**Convolutional Neural Networks and Apache Spark vs Hadoop**

* Overview (20)
  + Objective Statement and Research Question
* Lit Review (15)
* Critical Evaluation (30)
  + Key findings
    - Implications and limitations
    - Contradicting viewpoints and research gaps
* Conclusions (15)

Big Data Storage

* Critically assess data storage and management requirements from a modern perspective and evaluate limitations for legacy approaches
* Assess design concepts and architectural patterns of distributed big data systems and analyse components that form their technology stack
* Critically evaluate and select big data environment for retrieving…select appropriate algorithms for the required scale and speed

Advanced Data

* Debate the theory and application of neural networks
* Analyse a set of requirements to determine the type of neural network for a problemset. Document and justify choices made to stakeholders and peers

Intro

Lit Review

Methodology

Results

Discussion

Conclusion

**CNN’s for Image Classification**

Image classification is a hot topic within data science that has been researched thoroughly in artificial intelligence.

Kayed *et al.* (2020) demonstrated image classification with machine learning and deep learning models, where the SVC reached a test accuracy of 89.70% and the deep learning CNN models reached 98.80%. (Kayed, Anter and Mohamed, 2020)

Image classification is an important tool within many industries such as medical and fashion. Classification within the fashion industry is challenging due to a multitude of clothes categories and depth of clothes properties, which can result in labels having similar features making it difficult to distinguish accurately.

* Clothes can be distorted by lengths
* Some garments have a small size
* Images can have different angles, lights and noise which changes peresepctive
* Garment can have similar features tights vs trousers, heel boot vs heels etc
* Images can be the individual item or item on a model, or item on a hanger

(Kayed, Anter and Mohamed, 2020)

The CNN architectures that have been used in deep learning include LeNet, Alex Net, Google Net, VGG Net and Res Net. (Kayed, Anter and Mohamed, 2020)

*CNN’s are a class of deep feed-forward artificial neural networks which is used mainly for image processing, classification, segmentation and others.* There are three types of layers, convolutional, pooling, and fully-connected. Three main concepts of CNNs are the local receptive areas, weights sharing and down sampling process. Confusion matrix was used to compare to other modesl. Other metrics include the kappa coefficient, informedness, mean square error, sensitivity, specificity and confusion metrics.(Kayed, Anter and Mohamed, 2020)

CNNs have demonstrated exceptional performance relating to domains of computer vision and image processing. *Some of the exciting application areas of CNN include Image Classification and Segmentation, Object Detection, Video Processing, Natural Language Processing, and Speech Recognition.* CNN has multiple feature extraction stages. Improvements in CNNs are continuously studied applying innovative techniques such as *exploiting spatial and channel information, depth and width of the architecture and multi-path information processing*. (Khan *et al.*, 2020)

CNNs have become a popular deep learning method due to its ability to *exploit spatial or temporal correlation* in data. It has three main layers, convolutional layers, non-linear processing units and subsampling layers. CNN has gained popularity due to the feature extraction ability which resembles that of the neocortex in a human brain. CNN can learn through backpropagation algorithm. Innovations within CNNs are mainly around the restricting of processing units and designing new blocks. Layers 🡪 convolutional layer contains convolutional kernals where the neurons act as a kernal. It becomes a correlation operation when a kernal is symmetric. Receptive fields are small slices of the image by the convolutional kernal. Pooling layer extract the feature motifs form the convolution layer and sums up information to determine a reasons. Most important that the position relative to other features is maintained rather than the exact location. Pooling formulations can be utilised. Reducing the feature map size can help reduce overfitting within the model. Activation function learns patterns and provides a decision. Changes to the function can help accelerate the learning. Batch Normalization this is to support internal covariance shift in the feature maps. It unifies and improves generalization. Dropout is provided to improve generalization as a regularization function. It reduces overfitting, by randomly skipping over some connections. Fully connected layer is at the end and extracts info from feature extraction and analyses the outputs of the layers.(Khan *et al.*, 2020)

Two concerns with deep learning architectures are the high computational cost and memory requirement which is limitations of deep CNN models mainly due to the high multitude of multiplication that occur at the convolution operation level. Due to this, an important step is the selection of a suitable big data storage and processing solution. (Khan *et al.*, 2020)

Nowadays, the quantity of data being generated and consumed is continuously increasing which demands technologies that can store more data and also process data faster. One of the most common ways that data has been stored is with relational databases. In relational databases, the data is structured based on a pre-defined type and format and can be accessed easily and quickly. While these databases have benefits such as predictability, easy user interaction, and easy retrieval, they have high set-up costs, difficult to scale and difficult to incorporate unstructured data. Examples of relational databases utilised today are MySQL and Oracle.(Jatana *et al.*, 2012; Ergüzen and ünver, 2018)

In order to overcome the challenges of relational databases, non-relational systems were established. Non-relational systems don’t utilise SQL and are often referred to as NoSQL databases. These were designed to offer high performance, availability, and scalability. However, these benefits are at the cost of losing ACID (atomic, consistent, isolated, durable) attributes that come with the relational databases. Instead, they contain the BASE (basic availability, soft state, eventual consistency) attributes. (Tudorica and Bucur, 2011)

Distributed file systems (DFSs) are a type of non-relational systems that is distributed across multiple servers or locations. The main attributes of a DFSs are transparency, fault tolerance and scalability. (Depardon *et al.*, 2013) A commonly used solution for big data storage is Hadoop, an open-source software framework that is comprised of two layers, the storage layer which is called Hadoop Distributed File System (HDFS) and second layer which is processing layer called MapReduce. Hadoop is scalable, fault tolerant, cost-effective and support unstructured data. The main limitation is the slow processing speed. (Dwivedi and Dubey, 2014; Ghazi and Gangodkar, 2015)

HDFS is targeted towards batch processing. The architecture behind HDFS is a NameNode, where the metadata is managed and DataNode, where the data is stored. The NameNode is broken down into blocks of data that get distributed to multiple DataNodes, and often blocks are replicated across nodes as system backup. MapReduce is a programming model which writes to applications and is capable of parallel processing. There are two phases, the map phase whose input arises from the HDFS and the reduce phase whose input is the map phase output. In addition, MapReduce utilised JobTracker daemon and TaskTracker daemon similar to the master/slave architecture. (Dwivedi and Dubey, 2014; Ghazi and Gangodkar, 2015)

Apache spark is a data processing platform that implements a hybrid framework that can support batch and real-time processing. Apache Sparks architecture is made up of driver program, cluster manage and slave nodes. The driver program is the master node and entry point for the application. The cluster manager is responsible for resourcing, splitting the jobs into the slave node clusters which then execute. Resilient Distributed Datasets (RDDs) are a core structure to Apache Spark which supports in-memory processing providing a fault tolerant framework. As a result of using RDDs, Apache Spark is a hundred times faster than Hadoop. Additional benefits of Apache Spark include use with multiple programming language, and real time processing. (Shaikh *et al.*, 2019)

**Introduction**

Machine learning (ML) is a form of artificial intelligence that utilises algorithms and statistical models to train a computer. Deep learning is a subset of ML which utilises neural networks to explore patterns and derive relationships within complex datasets. It closely resembles how the human brain learns and is utilised in tools we use everyday such as speech recognition, natural language processing. The neural network algorithms are based on the human brain and how it learns. to machine learning algorithms but contain many more layers which allow for the analysis of complex datasets.

Image classification is one of the many challenges that deep learning tools have been employed for particularly within medical imaging to identify diseases. Another industry is retail, where neural networks are utilised to categorise products in online stores. For example, ASOS a large online retailer could have almost 1000 new products per day which are then required to be sorted into over 100 different categories across product type, style, colour.

The main types of neural networks are Artificial Neural Networks (ANN), Convolution Neural Networks (CNN), Recurrent Neural Networks (RNN), Perceptron, Long Short Term Memory Networks and Radial Basis Functional Neural Networks.

Convolutional Neural Networks, a feed forward neural network, is the main method used for image classification. CNNs mimic the way the human eyes and brain works, by taking small subsections of an image and analysing them. There are three main layers that occur with a CNN, the convolution layer, the pooling layer and the fully connected layer. The convolution layer essentially extracts the features of the image. A filter size and stride size are defined and used to divide up the image into a number or pixels. Next is the pooling layer which takes the output of the convolution layer and reduces the dimensionality again where the average or max values from the results are taken. Then the fully connected layer, where the neurons in each layer are taken and connected to the subsequent layers. This layer then learns how to classify particular objects (Lang, 2021) (Li *et al.*, 2022) (Taye, 2023)

An activation function forms part of the neural network that determines what information is communicated to the next neuron. Activation functions common within neural networks are linear, sigmoid, tanh, ReLu and swish. Li *et al.*, 2022 evaluated the activation functions on the fashion MNIST data and concluded that the linear activation performs the worst as expected with a multi-layer model and that ReLU performed the best overall with relation to accuracy, training time and stability. ReLU is the most commonly used activation function as due to time and resource savings, simpler gradient definitions and sparser representation. (Li *et al.*, 2022) (Taye, 2023)

Loss functions are a measurement that calculates the distance between the predicted and actual value. Models are optimized to reduce the value of the loss function. For classification, cross entropy is the most common used which takes the predicted probability and the output value and the distance between them is used to calculate the penalty value. In CNNs, it is incorporated in a softmax layer. Other loss functions that have been introduced to account for the cross entrophy disadvantages include contrastive loss, triplet loss, center loss and the large margin softmax.(Li *et al.*, 2022)

Optimizers are functions that are used to reduce loss functions in neural networks. Gradient descent optimizers work to converge the model to a set of parameters that minimizes the loss on training. Examples of optimizers include stochastic gradient descent, adam, and RMSprop. Each optimizer performs to its own strength and weakens and should be decided based on data distribution, computing cost and accuracy. (Li *et al.*, 2022)

Hyperparameter tuning is a crucial part of any models development and is no different for CNNs in order to achieve the best performance. For CNNs the main hyperparameters to be tuned include, learning rate, epoch, min-batch size, number of layers and kernels and size of kernels. (Li *et al.*, 2022)

**Literature Review**

Kadam *et al.* (2020), proposed five different CNN architectures for image classification. The architectures were varied through activation methods, dropout, learning rate, batch size and layers. A testing accuracy of 99.55% was obtained for the MNIST dataset and 93.56% for the Fashion MINST dataset.(Kadam, Adamuthe and Patil, 2020)

Kayed *et al.* (2020) proposed a CNN model with a LeNet-5 architecture that obtained an accuracy over 98%. The architecture contains five layers which are a combination of convolutional layers with 5x5 filters and pooling layers with 2x2 and a stride of 2. An additional fully connected layer was v

Kayed *et al.* (2020) demonstrated image classification with machine learning and deep learning models, where the SVC reached a test accuracy of 89.70% and the deep learning CNN models reached 98.80%. (Kayed, Anter and Mohamed, 2020)

Sharma *et al.* (2018) assessed the performance of the most popular CNN models, Alex Nets, GoogLeNet and ResNet50 across various image data sets for object detection in real world scenes. The objective was to assess the accuracy and prediction consistency of each CNN. It was concludes that the higher number of layers were favourable. (Sharma, Jain and Mishra, 2018)

Nocentini *et al.* (2022) proposed four different CNN models for image classification using the Fashion MNIST dataset. The models were varied and tuned with respect to batch size, kernal size, number of filters and fully connected layers. They obtained an accuracy of 94.09% with their MCNN15 model. (Nocentini *et al.*, 2022)

AlexNet [11], the best model in ILSVRC-2012, uses ReLU as the activation function of the CNN-based model, which mitigates gradient vanishing problem when the network is deep and verifie s that the use of ReLU surpasses sigmoid in deep networks

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Description | Test Accuracy (%) | Test Loss (%) | Training Time (secs) |
| CNN\_Model\_One | Basic CNN model with Three convolutional and pooling layers and two connect layers | 91.24 | 28.34 | 606.90 |
| CNN\_Model\_One-Epoch Increase | CNN\_Model\_One but epoch increased from 10 to 30 | 90.94 | 83.57 | 1752.41 |
| CNN\_Model\_Two | CNN\_Model\_One with 5 extra fully connected layer | 91.04 | 32.64 | 607.89 |
| CNN\_Model\_Three | CNN\_Model\_Two with dropout of 0.3 added after each layer | 89.94 | 29.54 | 656.97 |
| CNN\_Model\_Four | CNN\_Model\_Two with batch normalization added and dropout | 91.49 | 32.50 | 973.83 |