Machine Learning strategies have become an integral part of the biomedical informatics domain to better model, forecast, and classify complex data. For classification tasks on such data, researchers in the biomedical field have often opted to use more traditional machine learning techniques (Random Forest, Support Vector Machines, Logistic Regression, etc.) as they are inherently more explainable in terms of the methods through which the classification is generated.

More recently, however, the fast-growing field of deep learning (DL) has gained traction, particularly in the ability to accommodate a variety of tasks (binary classification, multiclass classification, multilabel classification, regression, etc.) and data modalities (image, text, tabular, speech, etc.).

Furthermore, these models tend to vastly outperform traditional models on larger datasets.

While deep learning models are notably less interpretable (often considered “black box” methods), the higher accuracies presented by these models have rendered the need for more interpretable models obsolete.

In order to construct a CNN as part of AutoMLPipe-DL, we will use the SKORCH library. We use the DeepInsight implementation of CNNs in AutoMLPipe-DL, which applies the t-distributed stochastic neighbor embedding (t-SNE) method for feature visualization. This method reduces the dimensionality of the feature vectors so that they can be easily visualized.

Fig. 9. A binary-valued classification is generated.

Several CNN architectures designed for image classification tasks.

- The convolutional layer in CNN.
- The pooling layer in CNN.
- The fully connected layer in CNN.

The network learns in gradient descent through backpropagation, in which a chain rule of partial derivatives is used to calculate the gradient iteratively computed to update the model.

Analysis and Results on Benchmark Data

We used the functionality of the implemented deep learning algorithms on the UC Irvine Hepatocellular Carcinoma (HCC) dataset (Table 1).

Successful implementation of the AutoMLPipe-DL pipeline with 12 deep learning models (KNNClassifier, SKORCH MLP, CNN, Supervised TabNet, Semi-Supervised TabNet, RM with a different downsampled feature set, etc.), as well as a number of baseline models (MLPClassifier, SKORCH MLP, CNN, Supervised TabNet, Semi-Supervised TabNet, RM with a different downsampled feature set, etc.), achieved promising results. The Table 1 implementation was used either on a supervised or semi-supervised approach, whereas the unsupervised pre-trainer for representation learning is followed by a supervised fine-tuner through a masked self-supervision procedure (Fig. 3).

Future Work:

- Continue testing and debugging of SKORCH models to integrate them in the pipeline and apply Opaque hyperparameter optimization for all models.
- Expand upon the Opaque model implementation to incorporate other notable OPAW architectures, including TabNet, Masker, flexnet, Inception-v3, and DenseNet.
- Implement the Transformer, an effective self-attention model which requires creating an implementation using SKORCH.
- Complete the minimal changes necessary to adapt AutoMLPipe-DL to multiclass classification tasks, and exercise performance on other biomedical datasets.

References:


Acknowledgements

This work was supported by the University of Pennsylvania Center for Undergraduate Research and Fellowships (CURF) as part of the FIRE program.

Lab/Resources

www.med.upenn.edu/urbslab/
www.ryanurbanowicz.com

Code Availability

github.com/Urbslab

www.med.upenn.edu/urbslab/AutoMLPipe-DL