EEG Classification Model

IE6400

—

Group 33
Akshit Dubey
Jaque Sarnoff
Rashmi Daga
Rishabh Madhani
Shuhan Li

Overview

The project at hand delves into the development of a sophisticated classification model for the analysis of electroencephalogram (EEG) data, with a primary focus on enhancing the diagnosis of epilepsy. Our exploration is grounded in the invaluable insights drawn from two robust datasets: the CHB-MIT EEG Database and the Bonn EEG Dataset. By leveraging these diverse sources, we aim to contribute to the advancement of technology in the medical field, particularly in the domain of neurological diagnostics.

EEG data, capturing the intricate electrical activity of the brain, holds immense potential for deciphering neurological conditions. With an overarching goal of harnessing machine learning to build a model capable of discerning and classifying seizures, our project aligns with the broader objective of addressing critical medical challenges through technological innovation.

Background Information

Epilepsy, a disorder of the central nervous system, poses multifaceted challenges for individuals experiencing recurrent seizures. These seizures, characterized by temporary deviations in the brain's electrical activity, occur unpredictably and without prior alert. Manifesting in various forms from momentary lapses in attention to full-body convulsions, the potential risks associated with seizures, including physical injuries and mortality, underscore the critical need for effective diagnostic tools and interventions.

The development of a device capable of swiftly detecting and responding to seizures is paramount in mitigating these challenges. The significance of our project extends beyond epilepsy alone, aiming to contribute to healthcare efficiency, patient safety, and the overall improvement of neurological diagnostics through advanced technological solutions.

Datasets Utilized

The Bonn EEG Dataset, alongside the CHB-MIT EEG Database, supplements our exploration into EEG-based epilepsy diagnosis. The Bonn EEG Dataset, detailed in Andrzejak et al.'s study (Phys. Rev. E, 64, 061907), presents a unique perspective on brain activity, offering opportunities for advanced analyses and diagnostic methodologies.

Objective

In this project, our primary objective is to construct a classification model capable of analyzing EEG data and categorizing it into distinct classes. Given the widespread application of EEG data in neuroscience and medical fields, particularly in the diagnosis of epilepsy, our aim is to contribute a model that transcends the limitations of individual datasets.

Leveraging the richness of both the CHB-MIT EEG Database and the Bonn EEG Dataset, our model aspires to encapsulate the diversity of EEG signals. By doing so, we seek to enhance the model's efficacy, robustness, and applicability, ultimately advancing the state-of-the-art in EEG-based diagnostic tools.

Significance of the project

The significance of our project lies in its potential to revolutionize neurological diagnostics. Swift detection of seizures through advanced machine learning models not only addresses the challenges posed by epilepsy but holds promise across a spectrum of disorders where EEG data can provide profound insights.

Outline

Our exploration encompasses data preprocessing, feature extraction, model architecture, training specifics, evaluation results, and avenues for future work. By meticulously documenting our journey through these sections, we aim to provide not only a comprehensive understanding of our methodologies but also insights that contribute to the broader field of EEG-based diagnostics.

Tasks

1. Data Preprocessing

Our initial phase of examination commenced with the retrieval of data. This process entailed downloading datasets from their individual sources. Subsequently, we proceeded to delve into the data through exploration, comprehending its characteristics and structural organization. This stage was pivotal as we aimed to analyze two datasets collectively. Subsequent to this, we scrutinized the datasets for any instances of missing values that necessitated attention.

2. Feature Extraction

Given that our analysis involves EEG scans, the extraction of particular features from our data holds significant importance. In this context, we identified time and frequency as pertinent features that will play a crucial role in our analytical process.

3. Data Splitting

The preprocessed data undergoes segmentation into validation, training, and test sets. This segmentation allows us to build and fine-tune the chosen model on one subset, validate its performance on another, and ultimately test its efficacy.

4. Model Selection

The selected model is LSTM (Long Short-Term Memory), a type of recurrent neural network designed to handle sequential data with long-term dependencies. This modeling approach proves highly effective in EEG scan analysis due to its capability to capture and retain information over extended periods, adapt to variable-length data, and proficiently learn and extract relevant features from the input data.

```python
## Model
import numpy as np
import torch

class LSTMModel(torch.nn.Module):
    def __init__(self):
        super(LSTMModel, self).__init__()
        self.lstm1 = torch.nn.LSTM(input_size=10000, hidden_size=100, num_layers=2, batch_first=True)
        self.relu = torch.nn.ReLU()
        self.dropout = torch.nn.Dropout(0.2)
        self.fc1 = torch.nn.Linear(100, 32)
        self.out = torch.nn.Linear(32, 2)
        self.softmax = torch.nn.Softmax(dim=1)

    def forward(self, x):
        h_t = torch.zeros(2, x.size(0), 100, dtype=torch.float32).to(x.device)
        c_t = torch.zeros(2, x.size(0), 100, dtype=torch.float32).to(x.device)
        x, _ = self.lstm1(x, (h_t, c_t))
        x = x[:, -1, :]
        x = self.relu(x)
        x = self.dropout(x)
        x = self.fc1(x)
        x = self.out(x)
        x = self.softmax(x)
        return x
```

Input size: The number of features in the input at each step. In this case, it corresponds to the number of time-domain samples extracted from the EEG signals.

Hidden size: The number of hidden units in the LSTM layer. It shows the memory capacity of the LSTM. A larger value allows for the network to capture more complex patterns but may also increase computational requirements.

Num layers: The number of LSTM layers stacked on top of each other. Stacking multiple layers helps the model learn hierarchical representations.

Batch first=True: This is just a parameter in which it indicates if the input data is batched with a size. The input and output tensors are typically batched with the batch dimension as the first dimension depending on the datasets.

5. Model Training

Once the architecture of our model was defined, we set the 'Loss Function' as cross-entropy, aiding in predicting the model's probability. Subsequently, we opted for the Adam optimizer algorithm to regulate the learning rate, determining the step size in the parameter space. The selection of this optimizer enables us to minimize the loss function effectively during the training process.

```
model = LSTMModel()

model.train()

epochs = 10

loss_fn = torch.nn.CrossEntropyLoss()
optimizer = torch.optim.Adam(model.parameters(), lr=3e-4)
for _ in tqdm(range(epochs), leave=True):
    running_loss = 0
    for batch in tqdm(train_dataloader, leave=True):
        x = batch['signal']
        y = batch['label']
        x = torch.tensor(x.reshape(-1, 1, 10000))
        y_hat = model(x)
        loss = loss_fn(y_hat, y)
        optimizer.zero_grad()
        loss.backward()
        optimizer.step()

        running_loss += loss.item()/len(train_dataloader)
    print(f"Loss: {running_loss}")
```

6. Model Evaluation

The objective of our model evaluation is to assess its performance. During this evaluation phase, 'dropout layers' are deactivated, ensuring that batch normalization layers utilize pre-computed statistics rather than updating during evaluation. The chosen metric for evaluation is 'accuracy,' indicating the proportion of correct predictions out of the total predictions made.

Results

```
model.eval()

preds = []
labels = []
for batch in test_dataloader:
    x = batch['signal']
    y = batch['label']
    x = torch.tensor(x.reshape(-1, 1, 10000))
    y_hat = model(x)
    pred = y_hat.argmax(dim=1)
    labels.extend(y.tolist())
    preds.extend(pred.tolist())

correct = [1 if p==l else 0 for p, l in zip(preds, labels)]


print(sum(correct)/ len(preds))
print(classification_report(labels, preds))
```

```
<ipython-input-7-7e29ccdc2a9e>:8: UserWarning: To copy construct
  x = torch.tensor(x.reshape(-1, 1, 10000))
0.7688172043010753
              precision    recall  f1-score   support

           0       0.79      0.96      0.87       147
           1       0.25      0.05      0.09        39

    accuracy                           0.77       186
   macro avg       0.52      0.51      0.48       186
weighted avg       0.68      0.77      0.70       186
```

Conclusion

In summary, the EEG analysis proved successful, involving a sophisticated process that encompassed tasks from pre-processing to model testing. This comprehensive approach was executed diligently, considering the challenge of working with multiple datasets in a domain unfamiliar to the team.

Each dataset presented a diverse range of subjects, features, and recordings, laying a foundation for success in selecting LSTM modeling. The inherent adaptability of LSTM effectively complements the diversity observed within the datasets.

The rationale behind opting for the LSTM model is to enhance the comprehension of optimizing the model's parameters, enabling it to learn patterns, dependencies, and representations from sequential data effectively. Additionally, the evaluation process allows us to assess the accuracy of the predictions made by the model.

This project strives to advance EEG-based diagnostics, merging insights from the CHB-MIT EEG Database and the Bonn EEG Dataset. Our classification model aims to enhance epilepsy diagnosis and holds promise for broader applications in neuroscience, marking a significant stride towards technologically-driven healthcare solutions."