**Overview**

Recent work has shown that explicitly modeling the co-occurrence relationship between classes is critical for achieving good performance on multi-label classification task. We propose an end-to-end model by adopting the transformer-based feature-extraction backbone with a novel and efficient association module.

**Highlights:**
- A simple yet effective end-to-end transformer-based framework for multi-label classification.
- A new association module to explore label correlation.
- The module is learnable and is computation-efficiency evaluated on different benchmark datasets.
- The proposed model on different benchmark datasets: MS-COCO and PASCAL VOC and obtains superior or comparable performance.

**Introduction**

**Background:** The goal of Multi-Label Classification (MLC) is to predict a set of labels for a single image.

**Challange:** 1). tiny object detection and 2). positive and negative label imbalance.

**Motivation:** The missing tiny object detection is one of the main challenges of MCL. We propose to overcome this issue by adopting the Vision Transformer (ViT) as backbone with label association information to boost the final prediction.

**Approach**

**Architecture Overview:** We leverage the transformer as the backbone feature extractor to get the extracted feature $L \in R^d$. We forward $L$ through the Association Module (AM) to calculate the association matrix $A \in R^C \times C$ and output $N_1 \in R^C$. Meanwhile, we get another output $N_2 \in R^C$ by forwarding $L$ through a fully connected layer. The final prediction is the fusion operation of the $N_1$ and $N_2$.

**Operation of AM:**

We first unsqueeze the feature $L$ and conduct a 1D convolution to project 1D embedding to 2D, \( \{K, M\} \in R^C \times d \):

\[ K, M = \text{Conv1D}(\text{unsqueeze}(L, 1)) \]

We transpose the feature $K$ and conduct multiplication with $K$ itself attached a sigmoid function to calcualte the association matrix $A$:

\[ A = \text{sigmoid}(K \times K^T) \]

We finally multiply feature $M$ with association matrix $A$ and apply another Conv1D to get the output $N_2$:

\[ N_1 = \text{Conv1D}(M \times A) \]

**Data & Metrics:** We evaluate the model with two public datasets: MS-COCO and PASCAL VOC. To evaluate the performance of our model, we use mean average precision (mAP) as metric.

**Conclusion**

We proposed AssocFormer, which combines a transformer backbone with a light-weight association module, for the task of multi-label image classification. This approach outperforms prior work on two standard public benchmark datasets, while simultaneously being simpler to implement.