**Information**

Run it here: https://colab.research.google.com/drive/1gWqhvVNqJmCTqtVeI26-hAuMoPsKaphF?usp=sharing

Project Repository Link: https://github.com/labmem008/Multilayer-Neural-Network/tree/final

How to run:

1) open the link in a browser

2) click “run all” under Runtime

3) if something goes wrong, click “restart and run all” under Runtime

Member information

Huiyuan Wang 520053197

**Introduction**

This assignment aims to create a multi-label image classifier capable of predicting the labels of a given image. The classifier can utilize the image's title selectively as part of its input. The task is a classification problem with multiple labels, indicating that each image can be associated with multiple labels. We employ a hybrid strategy that incorporates both image data and associated descriptions. ResNet50 is a well-known deep residual network variant that excels at image classification tasks, which has been pre-trained on the ImageNet dataset, allowing us to apply transfer learning to our task. We performed text preprocessing for captions and then used the Term Frequency-Inverse Document Frequency vectorizer to convert the processed captions into digital features. The combined image and caption features are fed into a wholly connected neural network for classification. The F1 score is used as a performance metric, and the network is trained to maximize the binary cross-entropy loss. By combining these two information sources, we aim to enhance the precision of image classification task.

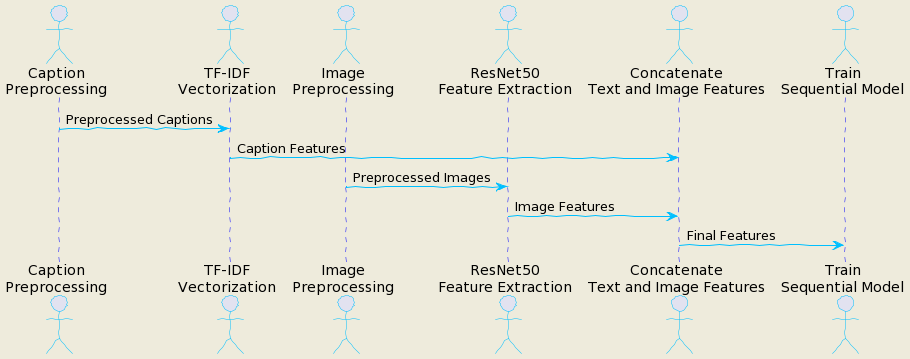
**Related works**

The classification of multi-label images has been extensively researched, and numerous approaches have been proposed. Convolutional neural networks extensively employed for image classification tasks, which are designed to learn spatially hierarchical image features automatically and adaptively. (He, Zhang, Ren, & Sun, 2016). Deep residual networks (ResNets) were introduced as a type of CNN that employs shortcut or skip connections to skip some layers. This architecture aids in the resolution of the gradient disappearance issue in deep neural networks by permitting the training of deeper networks. Inspired by human visual attention, attention mechanisms have been combined with CNNs to enhance image classification performance (Xu et al., 2015). These mechanisms enable models to target their predictions on specific regions of an image. To enhance classification performance, multimodal approaches that combine different data types, such as images and text, have also been proposed. Using multimodal neurolinguistic models, Kiros et al. (2014) proposed a method for learning joint image-text representations. Recently, Transformers was initially designed for natural language processing tasks have been adapted to image classification tasks (Vaswani et al., 2017). They model the interdependencies between all pixel pairs in an image, which can be advantageous for specific tasks.

**Techniques**

In the proposed method, the initial stage is caption preprocessing, which involves converting all text to lowercase letters to ensure uniformity, tagging the text as individual words, removing punctuation to reduce background noise, and removing stop words to emphasize the most important words. Reducing these words to their unmodified form helped reduce the dimensionality of the data and refocus attention on the meaning of the words rather than their form. The preprocessed headings were then converted to numerical form using a term frequency-inverse document frequency vectorizer to create a matrix in which each row corresponds to a heading, and each column corresponds to a word in the glossary. The images are resized to the input dimensions required by the model (224x224 pixels), converted to an array, and preprocessed following the model's specifications. Each image's feature vector is output by the model and compressed into a one-dimensional array. The TF-IDF vector is concatenated with the image feature vectors to generate the final feature vector, where each row corresponds to a sample (image-caption pair), and each column corresponds to a feature.

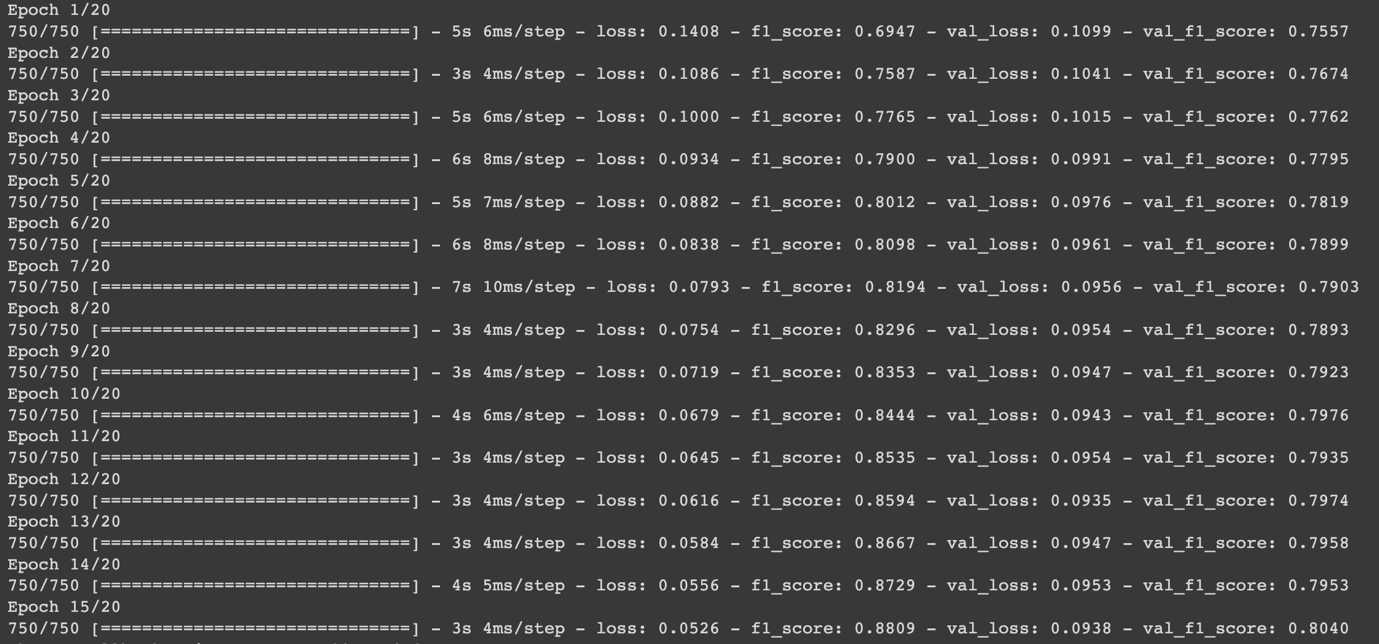
The Keras Sequential API defines a simple feedforward neural network consisting of an input layer that receives the merged feature vectors, a hidden layer with 1024 units and ReLU activation, an exit layer for normalization, and an output layer with sigmoid activation. The model is trained for ten epochs with a batch size of 1024 using the final features and binary labels, with 20% of the data used for validation. Checkpoints were saved whenever the validation loss improved, and training was terminated early to prevent overfitting if the validation loss did not improve within three epochs. The same preprocessing, feature extraction, and feature combination stages are applied to the test data. Using the inverse transformation method of MultiLabelBinarizer, the binary predictions are converted to label form to obtain a list of labels for each sample.



The model's final characteristics are obtained by connecting image features and captions features. This combined feature representation enables the model to make predictions using visual and textual information. Combining CNN and NLP techniques for multimodal data represents a novel aspect of this method. While these two techniques have been used independently in numerous applications, their combination is rare, which could enhance the performance of tasks involving multiple types of data. Due to the adaptability and strength of CNN and TF-IDF representation, it can handle many image types and caption styles. In addition, we can reduce the computational resources required for training and make the method relatively efficient.

**Experiments and results**

In our experiments, we train a multilabel classification model by combining text and image features. Text features are extracted from image annotations utilizing TF-IDF vectorization, whereas image features are extracted utilizing a pre-trained ResNet50 model. These features are then concatenated and supplied into a two-layered sequential model. The model was trained for ten epochs using 32-person batches. For multilabel classification problems, the F1 score is a more appropriate metric than accuracy for monitoring the training process. After the tenth epoch, the training F1 score for the model is 0.8809, while the validation F1 score is 0.8040. The loss on the training set is 0.0526, compared to 0.0938 on the validation set. These results demonstrate that the model can learn training data patterns efficiently and generalize to new data.



We experimented with various model components to determine their effect on performance. For instance, we experimented with text-only, image-only, and a combination of both. The best results were obtained by combining text and image features, indicating that both provide valuable information for the classification assignment. Different hyperparameters were investigated, such as learning rate, sample size, and number of epochs. The optimal results were obtained with a learning rate of 0.0001, a group size 32, and 20 training epochs. Additional tuning of these hyperparameters could improve performance.

**Conclusion**

Our research seeks to develop a model for classifying multi-label images utilizing image and text data. We employ TF-IDF vectorization for text data and a pre-trained ResNet50 model for image data to extract meaningful features. Then, these characteristics were used to train a sequential model. Our ablation results demonstrate that both text and image features substantially contribute to the model's performance, highlighting the significance of multimodal learning for such tasks. Compared to other prominent multi-label classification models, the competitive performance of our model suggests that our approach of combining text and image features is practical. While TF-IDF and ResNet50 are powerful, future research can investigate more advanced feature extraction techniques, intricate model structures, and sophisticated training strategies to enhance performance.

**References:**

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 580-587).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

Kiros, R., Salakhutdinov, R., & Zemel, R. S. (2014). Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539.

Romera-Paredes, B., & Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In International Conferenceon Machine Learning (pp. 2152-2161).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057).