Motivation
➢ The complex entanglement among different attributes' visual features in the embedding space.
➢ The diversity of attributes in images of the same category leads to inaccurate predictions.
➢ Using a single mapping function to map different granular visual features to semantic space degrades model performance.

Contribution
➢ We propose a visual feature augmentation that explicitly extracts attribute features and adopts a cosine similarity loss to disentangle them in the embedding space, enhancing the visual features.
➢ We propose a semantic feature augmentation containing a bias learner, which estimates an offset to alleviate the difference between the actual and predicted attributes, leading to improved class-level semantic features for each image.
➢ We employ two mapping functions to avoid inconsistencies in the mapping process of different granular visual features, thus reinforcing the semantic features corresponding to varying visual features.

Method

\[ L_{\text{cvx}} = \| \hat{Z}^I - I \|^2 \]

where \( I \in \mathbb{R}^M \), \( \hat{Z} = [\hat{z}_1, \ldots, \hat{z}_n] \), and \( \hat{z}_i \) is the predicted attribute feature.

\[ L_{\text{attr}} = -\log \frac{\exp(\hat{a}_i \cdot \hat{a}_j)}{\sum_{k=1}^{n} \exp(\hat{a}_k \cdot \hat{a}_j)} \]

where \( a_j \) is the ground truth semantic feature.

\[ L_{\text{cls}} = -\log \frac{\exp(\hat{a}_j \cdot \hat{a}_j)}{\sum_{k=1}^{n} \exp(\hat{a}_k \cdot \hat{a}_j)} \]

Results

Experiments

Comparison with State-of-the-Art Methods

- In fine-grained datasets, the CZSL accuracy and the harmonic mean increase first and then decrease.
- For the coarse-grained dataset, as the value of \( \beta_1 \) gradually increases, the CZSL accuracy and the harmonic mean decrease and reach a minimum when \( \beta_1 \) equals 1.