Audio-Visual Predictive Coding for Self-Supervised Visual Representation Learning

University of Nottingham
UK | CHINA | MALAYSIA

Mani Kumar Tellamekala[1], Michel Valstar[1], Michael Pound[1], Timo Giesbrecht[2]

[1] University of Nottingham, [2] Unilever R&D Port Sunlight, UK

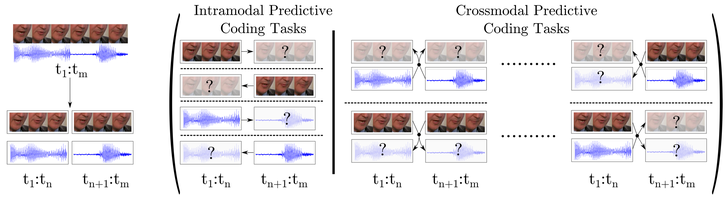
Introduction

- **Goal:** To learn semantic **visual features** from **unlabeled audio-visual** data
- **Motivation:** To make downstream task learning more labeled data efficient

Our Approach

- We propose a self-supervised learning method with a **multimodal proxy task.**
- A **Proxy Task** (e.g., jig-saw puzzle[3]) is designed based on **intrinsic correspondences** between unlabeled datapoints (intra- or cross-modal).
- Our proxy task learns builds on Contrastive Predictive Coding[1]
- **Predictive Coding:**
  - Given an unlabeled sequence ($X_1, X_2, X_3, \ldots, X_m$), predict future frames ($X_{n+1}, X_{n+2}, \ldots, X_m$) from past frames ($X_1, X_2, \ldots, X_n$) in the **feature space.**

Our Proxy Task: Audio-Visual Permutative Predictive Coding

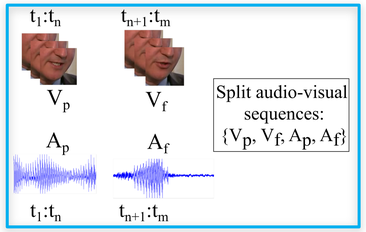


Overview of our approach to Self-Supervised Visual Feature Learning

- ❖ A multi-task learning framework designed to exhaustively exploit the **temporal (intra-modal) and cross-modal correspondences jointly.**
- ❖ Learning multiple predictive coding tasks could be less vulnerable to shortcuts or trivial representations than a single predictive coding task.

Method: Self-Supervised Learning

- We train a multi-task learning model with shared audio-visual feature encoders
- **Loss function:** A sum of **Noise Contrastive Estimation**[4] losses computed for all the below listed permutative predictive coding sub-tasks.

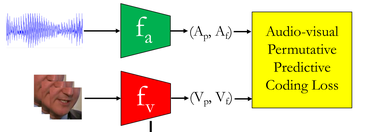


Dataset: Unlabeled word-utterance audio-visual sequences from LRW[2]

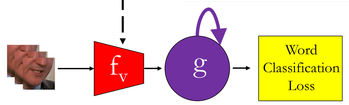
List of predictive coding sub-tasks

Permutative predictive coding sub-tasks (Input ➢Target)

One-to-One (#12 tasks)	One-to-Two (#12 tasks)
V_p ➢ V_f	V_p ➢ (V_f,A_f)
A_p ➢ V_p	V_f ➢ (V_p,A_p)
A_p ➢ A_f	A_p ➢ (V_f,A_f)
. . .	. . .
. . .	. . .
V_f ➢ V_p	V_f ➢ (A_p,A_f)
Two-to-One (#12 tasks)	**Two-to-Two (#6 tasks)**
(V_p,A_p) ➢ V_f	(V_p,V_f) ➢ (A_p,A_f)
(A_p,A_f) ➢ V_p	(A_p,A_f) ➢ (V_p,V_f)
(A_p,V_f) ➢ A_f	(V_p,A_p) ➢ (V_f,A_f)
. . .	(V_f,A_f) ➢ (V_p,A_p)
. . .	(V_p,A_f) ➢ (A_p,V_f)
(V_f,A_f) ➢ V_p	(A_p,V_f) ➢ (V_p,A_f)



Downstream Task: Lip-Reading



Task: Predict the word uttered in a video
Dataset: LRW with 500-word classes
Metric: Word Classification Rate (WCR)

Visual Encoder (2D ResNet-34)
Audio Encoder (1D ResNet-18)
Temporal Model (GRU/Temporal Conv.)

Evaluation Protocol: Measure WCR
- ❑ before finetuning the visual encoder
- ❑ after finetuning the visual encoder
  - ❑ using the entire train data and
  - ❑ using small amounts of train data.

Key Experimental Results

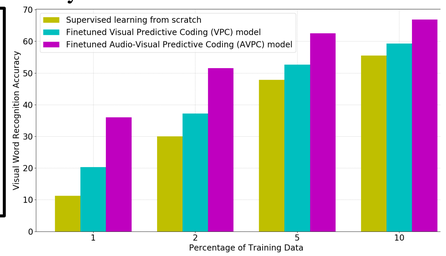
Word Classification Rates before finetuning (after finetuning) the visual feature encoder (f_v) on the lip-reading labeled data (LRW Test Set)

Proxy Task	Using Temporal Conv	Using GRU
AV Synchronization	50.70 (74.17)	55.26 (76.92)
Time-Arrow	52.42 (75.80)	59.88 (78.26)
AV Correspondence	56.22 (74.23)	61.90 (77.90)
Visual Predictive Coding	60.77 (77.95)	67.62 (81.76)
Audio-Visual Predictive Coding (ours)	**76.47 (80.44)**	**80.30 (83.16)**

Data-Efficiency Evaluation

- Number of labeled instances required to learn lip-reading task

- With 1% of train data (10 instances per word class),
  - Our method: 38% WCR
  - Fully-supervised: 11% WCR



Conclusion:

- Temporal and cross-modal correspondences used as natural supervision signals jointly lead to semantic visual features that
  - ➢ generalize well to the downstream supervised learning tasks and
  - ➢ highly efficient in terms of labeled data requirement

References:

1. Oord et al. "Representation learning with contrastive predictive coding." arXiv preprint :1807.03748 (2018).
2. Chung et al. "Lip reading in the wild." ACCV . Springer, Cham, 2016.
3. Noroozi et al. "Unsupervised learning of visual representations by solving jigsaw puzzles." ECCV, 2016.
4. Gutmann et al. "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models." AISTATS, 2010.

Contact: <mani.tellamekala,michel.valstar,michael.pound>@nottingham.ac.uk, timo.giesbrecht@unilever.com