Transfer Learning Document: NLP, Image Classification, and Time Series Forecasting

Overview

This document explores how **transfer learning** can be applied across three domains:

1. **Natural Language Processing (NLP)** – text classification using sentiment data.
2. **Computer Vision (CV)** – image classification using pre-trained CNNs.
3. **Time Series Forecasting** – predicting gold prices using sequential models.

The aim is to demonstrate how pre-trained models and advanced architectures can be adapted for new tasks with minimal training data, improving both efficiency and accuracy.

Datasets

1. **NLP (Text Classification)**
    - Dataset: UCI Sentiment Labelled Sentences
    - Alternative: IMDB movie reviews (Kaggle).
2. **Computer Vision (Image Classification)**
    - Dataset: 5 Flower Types Classification Dataset (Kaggle)
    - Alternative: CIFAR-10 or custom datasets.
3. **Time Series (Gold Price Forecasting)**
    - Dataset: XAUUSD Gold Price Historical Data (2004–2024)

- Alternative: Fetch live financial data using `yfinance`.

Workflow

1. Data Understanding

NLP:

- Inspect sentiment dataset (positive vs negative labels).
- Analyze word counts and class distribution.

CV:

- Visualize images from different categories.
- Check dataset balance across flower classes.

Time Series:

- Load gold price dataset.
- Convert `Date` to datetime, set as index.
- Explore columns: Open, High, Low, Close, Volume.

2. Exploratory Data Analysis (EDA)

NLP:

- Plot class balance.
- Visualize frequent words with word clouds.

CV:

- Display sample images per class.
- Plot number of images per category.

Time Series:

- Plot closing prices over time.
- Calculate and plot moving averages (7-day, 30-day).
- Visualize daily returns for volatility.

3. Preprocessing

NLP:

- Tokenize text using Hugging Face `AutoTokenizer`.
- Convert to input IDs and attention masks.
- Train/test split (80/20).

CV:

- Resize images to 224×224.
- Normalize pixel values.
- Apply augmentations (rotation, flip, zoom).

Time Series:

- Scale values using `MinMaxScaler`.
- Create sliding window sequences (e.g., past 60 days → next day).
- Reshape into `(samples, timesteps, features)` format.

4. Model Building

Baselines:

- NLP: Logistic Regression / RNN.
- CV: Small CNN trained from scratch.
- Time Series: Basic RNN.

Transfer Learning Models:

- **NLP (Transformers)**
    - Fine-tune `bert-base-uncased` using Hugging Face `Trainer`.
    - Evaluate using accuracy and F1-score.
- **CV (Pre-trained CNNs)**
    - Use ResNet50, VGG16, or MobileNet as feature extractors.
    - Add dense classification head for flower categories.
- **Time Series (Gold Price)**
    - Build LSTM and GRU models for forecasting.
    - Use dropout, learning rate schedulers, and callbacks.

5. Evaluation

- **NLP & CV:** Accuracy, Precision, Recall, F1-score, and Confusion Matrix.
- **Time Series:** RMSE, MAE, predicted vs actual price plots.
- Compare **baseline vs transfer learning** results.

- Visualize training and validation curves.

Key Takeaways

- Transfer learning reduces training time and improves performance across domains.
- Pre-trained models capture **language semantics, image features, and temporal dependencies** effectively.
- Fine-tuning strategies (freezing/unfreezing layers, adjusting learning rates) are crucial for optimal results.
- Applications span **sentiment analysis, image recognition, and financial forecasting**, showcasing the versatility of transfer learning.

References

- 5 Flower Types Classification Dataset (Kaggle)
- XAUUSD Gold Price Historical Data (Kaggle)
- UCI Sentiment Labelled Sentences Dataset