Multilabel Aerial Image Classification using Deep Learning on the AID Dataset

Group Member:

- **Tunahan Yazar:** Responsible for literature review, dataset exploration, model design and implementation, conducting experiments, and leading report writing.
- **Cansu Temizkan:** Responsible for data preparation and augmentation, assisting in model training and evaluation, and contributing to experiments.
- **Defne Koçulu:** Responsible for evaluation metrics, performance visualization, and result interpretation.
- **Yavuz Can Atalay:** Responsible for model optimisation, error analysis, and statistical analysis
- **Batuhan Güzelyurt:** Responsible for documentation and deliverables, assisting in statistical analysis.

a) Problem Definition and Motivation

Automatic analysis of aerial imagery is a critical component of modern remote sensing, with wide ranging applications in urban planning, agricultural monitoring, environmental science, and disaster response. A fundamental task in this domain is scene classification, which aims to assign semantic labels to aerial images to understand their content.

Historically, this task has been approached as a single label classification problem, where each image is assigned to a single, mutually exclusive category. However, this paradigm oversimplifies the rich and complex nature of aerial scenes. An aerial view of a suburban area, for instance, is not just 'residential'; it simultaneously contains multiple landcover types and objects such as 'buildings', 'roads', 'trees', 'cars', and possibly a 'pool' or 'park'. Assigning only a single label results in a significant loss of valuable semantic information.
This limitation motivates the shift towards multilabel aerial image classification, a more advanced and realistic problem formulation. In this paradigm, an image can be associated with a set of multiple labels, allowing for a far more descriptive and comprehensive understanding of its contents. The goal is to develop a model that can recognize the presence of any number of relevant labels within a single image.

b) Overview of Related Work

The methodologies for aerial scene classification have evolved considerably over the years, moving from hand crafted features to deep learning based approaches.

Early Methods: Hand Crafted Features Initial research focused on creating feature representations based on image properties. These methods can be grouped into:

- **Low Level Features:** These capture basic visual properties. Examples include Scale Invariant Feature Transform (SIFT), which describes local image patches based on gradient orientations, and Local Binary Patterns (LBP), which encode texture information.
- **Mid Level Representations:** To create a more holistic scene representation, low level features were often aggregated using models like the Bag of Visual Words (BoVW). This approach involves creating a visual vocabulary by clustering local features and then representing an image as a histogram of these visual words. Extensions like the Improved Fisher Kernel (IFK) provided more sophisticated encoding schemes [1].

While foundational, these methods often struggle to capture the high level semantic content required for complex scene understanding.

The Rise of Deep Learning The turning point in image classification was the success of deep Convolutional Neural Networks (CNNs), starting with AlexNet. As demonstrated in the original analysis of the AID dataset, pre trained CNNs like VGG-16 and CaffeNet decisively outperformed traditional methods, showcasing their superior ability to learn discriminative, hierarchical features directly from image data [1]. Subsequent architectures like ResNet [2] and DenseNet [3] introduced innovations such as residual and dense connections, enabling the training of even deeper and more powerful networks.

Modern Multi Label Approaches Current research in multilabel classification focuses on adapting and enhancing these powerful deep learning models. A key trend is the use of attention mechanisms, which allow a model to dynamically focus on the most relevant image regions for each potential label. Furthermore, researchers are exploring methods to explicitly model the relationships between labels. This includes using Recurrent Neural Networks (RNNs) or Graph Convolutional Networks (GCNs) to learn and leverage the cooccurrence patterns of different labels, improving predictive accuracy.

c) Dataset and Method

Dataset: This project will be based on the AID (Aerial Image Dataset), a large scale benchmark created for

aerial scene classification [1]. The original dataset consists of 10,000 aerial images, each 600x600 pixels, covering 30 distinct scene classes. Its key features, including high intraclass variation and small interclass dissimilarity, make it an excellent and challenging resource for developing robust models.

We will use a specific multilabel version of AID available from Hugging Face (jonathan-roberts1/AID_MultiLabel). This version is explicitly annotated for multilabel classification, providing a suitable foundation for our project.

Methodology: Our approach is centered on a fine tuned deep learning model, following a structured methodology.

1. **Data Preprocessing and Augmentation:** To improve model generalization and prevent overfitting, we will apply a series of data augmentation techniques to the training set. These will include random geometric transformations (rotation, horizontal/vertical flips, scaling) and appearance modifications (brightness, contrast, and saturation jittering).

2. **Model Architecture:** We will use a pretrained CNN as our feature extractor. Our primary candidates are ResNet-50 [2] and DenseNet-121 [3]. These architectures are chosen for their proven effectiveness and efficient gradient flow, which is crucial for training deep networks. The core idea is to adapt these models for the multilabel task:

    - The final classification layer of the pretrained model will be replaced with a new fully connected layer. The number of output neurons will match the total number of unique labels in our dataset.
    - A sigmoid activation function will be applied to this final layer. This is critical for multilabel classification as it outputs an independent probability for each label, allowing the model to predict the presence of multiple classes simultaneously.

3. **Training Strategy:** We will employ a transfer learning strategy to leverage the knowledge encoded in the pretrained models.

Loss Function: The model will be trained using a Binary Cross Entropy (BCE) loss function, which is the standard for multilabel problems as it evaluates the error for each label independently.

Fine Tuning: We will adopt a two stage fine tuning process. Initially, we will "freeze" the weights of the convolutional base and train only the newly added classification layer. This allows the classifier to adapt to the new dataset. Subsequently, we will "unfreeze" the entire network and continue training with a very low learning rate. This fine tunes the entire model for the specific features of aerial imagery.

Optimizer: We will use the Adam optimizer, a widely used and effective algorithm for training deep neural networks.

Hypothesis: We hypothesize that by fine tuning a pretrained, deep CNN architecture (like ResNet or DenseNet) on the multilabel AID dataset, our model will effectively learn the intricate spatial features and semantic correlations present in aerial imagery. We expect this approach to yield high accuracy in predicting multiple, cooccurring labels for any given scene, significantly outperforming simpler models.

d) Performance Metrics and Success Criteria

To rigorously evaluate the performance of our multilabel classification model, we will employ a set of standard and comprehensive metrics:

- **Precision, Recall, and F1 Score:** These will be calculated on a per label basis to assess performance for individual classes. We will also compute the micro and macro averages to get an overall sense of performance. The macro averaged F1 score is particularly important as it treats each class equally, regardless of its frequency.
- **Mean Average Precision (mAP):** This metric provides a single figure measure of quality across all classes, considering the ranking of predicted labels.
- **Hamming Loss:** This straightforward metric calculates the fraction of labels that are incorrectly predicted (a true label is missed or a false label is predicted). A lower Hamming Loss indicates a better model.

Success Criteria: We define a tiered set of goals for our project based on the macro averaged F1 score, which provides a balanced measure of precision and recall across all classes, including rare ones.

- **Minimum Success:** Achieve a macro averaged F1 score greater than 0.80.
- **Target Success:** Achieve a macro averaged F1 score greater than 0.85.
- **Exceptional Success:** Achieve a macro averaged F1 score greater than 0.90.

Meeting the target success criterion would demonstrate that our model is highly effective and robust for the task of multilabel aerial image classification.

References

[1] G.-S. Xia, J. Hu, F. Hu, B. Shi, X. Bai, Y. Zhong, L. Zhang, and X. Lu, "AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 7, pp. 3965-3981, July 2017.

[2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778.

[3] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 4700–4708.