CS 4701 Final Report

James Kim (jjk297)

Spring 2025

1 Loss Functions

We start by defining the following terms:

- $\mathbf{x}$ as the input image
- $\mathbf{f}(\mathbf{x})$ as the feature vector extracted by the ViT backbone
- $s \in \{0, 1, 2, 3\}$ as the season class (Spring, Summer, Autumn, Winter)
- $t \in \{0, 1, 2\}$ as the subtype within a season
- $c \in \{0, 1, ..., 11\}$ as the combined class, where $c = 3s + t$

1.1 Joint Loss

The joint loss approach directly supervises both levels of the hierarchy, using a weighted combination of losses:

$$\mathcal{L}_{\text{joint}}(\mathbf{x}, s, t, c) = \alpha \cdot \mathcal{L}_{\text{season}}(\mathbf{x}, s) + \beta \cdot \mathcal{L}_{\text{subtype}}(\mathbf{x}, t) + \gamma \cdot \mathcal{L}_{\text{full}}(\mathbf{x}, c)$$

Where:

- $\mathcal{L}_{\text{season}}(\mathbf{x}, s) = -\log P_{\text{season}}(s|\mathbf{x})$ is the cross-entropy loss for season prediction
- $\mathcal{L}_{\text{subtype}}(\mathbf{x}, t) = -\log P_{\text{subtype}}(t|\mathbf{x})$ is the cross-entropy loss for subtype prediction
- $\mathcal{L}_{\text{full}}(\mathbf{x}, c) = -\log P_{\text{full}}(c|\mathbf{x})$ is the cross-entropy loss for the full 12-class prediction
- α, β, γ are weighting parameters (typically $\alpha + \beta + \gamma = 1$) that will be tuned

1.2 Hierarchical Softmax Loss

The hierarchical softmax computes:

1. **Season probabilities:**
$$P(s|\mathbf{x}) = \frac{\exp(\mathbf{w}_s^T \mathbf{f}(\mathbf{x}))}{\sum_{s'=0}^{3} \exp(\mathbf{w}_{s'}^T \mathbf{f}(\mathbf{x}))}$$
where $\mathbf{w}_s$ are the weights of the season classifier.

2. **Subtype probabilities (conditional on season):**
$$P(t|s, \mathbf{x}) = \frac{\exp(\mathbf{v}_{s,t}^T \mathbf{f}(\mathbf{x}))}{\sum_{t'=0}^{2} \exp(\mathbf{v}_{s,t'}^T \mathbf{f}(\mathbf{x}))}$$
where $\mathbf{v}_{s,t}$ are the weights of the subtype classifier for season s.

3. **Joint probabilities for the full 12-class classification:**

$$P(c|\mathbf{x}) = P(s|\mathbf{x}) \cdot P(t|s, \mathbf{x})$$

where $s = \lfloor c/3 \rfloor$ and $t = c \mod 3$.

The hierarchical softmax loss is then defined as the negative log-likelihood of the true class:

$$\mathcal{L}(\mathbf{x}, c) = -\log P(c|\mathbf{x}) = -\log[P(s|\mathbf{x}) \cdot P(t|s, \mathbf{x})]$$

Which can be expanded as:

$$\mathcal{L}(\mathbf{x}, c) = -\log P(s|\mathbf{x}) - \log P(t|s, \mathbf{x})$$

Here we can see that if the probability assigned to the true classes s or t are low, then log will make that term larger, thus increasing the loss.

For a batch of N examples with inputs $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ and corresponding labels $\mathbf{C} = \{c_1, c_2, ..., c_N\}$, the total loss is:

$$\mathcal{L}(\mathbf{X}, \mathbf{C}) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(\mathbf{x}_i, c_i)$$

This loss function encourages the model to correctly predict both the season and the subtype, with the mathematical structure explicitly modeling the hierarchical relationship between them.