# **When to use Deep Learning or not over others?**

1. **Deep Learning** out perform other techniques if the ***data size is large***. But with small data size, traditional **Machine Learning** algorithms are preferable.
2. **Deep Learning** techniques need to have ***high end infrastructure*** to train in reasonable time.
3. When there is ***lack of domain understanding for feature introspection***, **Deep Learning** techniques outshines others as you have to worry less about feature engineering.
4. **Deep Learning** really shines when it comes to ***complex problems such as image classification, natural language processing, and speech recognition***.

While basic machine learning models do become progressively better at whatever their function is, they still need some guidance. If an AI algorithm returns an inaccurate prediction, then an engineer has to step in and make adjustments.

**With a deep learning model, an algorithm can determine on its own if a prediction is accurate or not through its own neural network.**

Q1) why CNN ?

Ans:

* CNN are complex feed-forward networks.
* They are used for image classification and recognition for their high accuracy.
* The CNN follows a hierarchical model which works on building a network, like a funnel, and finally gives out a fully-connected layer where all the neurons are connected to each other and the output is processed.
* CNNs are [regularized](https://en.wikipedia.org/wiki/Regularization_(mathematics)) versions of [multilayer perceptrons](https://en.wikipedia.org/wiki/Multilayer_perceptron). Multilayer perceptrons usually mean fully connected networks, that is, each neuron in one layer is connected to all neurons in the next layer.
* CNN is now the go-to model on every image related problem. In terms of accuracy they blow competition out of the water. It is also successfully applied to recommender systems, natural language processing and more. The main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision. For example, given many pictures of cats and dogs it learns distinctive features for each class by itself.
* CNN is also computationally efficient. It uses special convolution and pooling operations and performs parameter sharing. This enables CNN models to run on any device, making them universally attractive.

Convolution

* The main building block of CNN is the convolutional layer. Convolution is a mathematical operation to merge two sets of information.

## Pooling

* After a convolution operation we usually perform pooling to reduce the dimensionality. This enables us to reduce the number of parameters, which both shortens the training time and combats overfitting. Pooling layers downsample each feature map independently, reducing the height and width, keeping the depth intact.

## Fully Connected

* After the convolution + pooling layers we add a couple of fully connected layers to wrap up the CNN architecture.

## Training

* CNN is trained the same way like ANN, backpropagation with gradient descent. Due to the convolution operation it’s more mathematically involved

***Traditionally how has lip Reading been done and enhancements***

Traditional lip-reading systems usually consist of two stages: feature extraction and classification.

For the first stage, most previous feature extraction methods use pixel values extracted from the mouth region of interest (ROI) as visual information. Then, the abstract image features are extracted by discrete cosine transform (DCT) [6,7], discrete wavelet transform (DWT) [7] and principal componentanalysis (PCA) [7,8]. Therefore, the model-based methods, such as active appearance model (AAM) [9] and active shape model (ASM) [10] form non-rigid models and obtain a set of advanced geometric features which has the characteristics of lower dimensionality and stronger robustness.

In the second stage, the extracted features are fed into the classifiers of support vector machine (SVM) [11] and hidden Markov model (HMM) [12].

In recent years, researchers have introduced attention mechanisms on convolutional neural networks (CNN) to focus on areas of interest, and the classification and target detection of images have also achieved great success. For example, a CNN feature extraction method based on attention mechanism proposed by Vinyals et al. [13]. Furthermore, the mechanism of attention can be successfully applied in recurrent neural network (RNN) to find the relationship between the context. Since the changes between video frames of automatic lip-reading are continuous and happen in time series, the researchers use the long short-term memory (LSTM) network [14], which can find hidden association information in time series data such as video, audio and text. A multi-layered neural network structure of cascaded feed-forward layer and LSTM layer is proposed for word-level classification in speaker-based lip-reading.

In recent years, researchers have introduced attention mechanisms on convolutional

neural networks (CNN) to focus on areas of interest, and the classification and target detection of images have also achieved great success. For example, a CNN feature extraction method based on attention mechanism proposed by Vinyals et al. [13]. Furthermore, the mechanism of attention can be successfully applied in recurrent neural network (RNN) to find the relationship between the context. Since the changes between video frames of automatic lip-reading are continuous and happen in time series, the researchers use the long short-term memory (LSTM) network [14], which can find hidden association information in time series data such as video, audio and text. A multi-layered neural network structure of cascaded feed-forward layer and LSTM layer is proposed for word-level classification in speaker-based lip-reading.

**Dataset:** we're using GRID CORPUS dataset from university of sheffield. The Proposed model provides 100% accuracy for the trained videos for words it does not know it will predict word that sound similar from the dataset.(don’t say this until asked: For this the accuracy is typically around 80%).

GRID is a large **multitalker** audiovisual sentence corpus to support joint computational-behavioral studies in speech perception. In brief, the corpus consists of high-quality audio and video (facial) recordings of 1000 sentences spoken by each of 34 talkers (18 male, 16 female). Sentences are of the form "put red at G9 now". The corpus, together with transcriptions, is freely available for research use. GRID is described in more detail in this paper.

**Accuracy of human lip readers:**

Lip reading is a tricky business. Test results vary, but on average, most people recognize just one in 10 words when watching someone’s lips, and the accuracy of self-proclaimed experts tends to vary — there are certainly no lip-reading savants.

**Kanade – Lucas – Tomasi:It is shown mostly for dealing with the question that traditional techniques of image registration are usually expensive.**

**In Hidden Markov Model, it uses vectors of different lengths.. the three major problems are recognition and segmentation of ROI. it's good for face detection but not for feature detection**

**Viola Jones Algorithm: Viola-Jones has several advantages like feature selection that is sophisticated and an invariant detector that locates scales. We can scale the features instead of scaling the image itself. Since it’s a general scheme of detection, it can be trained for detecting other things like cars.**

**Youtube:** Youtube automatically recognises the speech and assigns captions (youtube cc) to videos using automatic speech recognition. “We’ve combined Google’s automatic speech recognition (ASR) technology with the YouTube caption system to offer automatic captions or auto-caps for short. Auto-caps use the same voice recognition algorithms in Google Voice to automatically generate captions for video.

**Movies:** Most existing systems also use ASR tech for audio to text conversion for the generation of subtitles.



