LEARNING AND ANALYSE ON MACHINE LEARNING

EC605 small project two

Dingjun Bian

## Introduction

Machine learning could be broadly described as computational methods using experience to improve performance or make accurate prediction. The word experience at here represents the past information that we have got. The information could be in many different forms , including digitalized human-labeled training sets. It currently includes three major directions, large scale machine learning ,deep learning and reinforcement learning.

There is a wide range of problems that could be solved through machine learning. It include text or document classification, natural language processing ,computer vision application and etcetera. Because of that reason, machine learning is already everywhere in our daily life. For example virtual personal assistants like Siri and Alexa, what they do is help us find the information that we need. These kind of assistants collect and refine the information of your previous interactions with them. Later, they can use the information to help you more efficiently. Another pretty influential application is on social media services. Whenever you read news or all kind of information. The platform will try to collect the data and analyze it. Later they will provide the features that they believe you would like. Facebook is a good specific good example, Facebook continuously notices the friends that you connect with, the profiles that you visit very often, your interests, workplace, or a group that you share with someone etc. On the basis of continuous learning, a list of Facebook users are suggested that you can become friends with. The last example is the prediction on commuting. Almost each and every of us have used apps like Uber and Lyft. When we are booking a cab, the app will estimates the price of the ride and show the result for us. When sharing these services, how do they minimize the detours? The answer is still machine learning. Jeff Schneider, the engineering lead at Uber ATC reveals in a an interview that they use ML to define price surge hours by predicting the rider demand. In the entire cycle of the services, ML is playing a major role. In general, machine learning is every where in our life, more further information will be provided and analyzed.

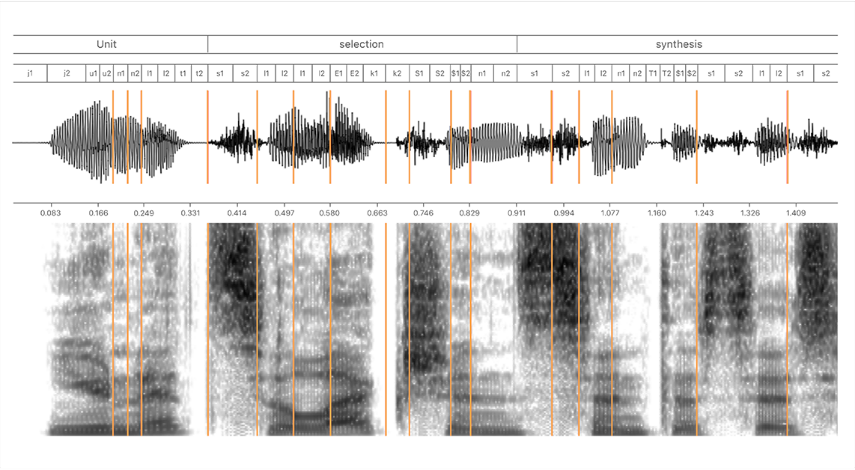
1. Reference
2. Foundations of Machine Learning, Mehryar Mohri, Afshin Rostamizadeh and Ameet Talwalkar
3. Apple Machine Learning Journal, Apple
4. Siri On-Device Deep Learning-Guided Unit Selection Text-to-Speech System, Tim Capes, Paul Coles, Alistair Conkie, Ladan Golipour, Abie Hadjitarkhani, Qiong Hu, Nancy Huddleston, Melvyn Hunt, Jiangchuan Li, Matthias Neeracher, Kishore Prahallad, Tuomo Raitio, Ramya Rasipuram, Greg Townsend, Becci Williamson, David Winarsky, Zhizheng Wu, Hepeng Zhang
5. Relative Learning and Analyses
6. Deep learning for Siri’s Voice:

Apple have its own website called Apple Machine Learning Journal. It is a kind of blog, which is used to detailing Apple’s work on machine learning, AI, and other related topics. The blog is written entirely by Apple’s engineers, and gives them a way to share their progress and interact with other researchers and engineers. Personal I am pretty curious about the technology that Siri is using, so I read a lot on the website, especially over machine learning.

First, in order to interact with people, there has to be synthesis of speech. There are essentially two speech synthesis techniques used in the industry: unit selection and parametric synthesis. Unit selection synthesis provides the highest quality given a sufficient amount of high-quality speech recordings, and thus it is the most widely used speech synthesis technique in commercial products. On the other hand, parametric synthesis provides highly intelligible and fluent speech, but suffers from lower overall quality. Therefore, parametric synthesis is often used when the corpus is small or a low footprint is required. Modern unit selection systems combine some of the benefits of the two approaches, and so are referred to as hybrid systems. Hybrid unit selection methods are similar to classical unit selection techniques, but they use the parametric approach to predict which units should be selected. In the meantime, Deep learning has also enabled a completely new approach for speech synthesis called direct waveform modeling (for example using WaveNet), which has the potential to provide both the high quality of unit selection synthesis and flexibility of parametric synthesis. However, given its extremely high computational cost, it is still not yet feasible for a production system.

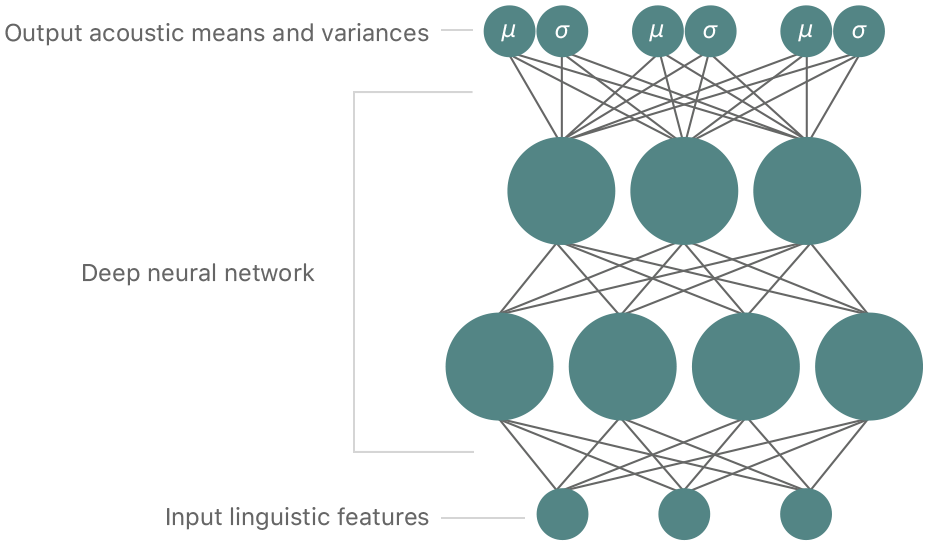
In order to building a high-quality text-to-speech (TTS) The first phase is to find a professional voice talent whose voice is both pleasant and intelligible and fits the personality of Siri. In order to cover some of the vast variety of human speech, they first record 10—20 hours of speech in a studio. The recording scripts vary from audio books to navigation instructions, and from prompted answers to witty jokes. Typically, this natural speech cannot be used as it is recorded because it is impossible to record all possible utterances the assistant may speak. Thus, unit selection TTS is based on slicing the recorded speech into its elementary components, such as half-phones, and then recombining them according to the input text to create entirely new speech. In practice, selecting appropriate phone segments and joining them together is not easy, because the acoustic characteristics of each phone depend on its neighboring phones and the prosody of speech, which often makes the speech units incompatible with each other.

Figure 1 illustrates how speech can be synthesized using a speech database segmented into half-phones.

figure 1

The goal of Siri’s TTS(text to speech) system is to train a unified model based on deep learning that can automatically and accurately predict both target and concatenation costs for the units in the database. Thus, instead of HMMs, the approach uses a deep mixture density network (MDN) to predict the distributions over the feature values. MDNs combine conventional deep neural networks (DNNs) with a Gaussian mixture models (GMM).

A conventional DNN is an artificial neural network with multiple hidden layers of neurons between the input and output layers. A DNN can thus model a complex and nonlinear relationship between input and output features. A DNN is trained by adjusting the weights of the network using backpropagation. In contrast, a GMM models the probability distribution of the output data given the input data using a set of Gaussian distributions, and is typically trained using the expectation maximization (EM) method. MDNs combine the benefits of DNNs and GMMs by using the DNN to model the complex relationship between input and output data, but providing probability distributions as output (see Figure 2).

figure 2

MDNs combine the benefits of DNNs and GMMs by using the DNN to model the complex relationship between input and output data, but providing probability distributions as output. Using the symbolic linguistic representation created by the text analysis module, the prosody generation module predicts values for acoustic features, such as intonation and duration. These values are used to select appropriate units. The task of unit selection has high complexity, so modern synthesizers use machine learning methods that can learn the correspondence between text and speech and then predict the values of speech features from the feature values of unseen text. This model must be learned at the training stage of a synthesizer using a large amount of text and speech data. The input to the prosody model are the numerical linguistic features, such as, phone identity, phone context, and syllable, word, and phrase-level positional features converted into convenient numerical form. The output of the model is composed of the numerical acoustic features of speech, such as spectrum, fundamental frequency, and phone duration. At synthesis time, the trained statistical model is used to map from the input text features into speech features, which are then used to guide the unit selection backend process where appropriate intonation and duration are crucial.

For Siri, they use an MDN-based unified target and concatenation model that can predict the distributions of both the target features of speech (spectrum, pitch, and duration) and the concatenation cost between the units to guide the unit search. Since the output of the MDN is in the form of Gaussian probability distributions, they can use the likelihood as a cost function for target and concatenation costs. The benefit of this approach becomes more clear when we consider the nature of speech. Sometimes the speech features, such as formants, are rather stable and evolve slowly, such as in the case of vowels. Elsewhere, speech can change quite rapidly, such as in transitions between voiced and unvoiced speech sounds. To take this variability into account, the model needs to be able adjust its parameters according to the aforementioned variability. The deep MDN does this using the variances embedded in the model. Since the predicted variances are context-dependent, we can see that they act as automatic context-dependent weights for the costs. This is important for improving the synthesis quality as we want to calculate the target and concatenation costs specific to the current context. The overall cost is a weighted sum of target and concatenation costs.

1. Learning and Analysis

For Siri, they chose to use deep mixture density network（deep MDM）as their neural network model. It is a feed-forward neural network model, built combining a convolutional neural network (CNN) and a with a Gaussian mixture models (GMM). It combined some strong points of both CNN and GMM. Conventional DNN has multiple hidden layers of neurons between the input and output feature ,so it can model a complex and adjusting the weights of the network using backpropagation. The GMM models the probability distribution of the date given the input data using a set of Gaussian distribution. The deep MDM achieved both of the strong points by using the DNN to model the complex relationship between input and output data, but providing probability distributions as output. Another advantage of using deep MDM is it has embedded variances. With the help of that we can deal with transitions between voiced and unvoiced speech more properly. In the meantime, the fundamental frequency of the speech region is highly dependent on the utterance as a whole. Because the deep MDM is recurrent, it can create natural and lively prosody in synthesized speech.