SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF ELECTRICAL AND ELECTRONICS

DEPARTMENT OF ELECTRONICS AND COMMUNICATIONS

UNIT – V – DEEP LEARNING – SECA4002

UNIT V APPLICATIONS OF DEEP LEARNING

Imagenet- Detection-Audio WaveNet-Natural Language Processing Word2Vec - Joint Detection BioInformatics- Face Recognition- Scene Understanding- Gathering Image Captions

5.1. Imagenet:

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. In Machine Learning and Deep Neural Networks, machines are trained on a large dataset of various images. Machines are required to learn useful features from these training images. Once learned, they can use these features to classify images and perform many other tasks associated with computer vision. ImageNet gives researchers a common set of images to benchmark their models and algorithms.

ImageNet is useful for many computer vision applications such as object recognition, image classification and object localization.Prior to ImageNet, a researcher wrote one algorithm to identify dogs, another to identify cats, and so on. After training with ImageNet, the same algorithm could be used to identify different objects. The diversity and size of ImageNet meant that a computer looked at and learned from many variations of the same object. These variations could include camera angles, lighting conditions, and so on. Models built from such extensive training were better at many computers vision tasks. ImageNet convinced researchers those large datasets were important for algorithms and models to work well.

5.1.1. Technical details of Image Net:

ImageNet consists of 14,197,122 images organized into 21,841 subcategories. These subcategories can be considered as sub-trees of 27 high-level categories. Thus, ImageNet is a well-organized hierarchy that makes it useful for supervised machine learning tasks. As many as 1,034,908 images have been annotated with **bounding boxes**. For example, if an image contains a cat as its main subject, the coordinates of a rectangle that bounds the cat are also published on ImageNet. This makes it useful for computer vision tasks such as object localization and detection. Then there's Scale-Invariant Feature Transform (SIFT) used in computer vision. SIFT helps in detecting local features in an image. ImageNet gives researchers 1000 subcategories with SIFT features covering about 1.2 million images. Images vary in resolution but it's common practice to train deep learning models on sub-sampled images of 256x256 pixels.

ImageNet did not define these subcategories on its own but derived these from WordNet. **WordNet** is a database of English words linked together by semantic relationships. Words of similar meaning are grouped together into a synonym set, simply called **synset**. Hypernyms are synsets that are more general. Thus, "organism" is a hypernym of "plant". Hyponyms are synsets that are more specific. Thus, "aquatic" is a hyponym of "plant". This hierarchy makes it useful for computer vision tasks. If the model is not sure about a subcategory, it can simply classify the image higher up the hierarchy where the error probability is less. For example, if model is unsure that it's looking at a rabbit, it can simply classify it as a mammal.

While WordNet has 100K+ synsets, only the nouns have been considered by ImageNet.

5.1.2. How the images are labelled in ImageNet?

In the early stages of the ImageNet project, a quick calculation showed that by employing a few people, they would need 19 years to label the images collected for ImageNet. But in the summer of 2008, researchers came to know about an Amazon service called Mechanical Turk. This meant that image labelling can be crowdsourced via this service. Humans all over the world would label the images for a small fee.

Humans make mistakes and therefore we must have checks in place to overcome them. Each human is given a task of 100 images. In each task, 6 "gold standard" images are placed with known labels. At most 2 errors are allowed on these standard images, otherwise the task has to be restarted.

In addition, the same image is labelled by three different humans. When there's disagreement, such ambiguous images are resubmitted to another human with tighter quality threshold (only one allowed error on the standard images).

5.1.3. How the images of ImageNet Licensed?

Images for ImageNet were collected from various online sources. ImageNet doesn't own the copyright for any of the images. This has implication on how ImageNet shares the images to researchers.

For public access, ImageNet provides image thumbnails and URLs from where the original images were downloaded. Researchers can use these URLs to download the original images. However, those who wish to use the images for non-commercial or educational purpose, can create an account on ImageNet and request access. This will allow direct download of images from ImageNet. This is useful when the original sources of images are no longer available.

The dataset can be explored via a browser-based user interface. Alternatively, there's also an API. Researchers may want to read the API Documentation. This documentation also shares how to download image features and bounding boxes.

5.1.4. Shortcomings of ImageNet:

Images are not uniformly distributed across subcategories. One research team found that by considering 200 subcategories, they found that the top 11 had 50% of the images, followed by a long tail.

When classifying people, ImageNet uses labels that are racist, misogynist and offensive. People are treated as objects. Their photos have been used without their knowledge. About 5.8% labels are wrong. ImageNet lacks geodiversity. Most of the data represents North America and Europe. China and India are represented in only 1% and 2.1% of the images respectively. This implies that models trained on ImageNet will not work well when applied for the developing world.

Another study from 2016 found that 30% of ImageNet's image URLs are broken. This is about 4.4 million annotations lost. Copyright laws prevent caching and redistribution of these images by ImageNet itself

Content Source: https://devopedia.org/imagenet

5.2. WaveNet:

WaveNet is a deep generative model of raw audio waveforms. We show that WaveNets are able to generate speech which mimics any human voice and which sounds more natural than the best existing Text-to-Speech systems, reducing the gap with human performance by over 50%. Allowing people to converse with machines is a long-standing dream of human-computer interaction. The ability of computers to understand natural speech has been revolutionised in the last few years by the application of deep neural networks. However, generating speech with computers — a process usually referred to as speech synthesis or text-to-speech (TTS) — is still largely based on so-called concatenative TTS, where a very large database of short speech fragments are recorded from a single speaker and then recombined to form complete utterances. This makes it difficult to modify the voice (for example switching to a different speaker, or altering the emphasis or emotion of their speech) without recording a whole new database.

This has led to a great demand for parametric TTS, where all the information required to generate the data is stored in the parameters of the model, and the contents and characteristics of the speech can be controlled via the inputs to the model. So far, however, parametric TTS has tended to sound less natural than concatenative. Existing parametric models typically generate audio signals by passing their outputs through signal processing algorithms known as vocoders. WaveNet changes this paradigm by directly modelling the raw waveform of the audio signal, one sample at a time. As well as yielding more natural-sounding speech, using raw waveforms means that WaveNet can model any kind of audio, including music.

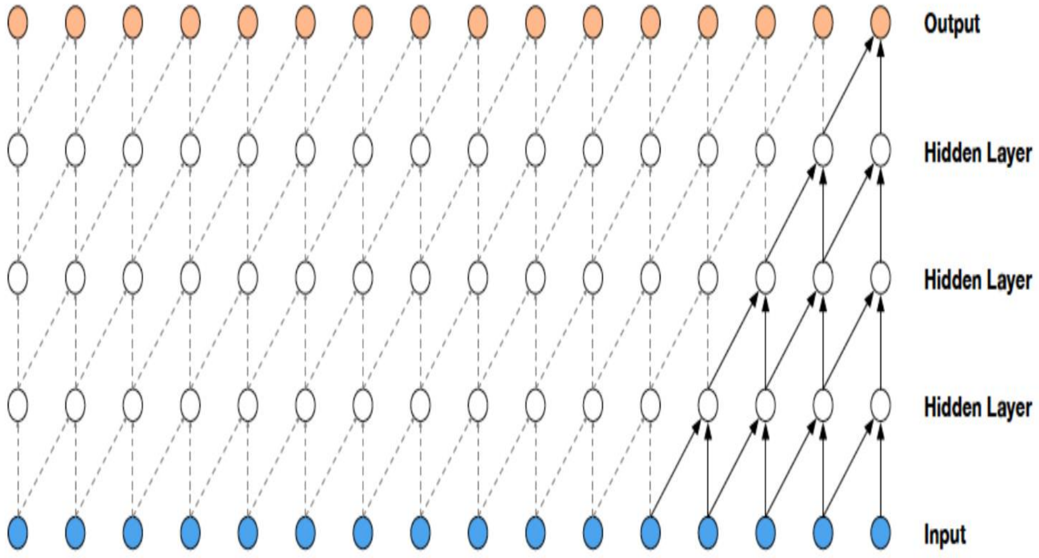

Figure 5.1: WaveNet Structure

The WaveNet proposes an autoregressive learning with the help of convolutional networks with some tricks. Basically, we have a convolution window sliding on the audio data, and at each step try to predict the next sample value that it did not see yet. In other words, it builds a network that learns the causal relationships between consecutive timesteps (as shown in figure 5.1)

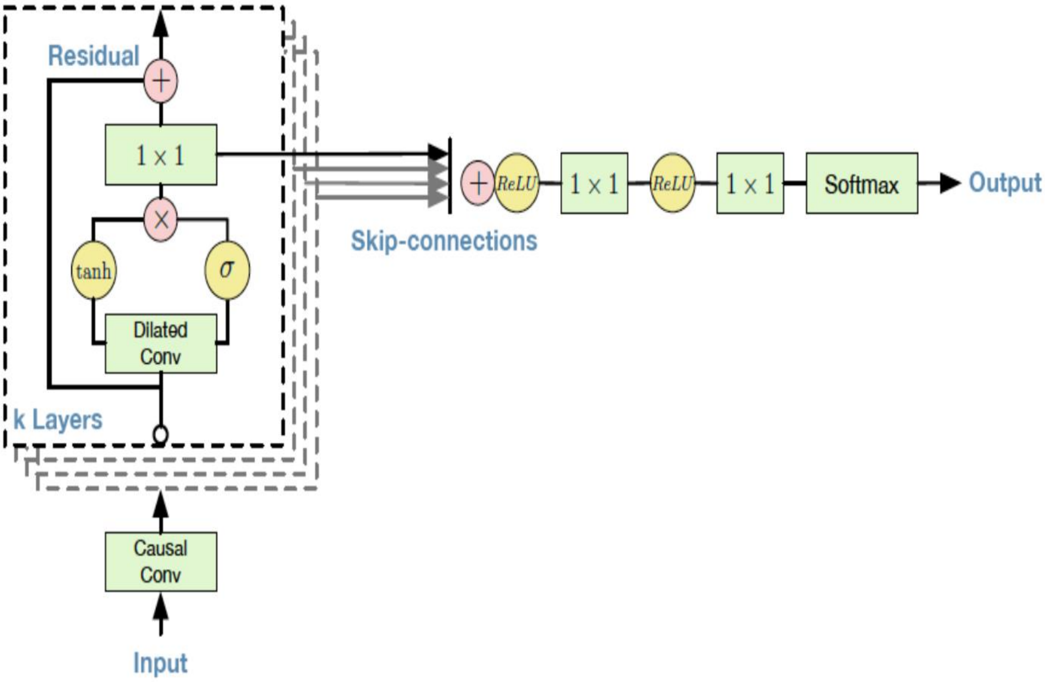

Figure 5.2: WaveNet Overall Model

Typically, the speech audio has a sampling rate of 22K or 16K. For few seconds of speech, it means there are more than 100K values for a single data and it is enormous for the network to consume. Hence, we need to restrict the size, preferably to around 8K. At the end, the values are predicted in Q channels (eg. Q=256 or 65536), which is compared to the original audio data compressed to Q distinct values. For that, the **mulaw quantization** could be used: it maps the values to the range of [0,Q]. And the loss can be computed either by cross-entropy, or discretized logistic mixture.

5.2.1. The Workflow of WaveNet:

- Input is fed into a causal 1D convolution
- The output is then fed to 2 different dilated 1D convolution layers with sigmoid and tanh activations
- The element-wise multiplication of 2 different activation values results in a skip connection

- And the element-wise addition of a skip connection and output of causal 1D results in the residual

Content Source: (1) https://www.analyticsvidhya.com/blog/2020/01/how-to-perform-automatic-music-generation/
(2)https://medium.com/@evinpinar/wavenet-implementation-and-experiments-2d2ee57105d5
(3)https://deepmind.com/blog/article/wavenet-generative-model-raw-audio

5.3. Natural Language Processing [NLP]:

Language is a method of communication with the help of which we can speak, read and write. For example, we think, we make decisions, plans and more in natural language; precisely, in words. However, the big question that confronts us in this AI era is that can we communicate in a similar manner with computers. In this sense, we can say that Natural Language Processing (NLP) is the sub-field of Computer Science especially Artificial Intelligence (AI) that is concerned about enabling computers to understand and process human language. Technically, the main task of NLP would be to program computers for analysing and processing huge amount of natural language data.

5.3.1. Natural Language Processing Phases:

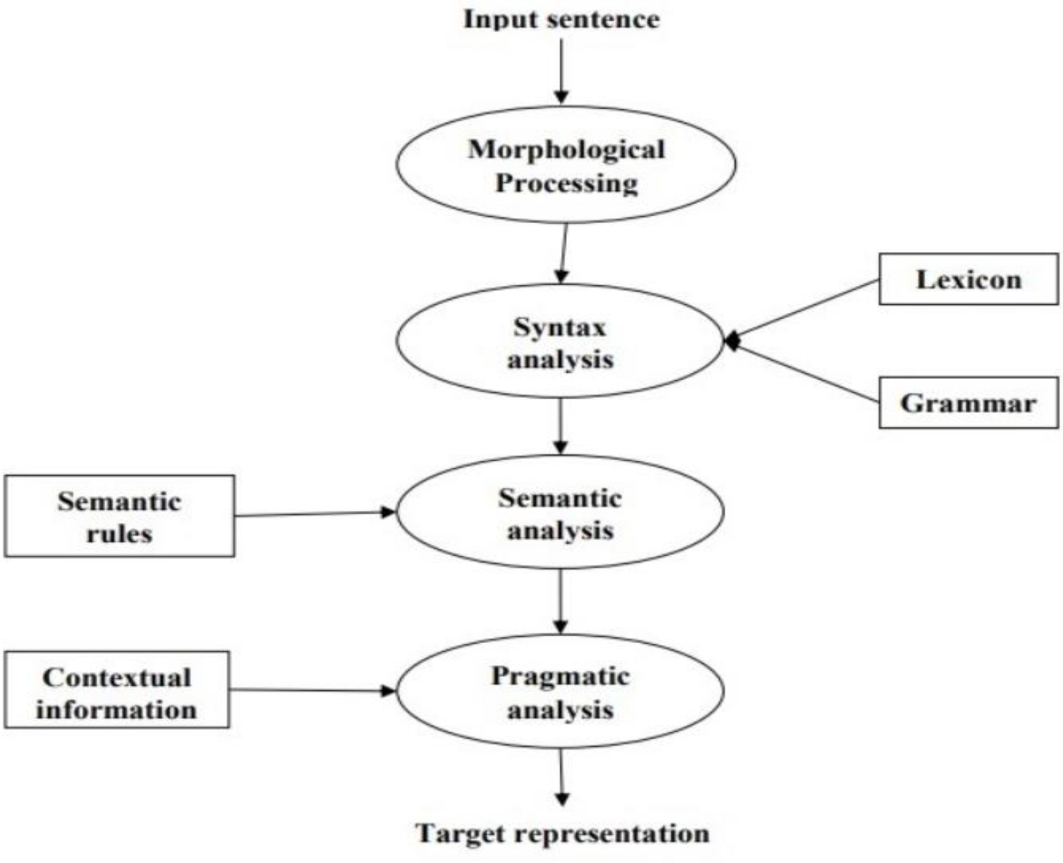

Figure 5.3: NLP Phases

The above diagram (Figure5.3) shows the phases or logical steps involved in natural language processing

5.3.1.1 Morphological Processing

It is the first phase of NLP. The purpose of this phase is to break chunks of language input into sets of tokens corresponding to paragraphs, sentences and words. For example, a word like **“uneasy”** can be broken into two sub-word tokens as **“un-easy”**.

5.3.1.2 Syntax Analysis

It is the second phase of NLP. The purpose of this phase is two folds: to check that a sentence is well formed or not and to break it up into a structure that shows the syntactic relationships between the different words. For example, the sentence like **“The school goes to the boy”** would be rejected by syntax analyser or parser.

5.3.1.3 Semantic Analysis

It is the third phase of NLP. The purpose of this phase is to draw exact meaning, or you can say dictionary meaning from the text. The text is checked for meaningfulness. For example, semantic analyser would reject a sentence like “Hot ice-cream”.

5.3.1.4 Pragmatic Analysis

It is the fourth phase of NLP. Pragmatic analysis simply fits the actual objects/events, which exist in a given context with object references obtained during the last phase (semantic analysis). For example, the sentence “Put the banana in the basket on the shelf” can have two semantic interpretations and pragmatic analyser will choose between these two possibilities.

Content Source: Https:// www.tutorialspoint.com/ natural_language_processing/ natural_ language_processing_quick_guide.htm

5.3.1.5. Different types based on Working:

1. Speech Recognition — The translation of spoken language into text.
2. Natural Language Understanding (NLU) — The computer’s ability to understand what we say.
3. Natural Language Generation (NLG) — The generation of natural language by a computer.

5.3.1.6. Applications of NLP:

- ✓ Spam Filters
- ✓ Algorithmic Trading
- ✓ Answering Questions
- ✓ Summarizing Information’s etc

Content Source: https://www.geeksforgeeks.org/natural-language-processing-overview/

5.4. Word2Vec:

Word embedding is one of the most popular representation of document vocabulary. It is capable of capturing context of a word in a document, semantic and syntactic similarity, relation with other words, etc. What are word embeddings exactly? Loosely speaking, they are vector representations of a particular word. Having said this, what follows is how do we generate them? More importantly, how do they capture the context? Word2Vec is one of the most popular technique to learn word embeddings using shallow neural network. It was developed by Tomas Mikolov in 2013 at Google.

The purpose and usefulness of Word2vec is to group the vectors of similar words together in vector space. That is, it detects similarities mathematically. Word2vec creates vectors that are distributed numerical representations of word features, features such as the context of individual words. It does so without human intervention.

Given enough data, usage and contexts, Word2vec can make highly accurate guesses about a word's meaning based on past appearances. Those guesses can be used to establish a word's association with other words (e.g. "man" is to "boy" what "woman" is to "girl"), or cluster documents and classify them by topic. Those clusters can form the basis of search, sentiment analysis and recommendations in such diverse fields as scientific research, legal discovery, e-commerce and customer relationship management. Measuring cosine similarity, no similarity is expressed as a 90 degree angle, while total similarity of 1 is a 0 degree angle, complete overlap.

Word2vec is a two-layer neural net that processes text by "vectorizing" words. Its input is a text corpus and its output is a set of vectors: feature vectors that represent words in that corpus. While Word2vec is not a deep neural network, it turns text into a numerical form that deep neural networks can understand.

Word2vec's applications extend beyond parsing sentences in the wild. It can be applied just as well to genes, code, likes, playlists, social media graphs and other verbal or symbolic series in which patterns may be discerned.

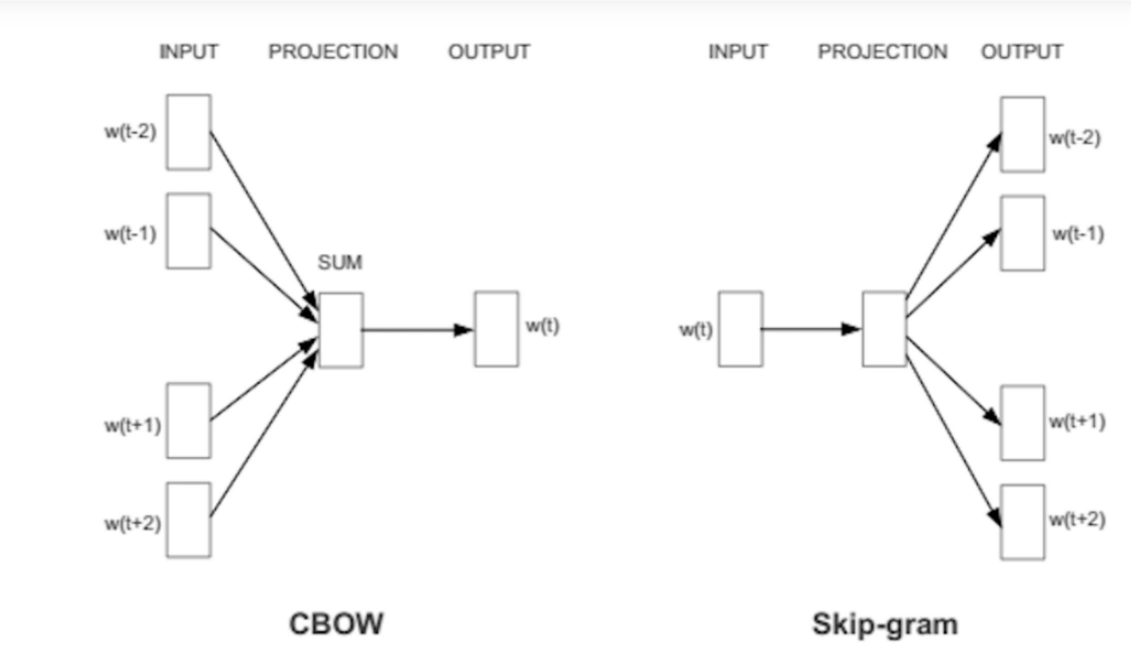

Figure 5.4: Two models of Word2Vec (A- CBOW & B- Skip-Gram model)

Word2vec is similar to an autoencoder, encoding each word in a vector, but rather than training against the input words through reconstruction, as a restricted Boltzmann machine does, word2vec trains words against other words that neighbour them in the input corpus. t does so in one of two ways, either using context to predict a target word (a method known as continuous bag of words, or CBOW), or using a word to predict a target context, which is called skip-gram.

When the feature vector assigned to a word cannot be used to accurately predict that word's context, the components of the vector are adjusted. Each word's context in the corpus is the *teacher* sending error signals back to adjust the feature vector. The vectors of words judged similar by their context are nudged closer together by adjusting the numbers in the vector.

Similar things and ideas are shown to be "close". Their relative meanings have been translated to measurable distances. Qualities become quantities, and algorithms can do their work. But similarity is just the basis of many associations that Word2vec can learn. For example, it can gauge relations between words of one language, and map them to another.

The main idea of word2Vec is to design a model whose parameters are the word vectors. Then, train the model on a certain objective. At every iteration we run our model, evaluate the errors, and follow an update rule that has some notion of penalizing the model parameters that caused the error. Thus, we learn our word vectors.

Content Source: (1) https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa
(2) https://wiki.pathmind.com/word2vec

5.5. Applications of Deep Learning Networks: Joint Detection

Deep Learning finds a lot of usefulness in the field of Biomedical and Bioinformatics. Deep Learning algorithms can be used for detecting fractures or anatomical changes in the human bones or in bone joints thereby early prediction of various diseases like arteritis can be done which helps in early curing of the so-called diseases also. Knee osteoarthritis (OA) is a very general joint disease that disturb many people especially people over 60. The severity of pain caused by knee OA is the most important portent to disable. Until now, the bad impact of osteoarthritis on health care and public health systems is still increasing.

Normal Neural Networks fails because of errors in the stages of Image Segmentation and Feature Extractions.TO avoid this we can build a Convolution based model as shown in the Figure 5.5 given below. In this example we had considered a CNN based Network. The input to this model is Knee Thermographs. Thermography is the image which senses or captures the heat intensity coming out from that particular region. Based on the patient's pressure points the color in the thermographs vary. Red regions denote more pressure locations and Yellow regions Denote less pressure locations. So, from the thermogram we can understand the effects of joint/Bone wear and tear or Damage occurred at particular spot.

The Convolution Filter is made to move over the image. The stride value here considered for this case study is 1. We have used Max Pooling and in the fully connected layer we have used Softmax aggregator.

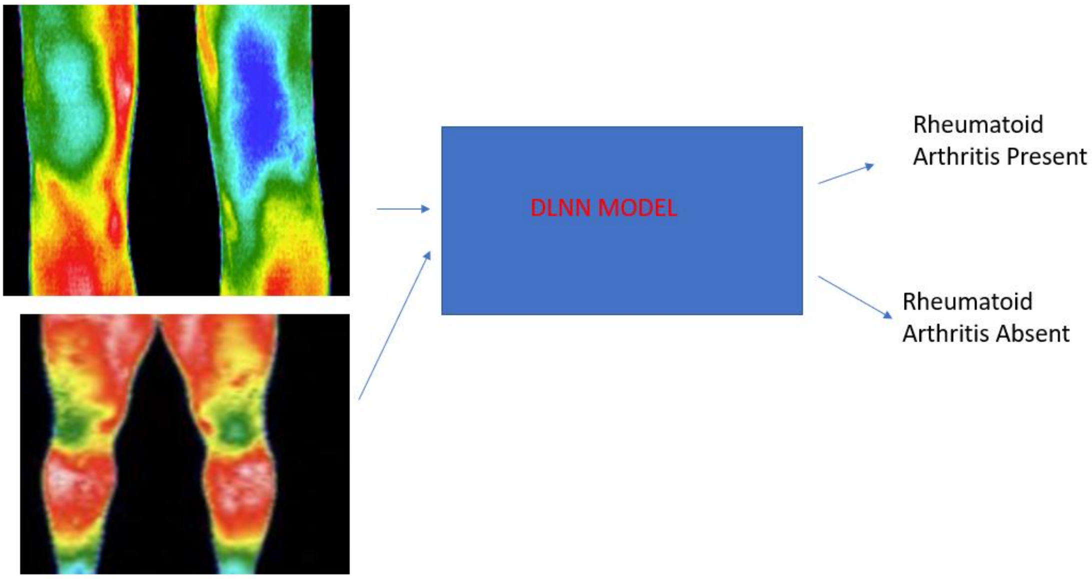

Figure 5.5: General representation of Bone Joint detection system

The above diagram is a schematic representation of a Deep Learning based network which can be used for human knee Joint deformities Identification purpose. Figure 5.6 shows the general Anatomical structure of a Human Knee Joint. Figure 5.6 shows the key structure skin texture of knee osteoarthritis. The left elevation of the image demonstrates the ordinary knee and the right elevation illustrates the contaminated joint. There have been many deep learning methods that can contribute well to the KOA diagnosis accurately as a part of early detection.

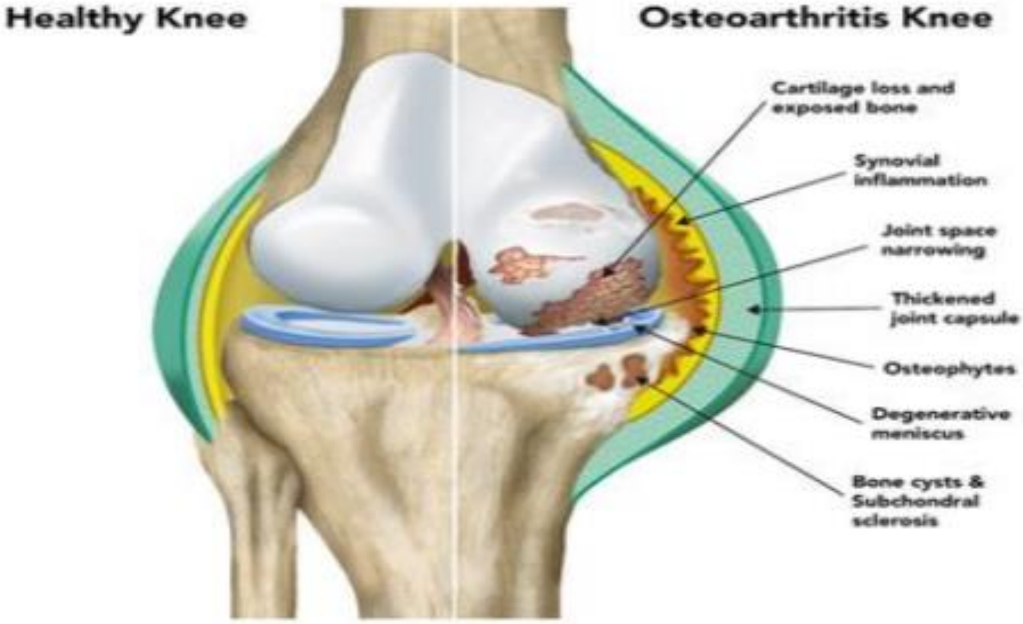

Figure 5.6: Human Knee Joint Structure

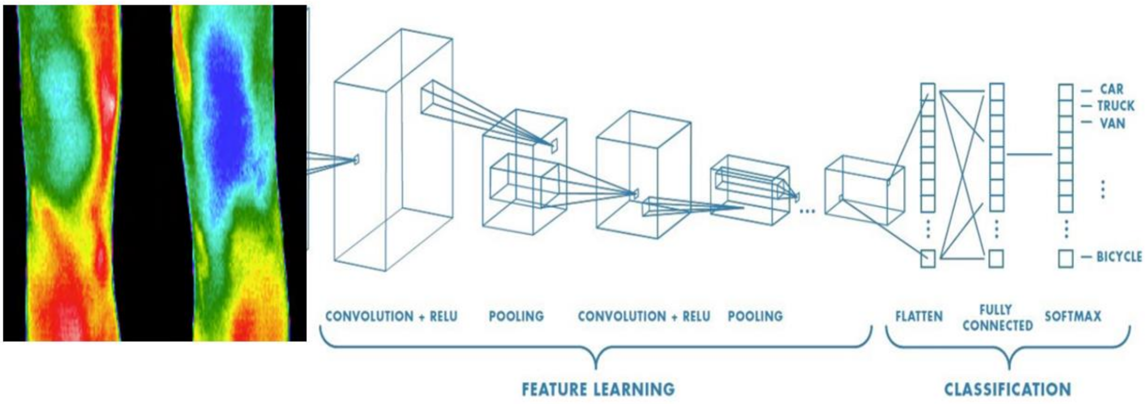

Figure 5.7: CNN based Knee Joint Detection Model

Figure 5.7 shows the full model of a joint detection procedure. The Convolution filter moves to the right with a certain Stride Value till it parses the complete width. Moving on, it hops down to the beginning (left) of the image with the same Stride Value and repeats the process until the entire image is traversed. The Kernel has the same depth as that of the input image. The objective of the Convolution Operation is to **extract the high-level features** such as edges, from the input image. Stride is the number of pixels shifts over the input matrix. When the stride is 1 then we move the filters to 1 pixel at a time. When the stride is 2 then we move the filters to 2 pixels at a time and so on

Pooling layers section would reduce the number of parameters when the images are too large. Spatial pooling also called subsampling or down sampling which reduces the dimensionality of each map but retains important information. This is to **decrease the computational power required to process the data** by reducing the dimensions

Types of Pooling:

- Max Pooling
- Average Pooling
- Sum Pooling

- The image is flattened into a column vector.
- The flattened output is fed to a feed-forward neural network and backpropagation applied to every iteration of training.

Over a series of epochs, the model is able to distinguish between dominating and certain low-level features in images and classify them using the Softmax Classification technique. The feature map matrix will be converted as vector (x1, x2, x3, …). These features are combined together to create a model.

Finally, an activation function such as softmax or sigmoid is used to classify the outputs as Normal and Abnormal.

5.5.1 Steps Involved:

- Provide input image into convolution layer
- Choose parameters, apply filters with strides, padding if requires. Perform convolution on the image and apply ReLU activation to the matrix.
- Perform pooling to reduce dimensionality size
- Add as many convolutional layers until satisfied
- Flatten the output and feed into a fully connected layer (FC Layer)
- Output the class using an activation function (Logistic Regression with cost functions) and classifies images.

5.6. Other Applications:

Similarly for the other Applications such as Facial Recognition and Scene Matching applications appropriate Deep Learning Based Algorithms such as AlexNet, VGG, Inception, ResNet and or Deep learning-based LSTM or RNN can be used. These Networks has to be explained with necessary Diagrams and appropriate Explanations.

Reference Books:

1. B. Yegnanarayana, “Artificial Neural Networks” Prentice Hall Publications.
2. Simon Haykin, “Artificial Neural Networks”, Second Edition, Pearson Education.
3. Laurene Fausett, “Fundamentals of Neural Networks, Architectures, Algorithms and Applications”, Prentice Hall publications.
4. Cosma Rohilla Shalizi, Advanced Data Analysis from an Elementary Point of View, 2015.
5. 2. Deng & Yu, Deep Learning: Methods and Applications, Now Publishers, 2013.
6. 3. Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, MIT Press, 2016.
7. 4. Michael Nielsen, Neural Networks and Deep Learning, Determination Press, 2015.

Note: For further reference, kindly refer the class notes, PPTs, Video lectures available in the Learning Management System (Moodle)

**************************** ALL THE BEST **