HYBRID ATTENTION-BASED PROTOTYPICAL NETWORKS FOR FEW-SHOT SOUND CLASSIFICATION

Paper Number: 4769

You Wang, David V. Anderson

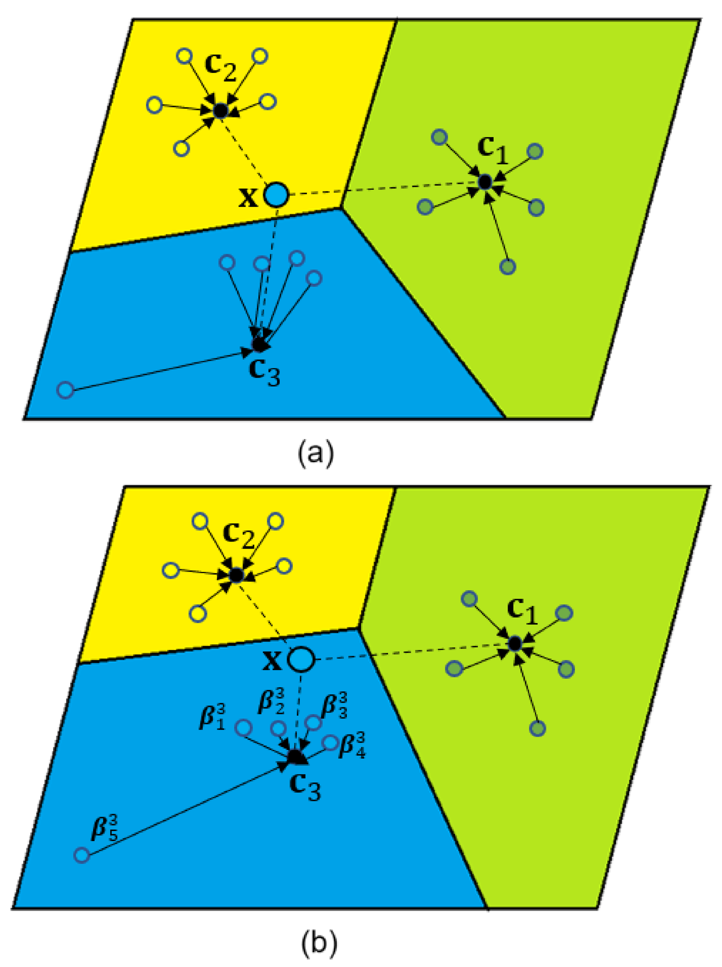
Georgia Institute of Technology, School of Electrical and Computer Engineering, Atlanta Georgia, USA

Abstract

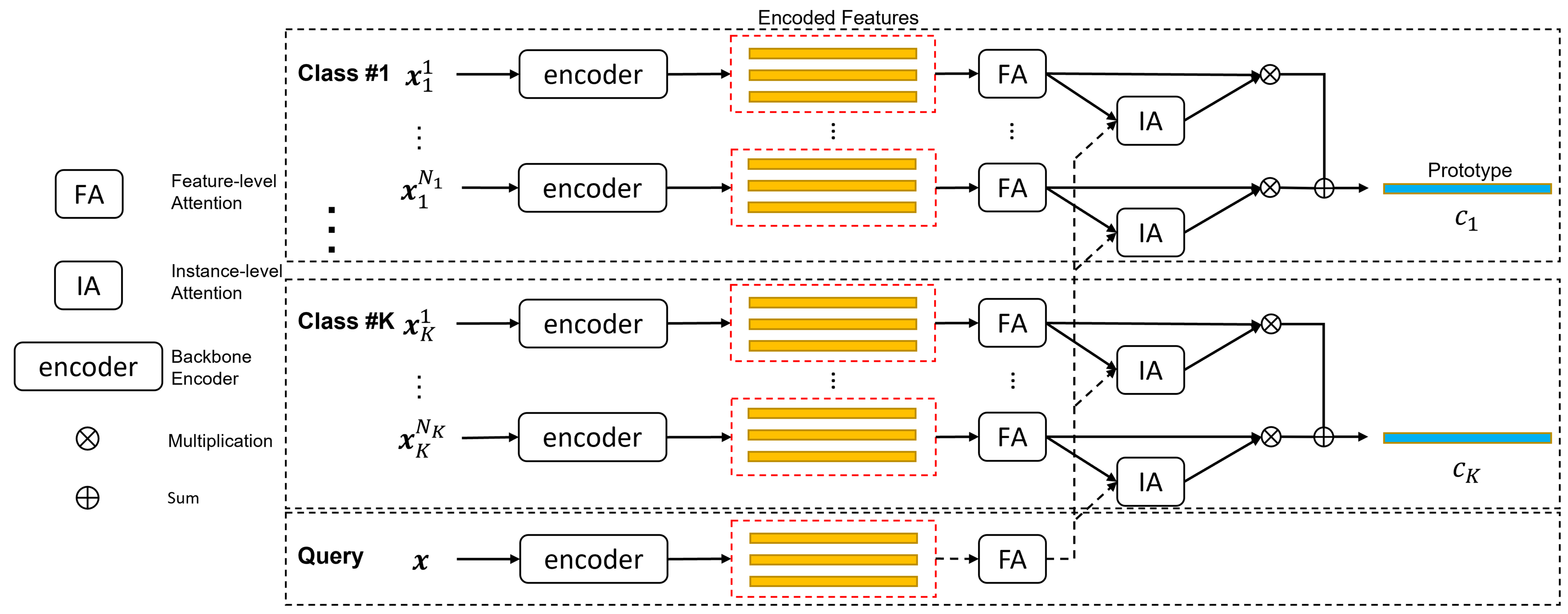
In recent years, prototypical networks have been widely used in many few-shot learning scenarios. However, as a metric-based learning method, their performance often degrades in the presence of bad or noisy embedded features, and outliers in support instances. In this paper, we introduce a hybrid attention module and combine it with prototypical networks for few-shot sound classification. This hybrid attention module consists of two blocks: a feature-level attention block, and an instance-level attention block. These two attention mechanisms can highlight key embedded features and emphasize crucial support instances respectively. The performance of our model was evaluated using the ESC-50 dataset and the noiseESC-50 dataset. The model was trained in a 10-way 5-shot scenario and tested in four few-shot cases, namely 5-way 1-shot, 5-way 5-shot, 10-way 1-shot, and 10-way 5-shot. The results demonstrate that by adding the hybrid attention module, our model outperforms the baseline prototypical networks in all four scenarios.

Motivations

- In audio classification tasks, attention is often used to emphasize certain temporal, channel, or spectral features.
- Prototypical networks, as a metric-based embedding learning, often suffers from bad feature vectors and outliers in the support instances.



Model Architecture

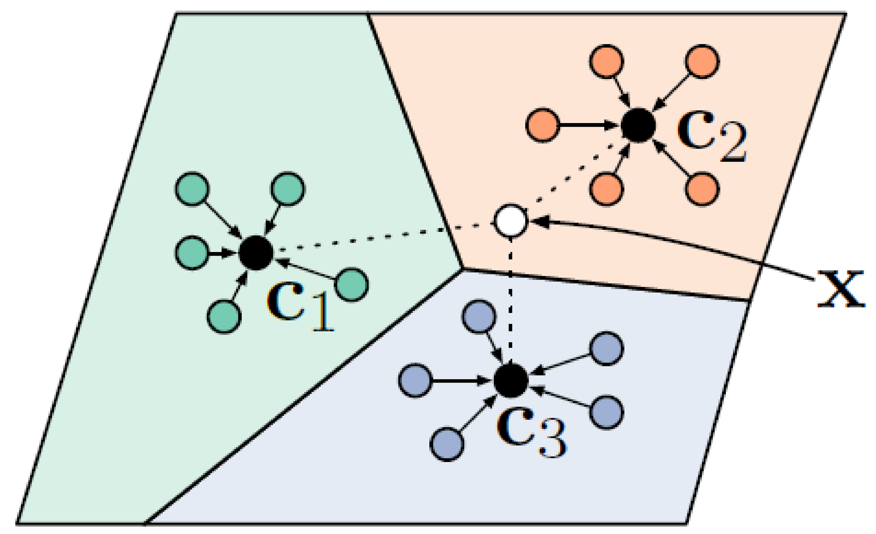


Related Work

- Prototypical Networks
  - The key idea is the computation of prototypes to represent each class by averaging the encoded feature vectors of support samples in each class.
  - The query sample is classified to the class of which the prototype is the nearest.

$$\mathbf{c}_k = \frac{1}{N_k}\sum_{i=1}^{N_k} f_\theta(\mathbf{x}_i^k), \quad k \in \{1, \ldots, K\}$$

$$P_\theta(y=k|\mathbf{x}) = \frac{\exp(-d(f_\theta(\mathbf{x}), \mathbf{c}_k))}{\sum_{k'} \exp(-d(f_\theta(\mathbf{x}), \mathbf{c}_{k'}))}$$

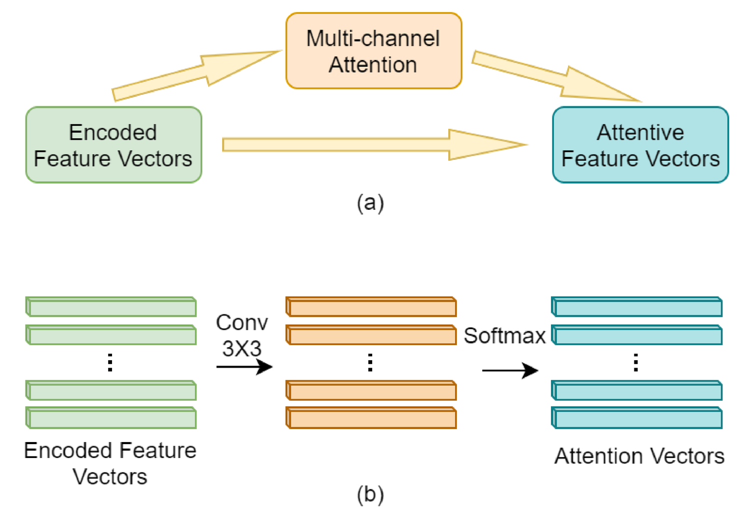


(a) Few-shot

Snell et al. 2017

Model Details

- Backbone Network
  - It contains 3 blocks consisting of a 3x3 convolutional layer, batch normalization, ReLU activation, and a max pooling layer consecutively.
  - Max pooling layer kernel sizes: 8x2, 8x2, and 2x1. Convolutional layer channel numbers: 128, 256, and 384.
- Feature-level Attention



Wang et al. 2021

- Instance-level Attention

$$\mathbf{c}_k = \sum_{i=1}^{N_k} \beta_i^k f_\theta(\mathbf{x}_i^k)$$

$$\beta_i^k = \frac{e_i^k}{\sum_{n=1}^{N_k} e_n^k}$$

$$e_i^k = sum\{\sigma(f_\phi(f_\theta(\mathbf{x}_i^k)) \circ f_\phi(f_\theta(\mathbf{x})))\}$$

Experimental Setup

- Datasets
  - ESC-50: 2000 5-second-long audio recordings organized into 50 balanced classes.
  - *noise*ESC-50: created in Chou et al. 2019 by mixing clean ESC-50 samples with random acoustic scenes from DCASE2016 dataset as additive background noise.
- Data Preparation
  - We randomly selected 35 classes for 10-way 5-shot training, 5 classes for 5-way 5-shot validation, and the remaining 10 classes for testing.
  - All audio clips were downsampled from 44.1kHz to 16kHz, and log mel-spectrograms with 128 mel bins were extracted.
  - The input features were z-score normalized using the mean and standard deviation of the training set before being fed into the model.

Results and Discussion

- ESC-50

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Prototypical Networks	64.40±1.38%	83.83±0.92%	47.83±1.10%	71.00±1.04%
Proto-FA (Ours)	**71.18±1.23%**	89.60±1.08%	**57.08±1.43%**	78.48±1.51%
Proto-HA (Ours)	–	**90.35±0.83%**	–	**80.08±1.31%**

- *noise*ESC-50

Model	5-way 1-shot	5-way 5-shot	10-way 1-shot	10-way 5-shot
Prototypical Networks	61.53±0.40%	81.03±0.57%	45.90±0.28%	64.98±0.52%
Proto-FA (Ours)	**71.35±0.90%**	88.00±0.63%	**56.55±1.22%**	78.55±0.75%
Proto-HA (Ours)	–	**88.78±0.45%**	–	**79.08±1.12%**

- Discussion
  - The feature-level attention module is capable of making data samples more distinguishable.
  - Instance-level attention module is able to focus on crucial support samples for both clean and noisy scenarios.
  - However, with *noise*ESC-50, when all the support and query samples are degraded, the advantage of instance-level attention module might not be as big as with clean data.