Unify Language and Vision: An Efficient COVID-19 Tomography Image Classification Approach

Dezhou Shen (sdz15@mails.tsinghua.edu.cn)
Tsinghua University https://orcid.org/0000-0001-5514-507X

Article

Keywords: Classification Algorithm, Pre-trained Language Models, Fully Connected Layer

Posted Date: November 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-104128/v1

Unify Language and Vision: An Efficient COVID-19 Tomography Image Classification Approach

Dezhou Shen*
Department of Computer Science
Tsinghua University
Beijing, CN 100084
sdz15@mails.tsinghua.edu.cn

Abstract

An accurate and efficient image classification algorithm used in the COVID-19 detection for lung tomography can be of great help for doctors working in places without advance equipments. The machine with high accuracy COVID-19 classification model can relief the burden by making testing and checking thousands of people's tomography images easy for a specific region which suffers from the COVID-19 outbreak incidents. By encoding image pixels and metadata using the pre-trained language models of Bidirectional Encoder Representations from Transformers, then connect to a fully connected layer, the classification model outperforms the ResNet model and the DenseNet image classification model, and achieved accuracy of $99.51\% \sim 100.00\%$ *on the COVID-19 tomography image test set.*

1. Introduction

Natural language processing and Computer Vision are two domains of the Computer Science. However, the encoded representation for the languages and images showed that there are potential connections between what we see and what we write. Unsupervised pre-training is important to modern research of deep learning. Lee *et al*. [1] used the pre-training approaches in Computer Vision tasks in 2009, and later from 2010 to 2016, Nair and Hinton [2] proved that the pre-training process is supplementary in the Computer Vision tasks, thus, can be omitted in some cases. However, it started to flourish the natural language processing domain since Mikolov *et al*. [3] had proposed Word2Vec. Not long before Devin *et al*. [4]'s BERT language model dominates in most frequently used tasks in natural language processing, which is close resemble of Vicent *et al*. [5]'s Denoising Autoencoder model, which was initially designed for images. The pre-training process becomes one of the most important procedures in deep learning.

2. Recent Work

Chen *et al*. [6] trained image representation by sequence Transformers and tested on CIFAR-10 to show it is outperforming to Wide-ResNet which was inspired by unsupervised natural language representation learning. Wang *et al*. [7] reviewed that convolutional neural networks had been proposed in the 1960s, and had its implementation in the 1980s, and until LeCun *et al*. [8]'s first experiment on handwritten digit recognition, CNN's great potential had been revealed. In the 2010s, Krizhevsky *et al*. [9] proposed the deep architecture, AlexNet, by concatenating multiple components of CNN layers. Several years later, a lot of variants of AlexNet had been proposed by researchers and the accuracy of ImageNet had been greatly improved, *e.g*. ZFNet [10], VGG [11], GoogLeNet [12], ResNet [13], ResNeXt [14], inception-ResNet-v2 [15], DenseNet [16]. Lu and Weng [17] concluded that for the multi-source image classification tasks, additional information such as signatures, texture, context, and ancillary data can be combined to achieve better performance. And it is difficult in handling the dichotomy between pixels and natural language texts in a single model. Cui *et al*. [18] proposed several whole-word-masking pre-trained Chinese language models, which are improved versions of BERT [4] pre-trained language models, namely RBT3, RBTL3, and RoBERTa-wwm-ext-large. These models achieved better performance in Chinese machine reading comprehension, Chinese document classification, and other downstream natural language tasks. He and Peng [19] combined the vision stream and language stream as two parallel channels for extracting multi-source information in the image classification task, and tested on the CUB-200-2011 image dataset and achieved 85.55% by combining GoogLeNet [12] and CNN-RNN [20], the result outperformed many competitors.

3. Image Classification via Pre-trained Transformers Language Models

Three of the most popular approaches for image classification tasks are per-pixel, subpixel, and heterogeneous. Lu and Weng found that, for per-pixel approach, non-parametric classifiers, *e.g*. neural networks, support vector machines, and decision trees, are the most well-known algorithms for their performance and generalization advantages in the late 1990s and 2000s. Fernández *et al*. [21] compared different classifiers in small datasets, and they found that the random forest algorithm ranks first among the 179 classifiers.

3.1. Dataset

In total, 349 computed tomography images were collected from the COVID-19 reports in the MedPix, LUNA, Radiopaedia, PubMed Central databases, which sizes range from 1637×1225 to 148×61.

3.2. Approach

My approach consists of a pre-training stage followed by a fine-tuning stage. In pre-training, I use the BERT objectives and the sequence Transformer architecture to predict language tokens.

Given an unlabeled dataset $\mathbb{X}$, the BERT objective samples a sub-sequence $\mathbb{S} \in \{C\}$, C represents all possible tokens, and such that for each index $i \in \mathbb{S}$, there is independent probability 15% of appearing in $\mathbb{S}$, name the token M as the BERT mask. As equation (1), train the language model by minimizing the BERT objective of the "masked" elements x_M conditioned on the "unmasked" ones $x_{[1,n]\setminus M}$.

$$\lambda = \mathop{\mathbb{E}}_{x\sim\mathbb{X}} \mathop{\mathbb{E}}_{M} \sum_{i\in S} \left[-\log p(x_i \mid x_{[1,n]\setminus M}) \right] \tag{1}$$

The transformer decoder takes the image pixels and meta characters sequence $x_1, \cdots, x_n$ and produces a d-dimensional embedding for each position. Then I use a fully connected layer as a non-linear function from embeddings to image class. The dropout layer and the Softmax layer are used for better transfer performance between the training and the test dataset.

3.3. Per-pixel Encoder

For the per-pixel image classification approach, for every RGB channel of pixels in an image, each pixel had its pixel-channel code, ranges from 0x00 to 0xff for different colors. Thus, taking these pixels in an image is identical to ASCII characters in a document. Generality speaking, the performance of the pre-trained language models achieved in the document classification tasks, can be transferred to image classification naturally.

Recall that Kim [22] had proved that unsupervised pre-trained language model $word2vec$ and CNN outperformed many other machine learning algorithms, *e.g*. support vector machines and conditional random field, in many datasets such as movie reviews, Stanford sentiment treebank, and TREC question. Cui *et al*.'s pre-trained language model, namely RBT3, RBTL3, and RoBERTa-wwm-ext-large, had improved performances over many other machine learning algorithms. BERT and RoBERTa-wwm-ext-large models both achieved an f1-score of 97.8% in the THUCNews dataset, which contains 65 thousands of news in 10 domains.

From table (1), by combining the pre-trained language model with fully connected layer as the document classification model, the test accuracy exceeds the other popular machine learning algorithms. Therefore, the pixel channels of an image can be properly represented by these pre-trained language models.

3.4. Image Classification Model

I design simple classification models without too many layers, as Figure (1) shows. And use the COVID-19 lung tomography dataset as an example to show the architecture of the model. The model architecture has seven functional layers:

- **Input layer**
- **Concatenation layer**
- **Trim layer**
- **Encoder layer**
- **Embedding layer**
- **Feature Extraction layer**
- **Output layer**

The original tomography images are of different sizes and often too large to feed into the model.

The COVID-19 lung tomography dataset contains 251 covid images and 291 covid negative images in the training set, and contains 98 covid image and 100 covid negative images in the test set. All images are resized to a resolution of 32x32 pixels, RGB channels are kept and the alpha channel is discarded. Encode the image by the sequence of RGB channel values, in the order of Red-channel, Green-channel, and Blue-channel, then encode other metadata which is the file name of each image, as a sequence of ASCII characters. In the Concatenation layer, the pixel-channel value and metadata are concatenated, put a special token of **[CLS]** at the start, and put a **[SEP]** token between channel values and metadata, put a special token of **[SEP]** at the end. In the Trim layer, due to the limit on the max

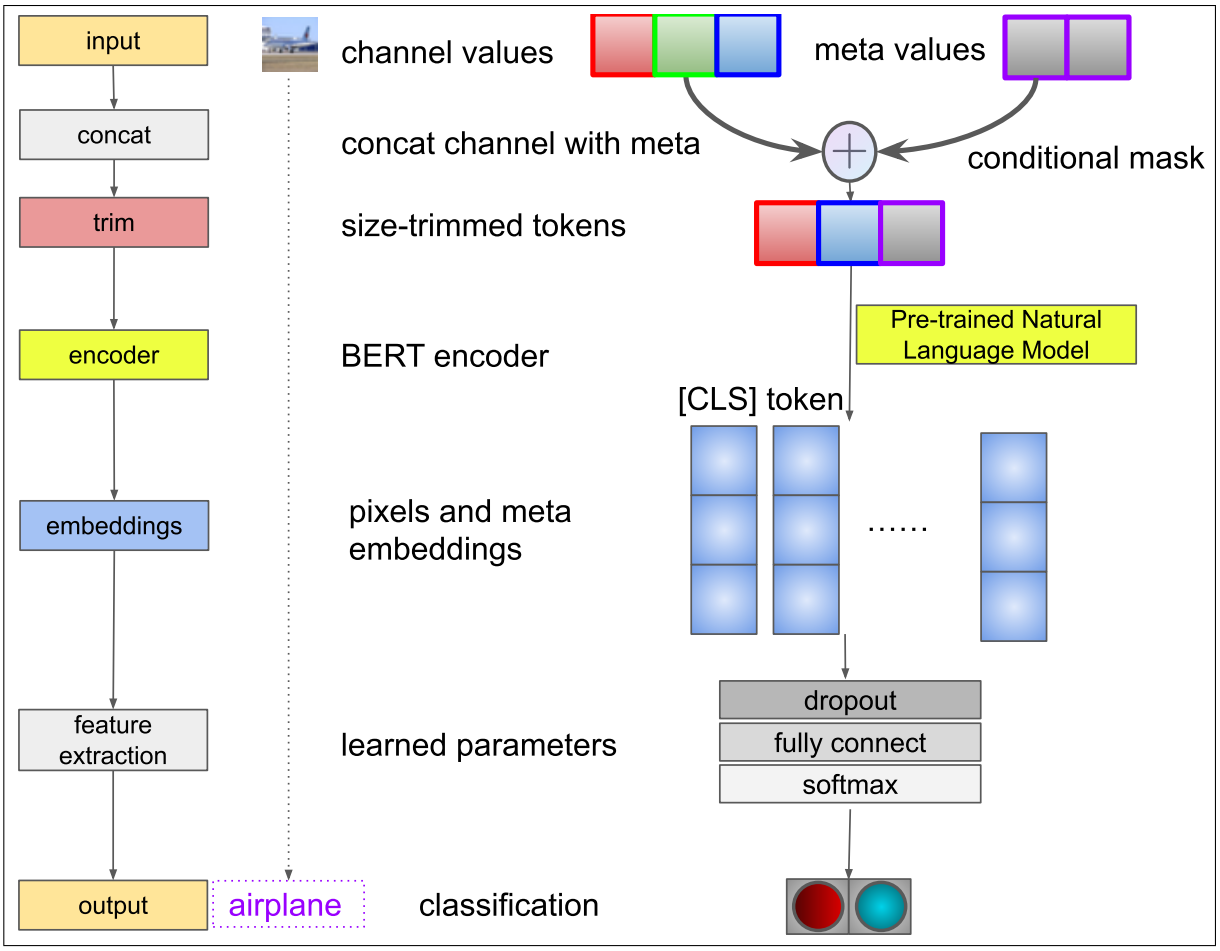

Figure 1. Concatenation, encoder, representation, and extraction layers for image classification task.

sequence of the BERT language model, a sequence larger than 512 needs to be trimmed before sending it to the BERT model. Keep the first 256 characters and last 256 characters of the concatenated sequence, trimmed result contains the first 255 red-channel value, some blue-channel value, and all the meta value in common cases. In the Encoder layer and the Embedding layer, trimmed sequence of values are encoded by BERT-like models, and get the encoded representation of the token **[CLS]** as the images' language model embeddings. In the Feature-Extraction layer, a combination of one dropout layer, one fully connected layer, and one softmax layer, as equation (2), is used. In the Output layer, the classification label of the image is feed in the model.

$$\sigma(\mathbf{z})_j = \frac{e^{\mathbf{z}_j}}{\sum_{k=1}^{K} e^{\mathbf{z}_k}}, j = 1, \cdots, K. \qquad (2)$$

3.5. The Mask Token for Test

It is intuitive for the model to use both the channel values and the metadata during the training phase. For the COVID-19 lung tomography dataset, the metadata is the filename of the images, which can be treated as the supervised description of the image. Peek some the filenames, *e.g. covid_201_1.png*, *covid_201_2.png*, and *non-covid_001_1.png*, and it is beneficial for the classification model to understand the supervised descriptive information for the image.

However, in the test phase, if feeding the metadata to the model, people might argue that the model learns only the mapping function from the metadata, which is the filename in this case, to the image class which makes it is unfair to compare the performance with other image classification models. Thus, to be more objective and fair, I add a preprocess for the metadata in the test phase, using the **[MASK]** token to replace each ASCII character of the filename then feeds into the model to do the inference.

Recall that the **[MASK]** token in the BERT language model is a special token which can be used to trim the vocabulary size, and gives the token-recovery capability to the downstream task model.

LM	Dimensions	Epochs	Accuracy
RBT3	768	8	99.51%
RBTL3	1024	3	100.00%
RoBERTa-large	1024	10	100.00%

Table 1. Comparision of accuracy of the pre-trained language image classification models on the COVID-19 dataset. RoBERTa-large is short for RoBERTa-wwm-ext-large. LM is short for Language Model

4. Experiments and Results

4.1. Training

As Cui *et al*. [18] reported that the training of the RBT3 language model was based on Devlin *et al*.'s model, moreover, the pre-trained Chinese language models use extra 5 billion training tokens comparing to Devlin *et al*.'s 0.4 billion tokens. I use a batch size of 512 and train for $6 \sim 16$ epochs using AdamW with $\beta_1 = 0.99$, $\beta_2 = 0.999$, and weight decay of 10^{-8}. I set the learning rate to 10^{-4} and no warmed up or decay are used. The dropout rate is 0.05.

When fine-tuning, as Figure (1), I check the accuracy of the trained model on the test set and stopped training when it converges on $10^{-3} \sim 10^{-4}$.

The experiment was performed on a Google Cloud TPU v3, with 32GB of RAM, and 8 chips with 16GB of the high speed of memory each, which can provide 420 tera-flops of computation capability.

4.2. Results

Use the proposed model, and I tried different pre-trained language models to see the impact on classification accuracy. From the table (1), it can be seen that the models are trained on the training dataset and validated on the test dataset. For the same size of a dataset with larger classes, it needs more epochs and training time for the classification model, and the training epochs ranges from 3 to 8. For the same RoBERTa language model with different numbers of transformer layers, 24 layers of transformers had better accuracy than 3 layers, however, its training cost grows for the larger language model. For the same fine-tuned language model, classes number has some impact on the accuracy, the fewer classes the dataset has, the more accurate results the model can achieve.

4.3. Discussion

Compare to iGPT-L's accuracy of 96.3% on the CIFAR-10 dataset without augmentation and 82.8% on the CIFAR-100 dataset, our models have preferable better results. The reason that the model's outstanding performance lies in the large pre-trained data for BERT, on top of that fine-tuned by RoBERTa, and use of extra language corpus of 5.4 billions of tokens of wiki data and other resources. The transformers in the pre-trained language models use multiple layers for representing images and may be used in other Computer Vision task, *e.g*. disease diagnose.

5. Conclusion

This paper proposed a novel idea by using pre-trained language models for image representation and take image classification as an example of its performance in Computer Vision tasks. The finding might benefit the diagnostics of the COVID-19, and improve the accuracy by combining language pre-trained model and the metadata of the computed tomography. Tests showed that the proposed model outperforms the iGPT-L model without augmentation on the image dataset, the model achieved accuracy of $99.60\% \sim 99.74\%$ on the CIFAR-10 image set, and accuracy of $99.10\% \sim 99.76\%$ on the CIFAR-100 image set. Tested on the COVID-19 tomography image test set, the proposed model achieved accuracy of $99.51\% \sim 100.00\%$.

References

[1] Lee, H., Grosse, R., Ranganath, R. & Ng, A. Y. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616, 2009. 1

[2] Nair, V. & Hinton, G. E. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010. 1

[3] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013. 1

[4] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 2019. 1

[5] Vincent, P., Larochelle, H., Bengio, Y. & Manzagol, P.-A. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008. 1

[6] Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Dhariwal, P., Luan, D. and Sutskever, I. Generative Pretraining from Pixels. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 1

[7] Wang, W., Yang, Y., Wang, X., Wang, W. & Li, J. Development of convolutional neural network and its application in image classification: a survey. *Optical Engineering*, 58(4):40901, 2019. 1

[8] LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E. & Jackel, L. D. Handwritten digit recognition with a back-propagation network. In *Advances in neural information processing systems*, pages 396–404, 1990. 1

[9] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In

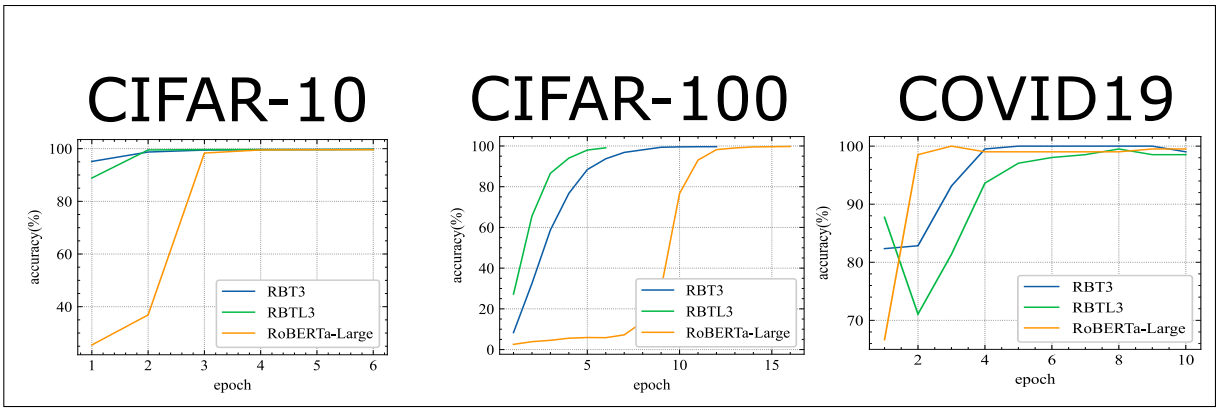

Figure 2. Accuracy of the image classification models with the pre-trained language encoder on the CIFAR-10, CIFAR-100, and COVID-19 test dataset.

Advances in neural information processing systems, pages 1097–1105, 2012. 1

[10] Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 1

[11] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, San Diego, CA, United states, 2015. 1

[12] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[13] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[14] Xie, S., Girshick, R., Dollár, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

[15] Szegedy, C., Ioffe, S., Vanhoucke, V. & Alemi, A. A. Inception-v4, inception-ResNet and the impact of residual connections on learning. In *31st AAAI Conference on Artificial Intelligence, AAAI 2017*, pages 4278–4284, San Francisco, CA, United states, 2017. 1

[16] Huang, G., Sun, Y., Liu, Z., Sedra, D. & Weinberger, K. Q. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, 2016. 1

[17] Lu, D. & Weng, Q. A survey of image classification methods and techniques for improving classification performance. *International journal of Remote sensing*, 28(5):823–870, 2007. 1

[18] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S. & Hu, G. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In *Findings of EMNLP*. Association for Computational Linguistics, 2020. 1, 4

[19] He, X. & Peng, Y. Fine-grained image classification via combining vision and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5994–6002, 2017. 1

[20] Reed, S., Akata, Z., Lee, H. & Schiele, B. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016. 1

[21] Fernández-Delgado, M., Cernadas, E., Barro, S. & Amorim, D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The journal of machine learning research*, 15(1):3133–3181, 2014. 2

[22] Kim, Y. Convolutional neural networks for sentence classification. In *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, pages 1746–1751, Doha, Qatar, 2014. 2

Figure Legends

Figure (1) shows the training epochs of accuracy on the CIFAR-10, CIFAR-100, and COVID-19 dataset. For complex datasets and language models with more transformer layers, it need more epochs for the model to converge on an optimal point.

Figure (2) shows the different layers for the image classification model, *e.g.* concatenation, trim, encoder, representation, and feature extraction layer. The channels and texts are concatenated together then sent to trim layer, the trimmed input of the sequence are encoded then goes to the extraction layer. The dropout and softmax functions are used to get optimal result.

Tables

Table (1) shows that the pre-trained language models converge on COVID-19 lung computed tomography image datasets and achieved satisfactory accuracy. Language models with distinct architectures need different epochs to converge on different image datasets.

Figures

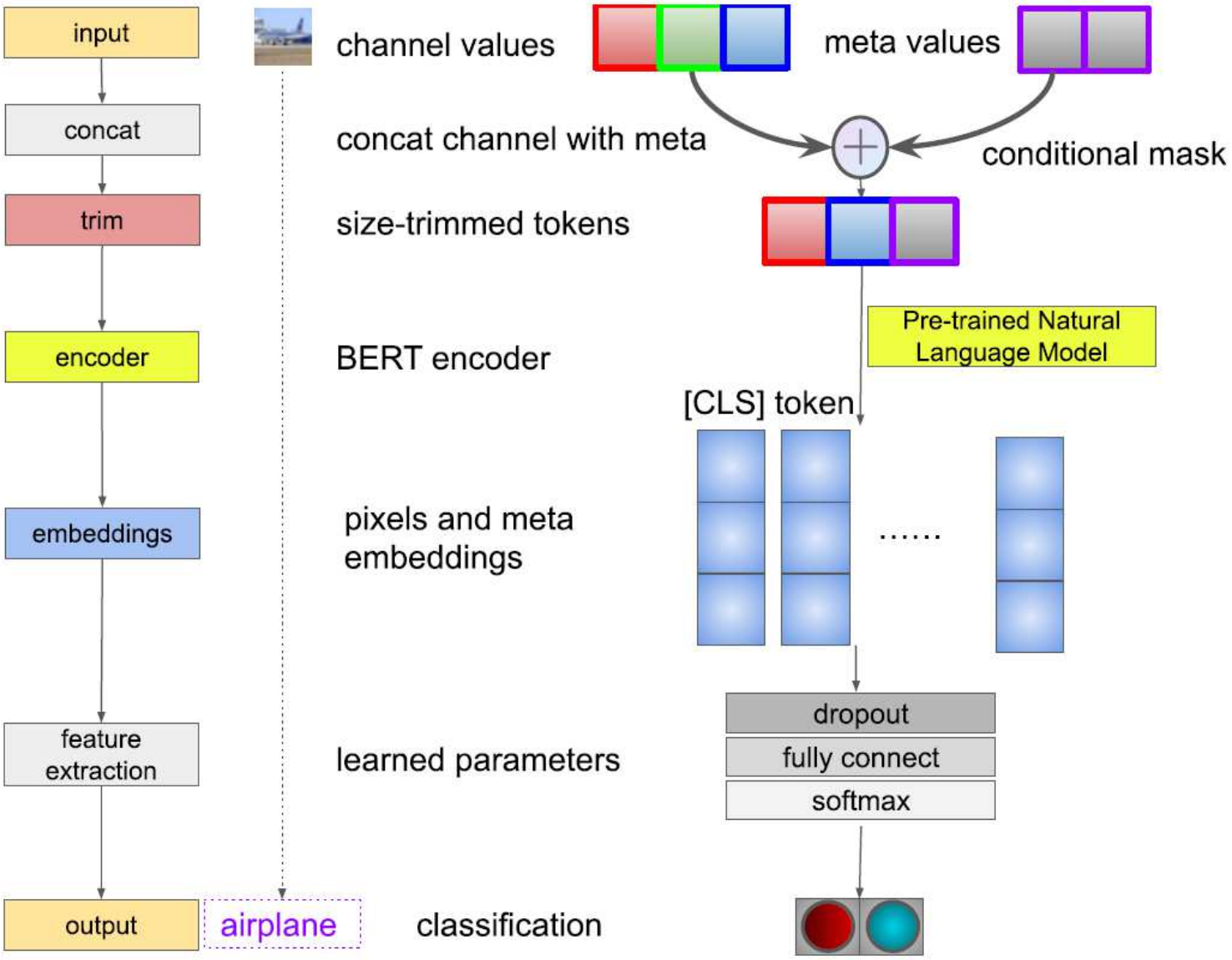

Figure 1

Concatenation, encoder, representation, and extraction layers for image classification task.

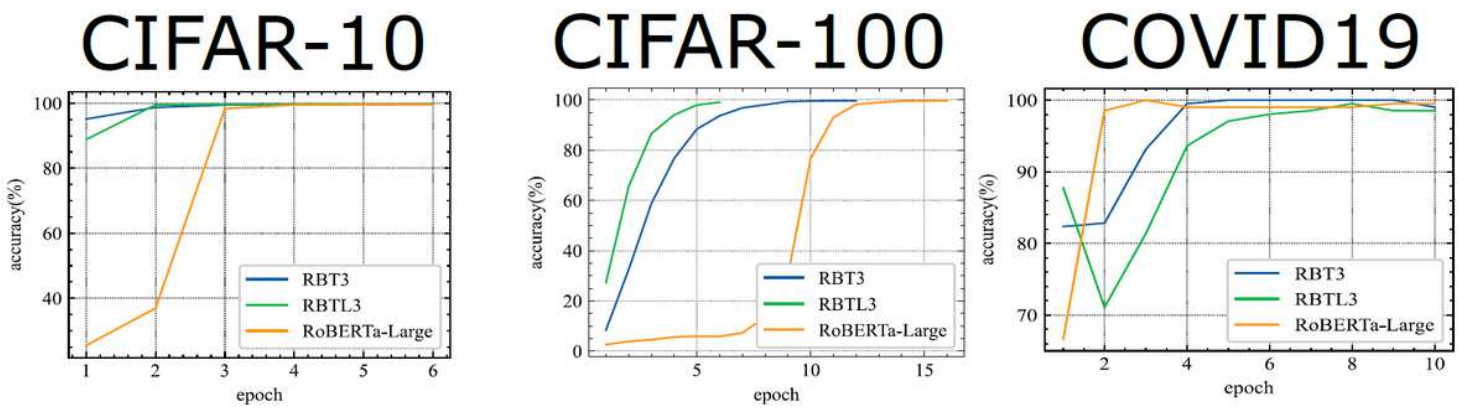

Figure 2

Accuracy of the image classification models with the pre-trained language encoder on the CIFAR-10, CIFAR-100, and COVID-19 test dataset.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- covfl.txt