UNIVERSIDAD DE MÁLAGA

Aprendizaje Computacional

Lab 2B:
Questions - Deep Learning

AUTHORS:

janeczko tomasz

lukáš hofman

November 10, 2024

1 Question 1

When developing a supervised method, why do we need to split the dataset into training and test sets?
In supervised machine learning, we split data into training and test set in order to assess generalization ability of a model: if the model can perform only on this particular training set or also on some new unseen data. The training set is for the model to learn and the test set which it never has seen is used for validation. It prevents the model from memorizing specific patterns rather than learning general trends, which makes them perform poorly in new data when it performs well on training data : overfitting. It's essential to understand how the model actually performs in the real world and so this validation step is critical.

2 Question 2

What is cross-validation?
Cross-validation is a process in machine learning used to measure how well a model generalizes to new, unseen data. It involves splitting the data into several subsets, or 'folds' in order to iteratively train on and test against the folds. Probably the most common form is k-fold cross-validation, where the dataset is divided into k parts. The model is trained on k-1 of these parts and tested on the remaining part. This is repeated k times so that every subset acts once as a test set.
This approach gives a more reliable evaluation than a single train-test split, reducing overfitting and enabling the understanding of model performance on different subsets of data. Cross-validation is primarily used during the model selection and tuning phase to help you discover the best hyperparameters or model configuration. It's not directly used in production or real-life deployment.

3 Question 3

What is artificial neural networks?
An artificial neural network (ANN or NN) is a computing system inspired by the neural networks of the human brain. These networks consist of interconnected layers of nodes, or 'neurons' which process information and learn to recognize patterns by being exposed to data. An ANN is structured as follows: there is one input layer, one or more hidden layers, and one output layer. Each neuron receives some inputs, processes them through weights and activation functions, and sends the result to the next layer.
ANNs are extremely flexible and have applications in image recognition, natural language processing, and predictive analytics. During training, they modify weights in order to decrease prediction errors learned from the data, resulting in a growth in accuracy the more data it consumes during the training of the model.

4 Question 4

What is Deep Learning (DL)?
Deep Learning (DL) is a subfield of machine learning that focuses on neural networks with many layers, known as deep neural networks. These networks are capable of learning complex patterns and representations in large, unstructured datasets. Unlike traditional machine learning methods that often require manual feature extraction, deep learning models can automatically learn feature hierarchies directly from raw data.
At its core, a deep learning model is made up of layers of artificial neurons, where each layer processes and transforms data in different ways. One of the biggest advantages of Deep Neural Networks (DNNs) is their ability to use specialized layers—such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer layers—to efficiently handle different types of data and tasks.
DNNs are often used in difficult tasks such as image recognition, natural language processing, and speech recognition.

5 Question 5

What is new in DL models with respect to traditional feedforward neural networks?
DL models bring several advancements over traditional feedforward neural networks (FNNs), allowing them to tackle more complex, large-scale problems with higher accuracy and efficiency. Here are the main advantages:

1. *Deep Architectures with Many Layers:*
DL models are characterized by their depth, often consisting of dozens or even hundreds of layers, compared to the limited number of layers in traditional FNNs. This depth allows DL models to learn hierarchical, multi-level representations of data, capturing complex features in ways that shallow FNNs cannot.
2. *Specialized Layer Types:*
Unlike traditional FNNs, which only have fully connected layers, DL models can incorporate specialized layers tailored to different types of data. For example:
- Convolutional layers (CNNs) excel in image and spatial data by detecting patterns such as edges and textures.
- Recurrent layers (RNNs, LSTMs) are used for sequential data, like text or time series, as they can handle dependencies across time steps.
- Transformer layers are particularly powerful for natural language processing, as they capture contextual relationships in sequences via attention mechanisms.
These specialized layers enable DL models to be far more effective across a range of complex tasks.
3. *Automatic Feature Learning:*
DL models can automatically extract features from raw data, such as edges in images or word embeddings in text, without requiring manual feature engineering. This is a significant advantage over FNNs, which rely heavily on hand-crafted features, making DL models more versatile and easier to deploy in varied domains.
4. *Usage of advanced training techniques:*
Advanced methods as regularization methods (dropout, batch normalization), data augmentation and transfer learning enable DL models to train on large datasets more effectively, providing robust, generalizable results.
5. *Parallel Processing and GPU Acceleration:*
DL models are designed to leverage GPUs and other specialized hardware, making it possible to train extremely large networks on big datasets efficiently. Traditional FNNs, which often don't require as much computational power, don't benefit from this level of parallelization.
6. *Advanced Architectures for Specific Use Cases:*
DL introduced architectures like Generative Adversarial Networks (GANs) for data generation, and Variational Autoencoders (VAEs) for unsupervised feature learning and data compression. These architectures go beyond what traditional FNNs offer, enabling applications in data synthesis, style transfer, and anomaly detection.
7. *Self-Supervised and Unsupervised Learning Capabilities:*
Many DL models can learn from large, unlabeled datasets using self-supervised or unsupervised learning methods, expanding their applicability beyond labeled data requirements. Traditional FNNs typically depend more heavily on labeled data, limiting their utility in real-world, large-scale applications.

6 Question 6

What is overfitting and how DL models avoid it?
Overfitting occurs when a model learns too much about the details and noise in the training data, negatively impacting performance on new, unseen data. The model becomes too complex and not only learns the underlying patterns but also the random fluctuations or noise within the training dataset. As a result, while the model looks great on training data, its predictive power on new data is poor. Overfitting can be dealt with in several ways:
Early Stopping: This technique consists of closing a learning phase when the performance of the model on the validation dataset no longer improves. A really nice way of stopping overfitting is to monitor the validation error during training and stop training before the model starts memorizing the training data.
Regularization: A regularization method used in L1 or L2 is to penalize the model with big weights, which reduces complexity and therefore reduces overfitting. L1 Regularization leads to sparse models (where many weights are zero) while L2 reduces the weights but leaves them small in absolute value.
Dropout: Dropout randomly deactivates a fraction of neurons during training, forcing the network to learn strong features so that no one neuron can dominate the decision. This technique very effectively reduces overfitting because it also creates somewhat differently looking "sub-models," each of which learns a different representation.
Data augmentation: By artificially increasing the size of the training data using rotation, scaling, or flipping of images, the model mitigates its memorization of respective training data and generalizes better.
These strategies form a powerful toolkit at the disposal of citizens in their endeavor to shore up the generalization capability of deep learning models when it comes to previously unseen data.

7 Problem 1

Google (among others) has produced astonishing results in the application of DL models in different domains. Mention two of these cases describing shortly the problem solved.
Over the years, companies like Google have created an impact on many complicated applications of DL models. Two such examples are:
AlphaFold for Protein Folding: AlphaFold is a deep learning model developed by Google's DeepMind that has solved the long-standing problems of predicting the 3D structures of proteins. This breakthrough will have a huge twofold impact: on our understanding of diseases and drug discovery. As such, predicting protein structures can catalyze scientific research in medicine and biotechnology.
Route Optimization with Google Maps: Google also employed deep learning methods to better its mapping services. By using inverse reinforcement learning, Google improved its route suggestion system, enabling an increase in global route match rates by 16-24These particular instances show how deep learning will change a different variety of fields: healthcare is one of them, navigational optimization is another.

8 Problem 2

Lack of data is a big limitation regarding the application of DL models to biomedical problems. What techniques can be applied to alleviate this problem.
Deep learning (DL) applications, in conjunction with biomedical problems, face a challenge of data scarcity. However, several techniques may mitigate this problem:
Data Collection: First and foremost, obtaining more data is the most direct way to alleviate data scarcity. In the biomedical field, this might involve collecting new samples, collaborating across institutions to access larger datasets, or expanding data collection efforts to include diverse populations. Although challenging and resource-intensive, a larger dataset provides a more comprehensive foundation for training deep learning models, leading to better generalization and improved robustness in biomedical applications.
Data Augmentation: Among the most widely used methods of limited data issue remediation is augmentation. This can refer to the generation of artificial data by creating variations from an existing dataset using a range of transformations. An example is medical image classification, in which applications involve rotation, flipping, and scaling of images to imitate variations that would reflect different conditions and angles. This process helps models generalize better by introducing higher variability without having to collect more data.
Generative Models: Techniques such as generative adversarial networks (GANs) can be employed in the generation of synthetic data. For instance, GANs have been used to synthesize realistic biomedical data, including images describing medical conditions or biological signals, resembling actual data distributions. The generated synthetic data can, in turn, be used to augment the training sets, enhancing model performance in data-scarce scenarios.
Transfer Learning: The pre-trained models, mostly trained on large, general datasets, may be fine-tuned on smaller biomedical datasets. In this way, the model learns generalizable features from a wider dataset base, which are then transferred onto a specific biomedical task. Transfer learning can significantly reduce the dependence on large labeled datasets in specialized domains.
Semi-supervised Learning: In cases where labeled data is limited, semi-supervised learning can be used to take advantage of the abundance of unlabeled data. These models learn using both labeled and unlabeled data to improve their performance given the scarcity of labeled data.
These techniques can work in conjunction to overcome the limitations posed by data scarcity and enable the use of deep learning in the biomedical field.

9 Used sources

- https://machinelearningmastery.com/train-test-split-for-evaluating-machine-learning-algorithms/
- https://www.mygreatlearning.com/blog/cross-validation/
- https://www.ibm.com/cloud/learn/neural-networks
- https://www.expert.ai/blog/deep-learning/
- https://www.freecodecamp.org/news/handling-overfitting-in-deep-learning-models/
- https://deepmind.google/discover/blog/2023-a-year-of-groundbreaking-advances-in-ai-and-computing/
- https://en.wikipedia.org/
- https://pytorch.org/docs/stable/generated/torch.optim.AdamW.html
- https://machinelearningmastery.com/cross-entropy-for-machine-learning/