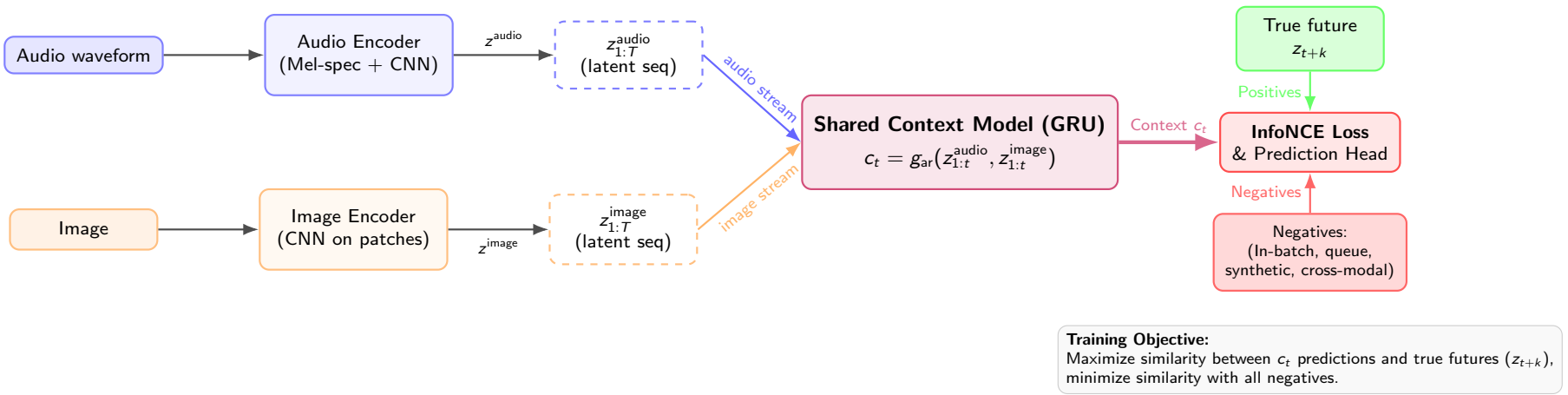
Shared context & loss

- Single GRU-based context model (256-dim) used for *both* modalities.
- Projection $c_t \rightarrow \hat{z}_{t+1}$, trained with InfoNCE.

Hybrid negative sampler

- In-batch negatives.
- MoCo-style **queue** of past latents (4096 entries).
- **Synthetic negatives:** Gaussian-perturbed and time-shuffled latents.
- Optional **cross-modal** negatives in multi-modal batches.

Model Overview: Multi-Modal CPC Architecture



Training Objective:
Maximize similarity between c_t predictions and true futures (z_{t+k}), minimize similarity with all negatives.

Algorithm Overview

Algorithm 1 Multi-Modal CPC Training with Hybrid Negatives

Require: Dataset $\mathcal{D}$, encoders $g_{\text{enc}}^{\text{aud}}, g_{\text{enc}}^{\text{img}}$, context GRU g_{ar}, negative samplers $\{\mathcal{S}_{\text{aud}}, \mathcal{S}_{\text{img}}\}$, epochs E, temperature τ

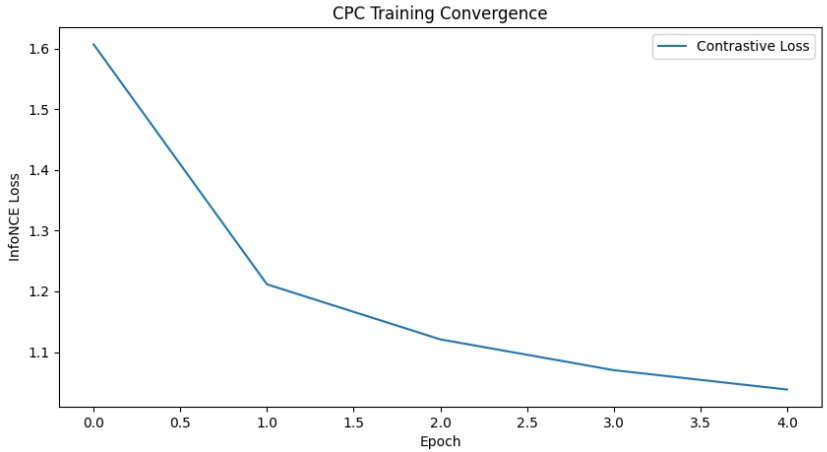
for epoch $= 1$ to E **do**
  for mini-batch B from $\mathcal{D}$ **do**
    Encode each modality:
      $Z^{\text{aud}} \leftarrow g_{\text{enc}}^{\text{aud}}(B_{\text{audio}})$
      $Z^{\text{img}} \leftarrow g_{\text{enc}}^{\text{img}}(B_{\text{image}})$
    Obtain context sequences:
      $C^{\text{aud}} \leftarrow g_{\text{ar}}(Z^{\text{aud}})$
      $C^{\text{img}} \leftarrow g_{\text{ar}}(Z^{\text{img}})$
    Build cross-modal negative pools from $Z^{\text{aud}}, Z^{\text{img}}$
    Compute audio loss:
      $L_{\text{aud}} \leftarrow \mathcal{S}_{\text{aud}}(C^{\text{aud}}, Z^{\text{aud}}, \tau, \text{cross-modal})$
    Compute image loss:
      $L_{\text{img}} \leftarrow \mathcal{S}_{\text{img}}(C^{\text{img}}, Z^{\text{img}}, \tau, \text{cross-modal})$
    $L \leftarrow \frac{1}{2}(L_{\text{aud}} + L_{\text{img}})$
    Backpropagate and update all trainable parameters
  end for
end for

Algorithm 2 Hybrid Negative Sampling for One Modality

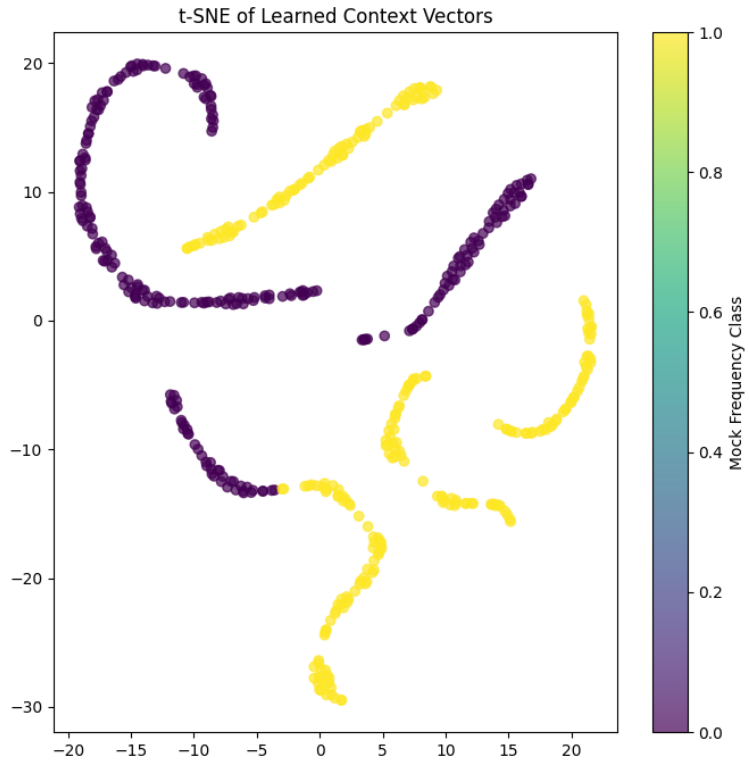
Require: Context $C \in \mathbb{R}^{B\times T\times d}$, latents $Z \in \mathbb{R}^{B\times T\times d}$, temperature τ, queue $\mathcal{Q}$, cross-modal pool $\mathcal{Z}^{\text{cm}}$

Form positives: $Z^{+} \leftarrow Z[:, 1:T-1, :]$, $C^{+} \leftarrow C[:, 0:T-2, :]$
Flatten: $Z_{\text{flat}}^{+}, C_{\text{flat}}^{+} \in \mathbb{R}^{N\times d}$, where $N = B(T-1)$
In-batch negatives: $Z_{\text{all}} \leftarrow Z_{\text{flat}}^{+}$
if $\mathcal{Q}$ not empty **then**
  $Z_{\text{all}} \leftarrow \text{concat}(Z_{\text{all}}, \mathcal{Q})$
end if
Generate synthetic negatives:
  $Z_{\text{pert}} \leftarrow Z_{\text{flat}}^{+} + \sigma \cdot \mathcal{N}(0, I)$
  $Z_{\text{shuf}} \leftarrow$ random permutation of Z_{flat}^{+}
$Z_{\text{all}} \leftarrow \text{concat}(Z_{\text{all}}, Z_{\text{pert}}, Z_{\text{shuf}})$
if $\mathcal{Z}^{\text{cm}}$ not empty **then**
  $Z_{\text{all}} \leftarrow \text{concat}(Z_{\text{all}}, \mathcal{Z}^{\text{cm}})$
end if
Compute logits: $\ell = C_{\text{flat}}^{+} Z_{\text{all}}^{\top} / \tau$
Labels: $y = [0, 1, \ldots, N-1]$
$L = \text{CrossEntropy}(\ell, y)$
Enqueue Z_{flat}^{+} into $\mathcal{Q}$ (FIFO)
return L

Results: Audio CPC (Synthetic Sine Waves)

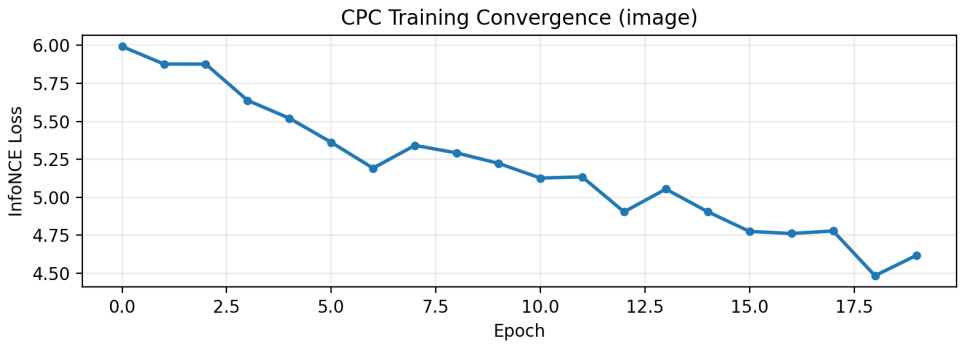


Training curve: InfoNCE loss vs. epoch.

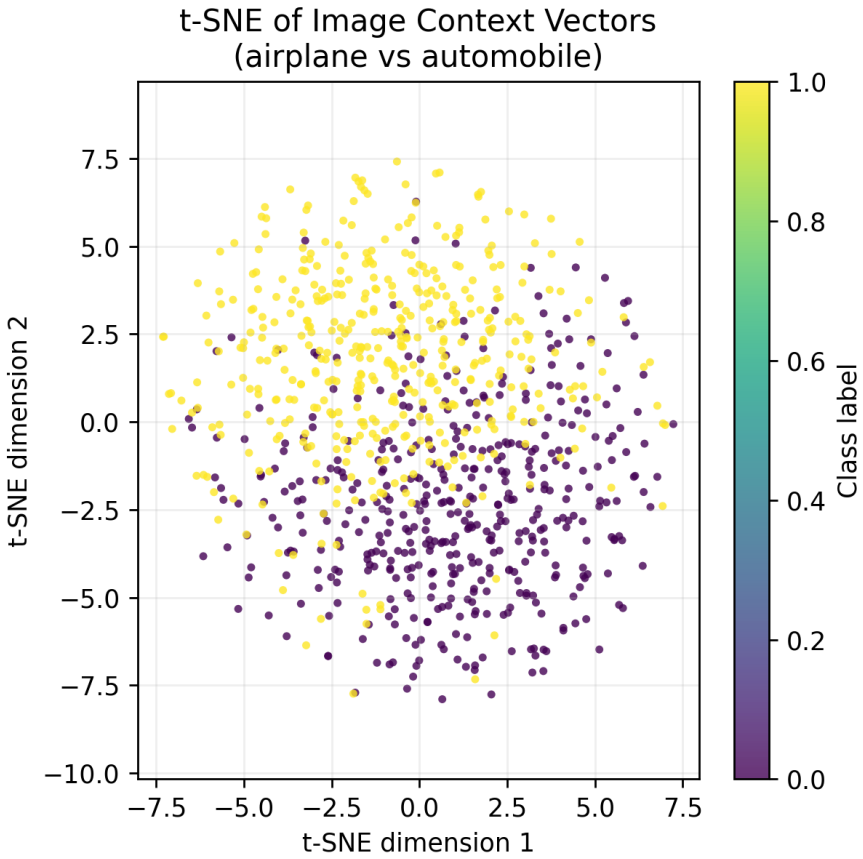


t-SNE of audio context vectors (colour = frequency class).

Results: Image CPC (Two-Class CIFAR-10)



Training curve for image CPC (InfoNCE loss vs. epoch).

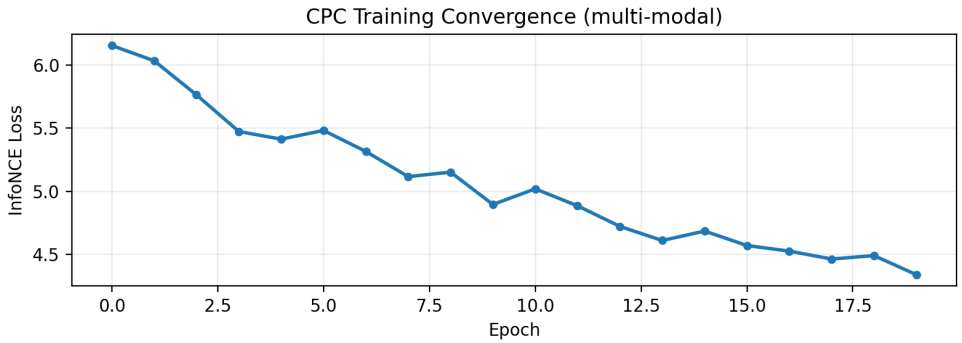


t-SNE of image context vectors (colour = airplane vs. automobile).

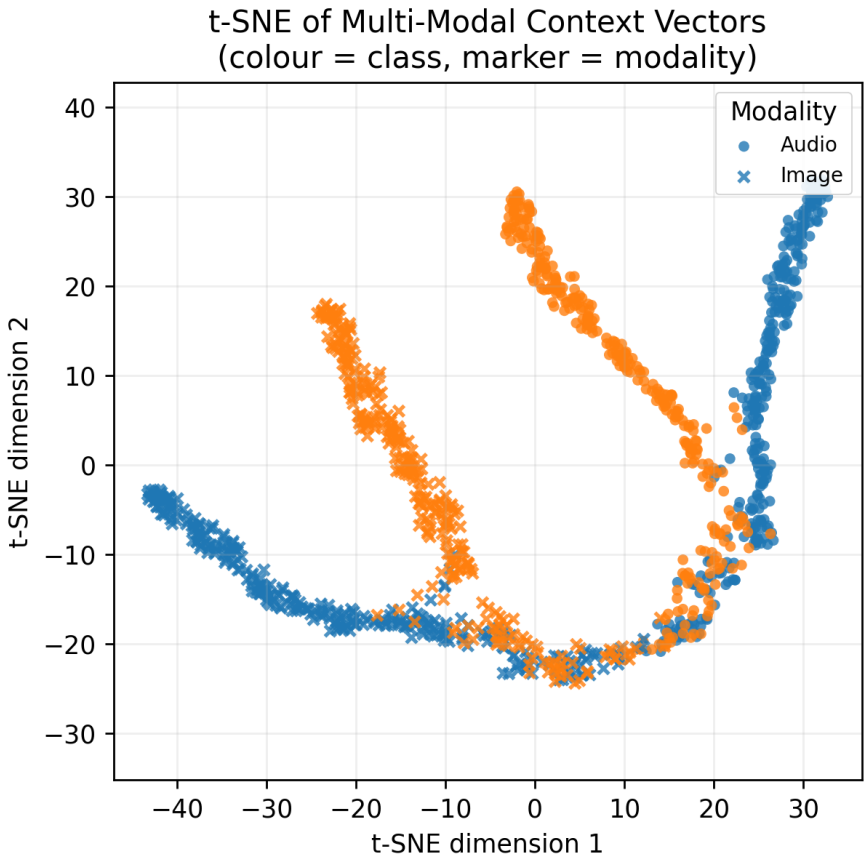
Results: Multi-Modal CPC (Audio + Image)

Multi-modal behaviour

- Train on paired (audio, bars-image) data with shared context GRU.
- Use cross-modal negatives (audio latents as negatives for image predictions and vice versa).



Multi-modal CPC training curve.



Summary of Observations

Representations

- Audio representation: **very clean geometry** (two smooth manifolds, almost linearly separable).
- Image representation: good but not perfect separation; reveals the difficulty of visual structure under same CPC hyperparameters.
- Multi-modal: snakes for each class touch and diffuse slightly; audio and image points intermingle along the same semantic manifold.

Conclusions & Takeaways

What we accomplished

- Implemented a **unified multi-modal CPC pipeline** with: separate audio/image encoders, shared context GRU, and hybrid negatives.
- Showed that CPC can learn **high-quality audio representations** and, after tuning, **strong image representations** on a realistic benchmark.
- Used **alignment, uniformity, linear probes, and t-SNE** to characterise context quality beyond raw loss.

Key lessons

- Negative sampling is *modality-sensitive*: settings that are perfect for audio can be too harsh for images.
- Multi-modal CPC with cross-modal negatives can encourage a shared, semantically meaningful context space.
- Controlled, compute-friendly setups (two-class CIFAR-10, synthetic audio) are very useful for understanding representation geometry.

Thank you!

Questions?