Covid CXR Hackathon - COSBI: Valerio Guarrasi, Paolo Soda

Interpreting imaging findings is multimodal by its very nature, and hence, AI needs to be able to process together data coming from various modalities to get models best suited to solve the tasks at hand. The potential of deep-learning shown processing unimodal data has recently motivated the rise of multimodal deep-learning (MDL), which aims to treat multimodal data by using deep-network-based approaches. In the specific context of COVID-19 pandemic, MDL can support the patients' stratification mining together images and clinical information.

The literature agrees that the three main open questions in MDL are how, when and which modalities must be joined to exploit the potential of each modality. Usually, MDL models are constructed by finding or combining the best model architecture for each modality which are then combined based on the nature of the data, on the task at hand and on the networks' structure by not obtaining necessarily the best ensemble. Therefore, we propose a novel joint fusion technique that automatically can find the optimal solution to which architectures must be combined, whatever the modality they belong to, maximizing the performance by also considering their diversity, exploring a novel locus of when the modalities are joint. To comprehend the validity of the proposed method, we applied it to the prognosis of the severity of COVID-19 positive patients exploiting both clinical and CXR scans, with the goal of improving the stratification of the disease.

Our optimized joint end-to-end framework is supported by a combination of the Explainable artificial intelligence (XAI) techniques illustrating the reasoning behind the decisions taken by the model, since a key impediment to using DL-based systems in practice is their black-box nature that does not permit to directly explain the decisions taken. In this way we improve trust and transparency by showing the relative contribution of each modality in making the decision.

Pre-processing

Regarding the clinical data, missing data were imputed using the mean and the mode for continuous and categorical variables, respectively. Finally, to have the features all in the same range, a min-max scaler was applied along the variables.

In the case of CXR scans we extracted the segmentation mask of lungs, using a pre-trained U-Net on two non-COVID-19 datasets: Montgomery County CXR set and the Japanese Society of Radiological Technology repository. The mask was used to extrapolate the minimum squared bounding box containing both lungs, which is then resized to 224x224 and normalized with a min-max scaler bringing the pixel values between 0 and 1. Since the data comes from a real-world emergency scenario, manual modifications of the scale of gray and inclination of scans were made to standardize the data both for the training and test set, as specified by the organizing committee.

Classification

To predict the severity outcome of COVID-19 patients exploiting both the pixel information of their CXR scan and their clinical information, we propose a novel supervised end-to-end joint fusion method, which maximizes the advantages brought by all the given modalities.

It first looks for the best combination of models, which are then joint and trained to carry out the given classification task. We consider the chance of having several models for each modality, a situation that researchers and practitioners usually meet in practice.

For the image modality we worked with 30 different CNNs that come from 8 different main architectures: AlexNet, VGG11, VGG11-BN, VGG13, VGG13-BN, VGG16, VGG16-BN, VGG19, VGG19-BN, ResNet18, ResNet34, ResNet50, ResNet101, ResNet152, ResNeXt50, ResNeXt101, Wide-ResNet50-2, Wide-ResNet101-2, DenseNet169, DenseNet161, DenseNet201, GoogLeNet, ShuffleNet-v2-x0-5, ShuffleNet-v2-x1-0, ShuffleNet-v2-x1-5, ShuffleNet-v2-x2-0, MobileNetV2, MNasNet0-5 and MNasNet1-0. In all the cases the weights were initialized using the values pre-trained on the ImageNet dataset; we also changed the output layer dimension to 2 neurons, one for each class.

In the case of clinical information, which are tabular data, we adopted 4 multi-layer perceptrons (MLPs) that differ in terms of depth and wideness of the model. In particular, the models' hidden layers have the following organizations:

- MLP-1: it has 3 hidden layers with 64, 64, 32 neurons;
- MLP-2: it has 5 hidden layers with 64, 128, 128, 64, 32 neurons;

- MLP-3: it has 7 hidden layers with 64, 128, 256, 256, 128, 64, 32 neurons;
- MLP-4: it has 9 hidden layers with 64, 128, 256, 512, 512, 256, 128, 64, 32 neurons;

A ReLU activation function is applied on all layers.

The first step is to detect which is the subset of models to be combined to get the best multimodal architecture providing the best multimodal data representation. We first train and evaluate in cross-validation each model on the corresponding unimodal dataset. Then, we calculate both an average evaluation metric and diversity metric of all possible ensembles on the validation set and find the optimal ensemble which maximizes both metrics. This means that we look for the ensemble that, on the one side, returns the best classification performance and, on the other side, reduces the incidence and effect of coincident errors among its members, this considering possible relationships between models and modalities. We adopt the accuracy and the correlation coefficient as the two objective functions. The accuracy is calculated on the models part of the ensembles in late-fusion via a majority voting aggregation function.

Once the models of the optimal ensemble are found, we concatenate the classification vectors, i.e. the output layers of the deep neural networks. This concatenated vector is then passed to a FC neural network, which has in the last layer a number of neurons equal to the number of possible classes. In this intermediate fusion, we exploit the advantages of both joint and late fusion, since the classifications of the sub-networks are aggregated via an end-to-end manner using the back-propagation during the training process that minimizes the overall loss function.

As a result of the optimization we obtain the optimum ensemble composed of two CNNs and one MLP: DenseNet121, VGG11-BN and MLP-1. By observing the results we notice that there is no redundancy in the model families, because each model extrapolates different importance information to satisfy the desired classification. Moreover, since we have at least one model for each modality, we understand that they all give useful and distinct information for the prognosis task. For computational restrictions for the joint fusions, we limited the number of possible architectures in the ensembles for each modality to a max of 3.

XAI: available at the following **link**
To open the black-box nature of the joint model, we extract the weights coming out of the classification vector, and by observing their mean relative intensities in CV, their distribution is the following: 49%, 49% and 2% for DenseNet121, VGG11-BN and MLP-1, respectively. This information not only makes us understand the importance of every single model for the final classification, but it also explains the hierarchy inter- and intra-modality. In particular, we notice that the image modality has more importance than the clinical one since the relative vector weights are 98% and 2% respectively. Results of the multimodal weights can be found at: xai/multimodal/weights.xlsx.

To enable physicians to explore and understand data-driven DL-based systems, we decided to work on XAI algorithms for single modalities. For each model composing the joint fusion, we can apply XAI algorithms which were realized for their specific modality. Considering only the clinical data we applied to MLP-1 the integrated gradients algorithm to show the features with their corresponding importance for a particular classification. Results of the clinical feature importance can be found at: xai/clinical.

Going forward, we can use the relative aforementioned weights of the classification vector for a specific modality to combine the results coming from a XAI algorithm. In xai/img/densenet121 and xai/img/vgg11_bn we show the feature maps extracted from DenseNet121, VGG11-BN by applying Grad-CAM, and by combining them via weighted normalized sum we obtain a resulting map for the modality: xai/img/densenet121;vgg11_bn. In this way we understand how the different components working on the image modality, and as a whole, interpret the pixel values, as shown in the following figure: