AI based Chat-bot using Azure Cognitive Services

Kapil Tajane
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
kapiltajane@gmail.com

Saransh Dave
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
saransh.dave@gmail.com

Pankaj Jahagirdar
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
pankajjahagirdar41@gmail.com

Abhijeet Ghadge
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
abhi360g@gmail.com

Akash Musale
Department of Computer Engineering
Pimpri Chinchwad College of Engineering
Pune, India
akashmusale100@gmail.com

***Abstract*— Letters ruled the earlier era in communication. Then with emergence of Telephones and subsequently mobile phones, voice conversations ruled the communication. However, currently, with the emergence of Internet and lots of social media, chat conversations are ruling the world. Think of your closest friend and ask yourself, have you talked more or chatted more? So, with popularity of chat in today's world, many technologists envisioned that chat couldn't just be a mode of communication between humans but also between a human and a computer. That's what chat-bot is. In some cases it is powered by machine learning (the more you interact with the chat-bot the smarter it gets). Or, more commonly, it is driven using intelligent rules (i.e. if a person says this, respond with that).**

A chat-bot can be useful in providing services in a variety of scenarios. These services include life-saving health messages, it may also include weather forecast or to purchase a new laptop, smartphone, and anything else in between. Many of the big companies like Google (Google Assistant), Amazon (Alexa), Microsoft (Cortana) and Oracle are spending good amount of energy and money for research on personal assistants. The following subjects would be touched upon for the development of chat-bot:

- **Using Azure Bot Architecture**
- **Using NLP for Language Understanding from the user and for the Language Generation**
- **Using Custom Vision services for the image recognition**

Keywords*— *Azure, NLP, Computer Vision

I. Introduction

The most glaring shortcoming of earlier chat-bots is they don't really understand what you're saying. They'll often misinterpret what you type, or ignore it completely. What more linguistically advanced chat-bots try to do is called natural language processing.

In fact, one of the greatest applications of the NLP in the current world is chat-bots [4] and therefore most of the big companies actively employ NLP researchers and every of the modern advanced personal assistants heavily use NLP.

We know that the real-life applications of cognitive services have been a breakthrough for software developers to use services like speech recognition, image recognition, natural language processing, etc. One such service currently being provided is Computer Vision Service. Computer vision focuses on giving visual capabilities similar to if not better to a computer system or even a machine.

Computer vision deals with the automatic extraction, understanding and analysis of useful information from a given single image or a set of images [9]. Automatic visual understanding is achieved by developing an algorithmic and theoretical basis.

When these services are used in real-life applications, it may be used for many applications of image recognition, expression analysis, object classification, OCR, etc. But one of the demerits of this services lies in the customised features provided by these services. For example, one cannot currently use these services for customising applications for recognizing themselves or their friends.

So, for this purpose a different Custom Vision Service can be used for customizing the application's visual capabilities to identify/recognize personalized objects through training. These services require very low training data set for training and very efficiently returns accurate outputs [16], [17].

A. Azure Bot Architecture.

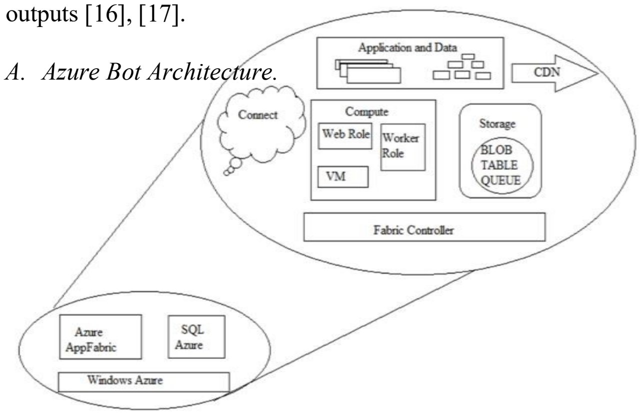


Fig 1: Azure Bot Architecture

978-1-5386-5257-2/18/$31.00 ©2018 IEEE

- Windows Azure platform has the following components:
    1. Windows Azure Compute
    2. Windows Azure Storage
    3. Fabric Controller
    4. Microsoft SQL Azure
    5. Windows Azure Platform AppFabric
    6. Content Delivery Network
    7. Connect

1. Windows Azure Compute: Windows Azure compute service runs applications on a Windows Server foundation. Any language that is supported by Microsoft Windows Azure can be used to create these applications. Visual Studio or other similar development tools can be used by developers; also they are free to use technologies like ASP.NET, PHP and Windows Communication Foundation. Windows Azure Compute consists of the following:

    1.1. Web Role has the primary purpose of running Web based applications. It is also possible to create applications using ASP.NET, Java, and also on-Microsoft technologies.

    1.2. Worker Role is designed so that they can run a variety of code. A worker role can be used to run a simulation, for example, video processing.

    1.3. Virtual Machine can be used to move some on-site Windows Server applications to Windows AZURE [17]. When an application is given to Azure to run by a developer, he submits configuration information along with it.

2. Windows Azure Storage: Windows Azure Storage consists of the following:

    2.1. Blob: It stands for Binary Large Object. This is the easiest method to store information. A storage account can have at least one container which contains one or more blobs. They store expansive amount of unstructured information. Blobs can be as huge as 1TB and to help transfer substantial number of blobs, they are additionally separated into blobs [17].

    2.2. Tables: To enable applications to work with data in an all the more fine-grained way, Windows Azure stockpiling gives tables. These are not relational tables. Be that as it may, the information they contain is really put away in set of elements or entities with properties. A table has no schema, rather properties can have different types, for example, int, string, bool, or date time. What's more, as opposed to utilizing SQL, an application can get to a table's information using SQL defined by OData. A single table can be very extensive, with billions of entities holding terabytes of information or data. Windows Azure storage can segment it crosswise over numerous servers if required to enhance execution.

    2.3. Queues: Queues are to some degree unique. They permit web role instance to interact with worker role instance. If a user wants to perform certain task using the Windows Azure Web role implemented Web page; the worker role after receiving the message reads the message and carries out the task.

B. *Using Natural Language Processing for Chat-bots*

Natural language processing also abbreviated as NLP is a field of artificial intelligence & computer science that deals with the interactions between humans and computers and vice-versa that is natural languages, specifically how to program computers to effectively process large amounts of natural language data.
Various challenges in NLP involve natural-language understanding, speech recognition, and also generation of natural language.

Although the whole semantic understanding is still considered a very distant goal [11], researchers are using the divide and conquer approach & have found out several sub-tasks which are useful for application development and their analysis. These range from the semantic, like word sense disambiguation [3], semantic-role labeling, named entity extraction and anaphora resolution, to the syntactic, such as part-of-speech tagging, chunking and parsing.

1. Multilayer perceptron (MLP)

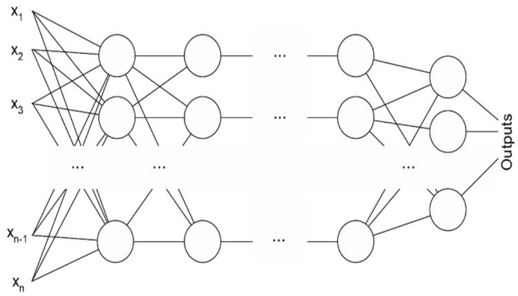
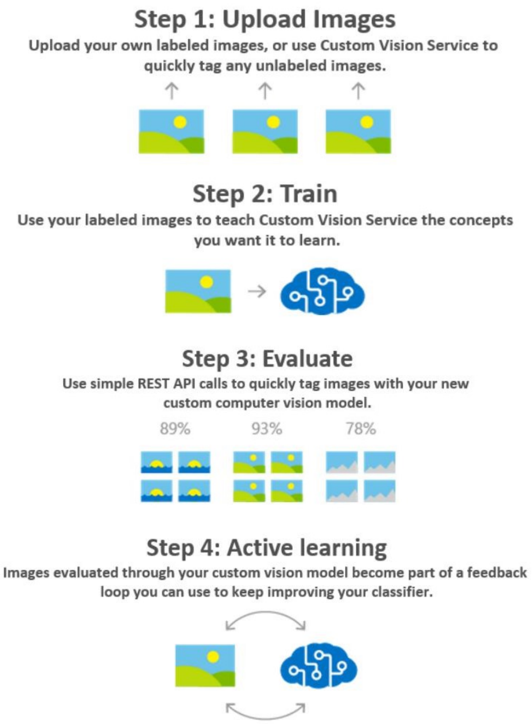


Fig 2: Multilayer Perceptron

A multilayer perceptron (MLP) has more than two layers. It makes use of a nonlinear activation function which is mainly hyperbolic tangent or logistic function that lets it classify data that is not linearly separable. Each node in a layer connects to each node in the following next layer making the network completely connected. For example, speech recognition and machine translation are some of the multilayer perceptron natural language processing applications [18].

C. *Azure Custom Vision Services*

Microsoft Azure Custom Vision Services: In the existing system of computer vision applications, there are various different kinds of API services available. But for our case we will consider the general Computer Vision API Version 1.0 provided by Microsoft Azure.



Step 1: Upload Images

Upload your own labeled images, or use Custom Vision Service to quickly tag any unlabeled images.

Step 2: Train

Use your labeled images to teach Custom Vision Service the concepts you want it to learn.

Step 3: Evaluate

Use simple REST API calls to quickly tag images with your new custom computer vision model.

Step 4: Active learning

Images evaluated through your custom vision model become part of a feedback loop you can use to keep improving your classifier.

Fig 5: Steps in Azure's Computer Vision

The Computer Vision API which is cloud based gives developers the access to various advanced algorithms for returning information & processing images [16]. Upload an image and the algorithm analyses visual content in various ways based on inputs and user choices. With the Computer Vision API users can analyze images to:

- Images will be tagged according to content.
- Categorizing of images.
- Quality of images will be identified.
- Human face detection.
- Recognize domain-based content.
- Generate descriptions of the content.
- Use optical character recognition to identify printed text found in images.
- Recognize handwritten text.
- Distinguish color schemes.
- Flag adult content.

D. *Automatic Speech Recognition Process*

Automatic speech recognition (ASR) is nothing but a method of transforming spoken language into text in real-time scenario [12], [13], [14]. It provides facility to store then process the words spoken by a person so that the stored words can be further analyzed. Improvement has been considerable in the field of speech recognition. This improvement happened due to lot research in this field. What makes possible to compare results from different ASR systems is the use of common speech corpora and large training sets. Progress has also been in acoustic modeling [11], such as contributions regarding context-specific Hidden Markov models (HMMs), changes in feature vectors over time or the presence of cross-word effects. Finally, progress in language modeling and search algorithms makes possibility of the better recognition of large vocabulary corpora and reduced experimentation cycles, respectively.

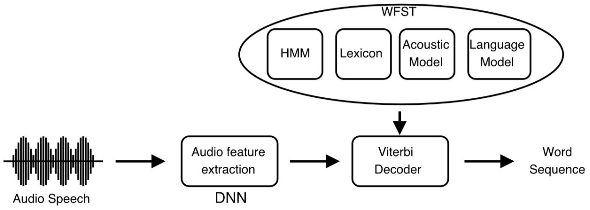


Fig 6: Speech to Text Conversion

Deep neural networks are used for this which refers to feed forward neural networks. Deep Neural Networks are feed forward Neural Networks with many layers. The DNN implements the acoustic model, it is the phase of the speech recognition which is in charge to translate the vector features of the audio signal to phonemes probabilities [6]. With the help of GPUs or dedicated accelerators there has been an increase in the performance of DNNs. Hence, achieving huge speedups and energy savings [10].

Hidden Markov Model: A Markov chain is an automaton with transitions and states where transitions have weights which are nothing but probabilities. It is used to assign probabilities to unambiguous sequences.

To compute a probability for a sequence of events that we can observe in the world a Markov chain can be useful when we need. In many cases, however, the events we are interested in may not be directly observable in the world. For example, in the case of parts of speech tagging, assigning tags like noun and verb to words we didn't observe part of speech tags in the world. The part-of-speech tags are called hidden because they are not observed. Speech recognition has a similar architecture; in which case we see acoustic events in the world and have to deduce the presence of hidden words that are the fundamental source of the acoustics.

An HMM is considered as a statistical Markov model in which the system being modeled is considered to be a Markov process having unobserved states. A process satisfies the Markov property if one can make predictions for the future of the process based only on its present state and also one could be knowing the process's full history, hence independently from such history; i.e., conditional on the present state of the system, its future and past states are independent.

The Viterbi algorithm is a dynamic programming algorithm which is used for identifying the most likely sequence of hidden states called the Viterbi path that results in a sequence of observed events, especially in the context of Markov information sources and hidden Markov models. For example, in case of speech recognition, the acoustic signal is treated as the observed sequence of events, and a string of text is considered to be the "hidden cause" of the acoustic signal. The Viterbi algorithm finds the most likely string of text given the acoustic signal.

Weighted Finite State Transducers also known as WFST is a Melay state machine widely used in speech recognition. This state machine allows to separately defining different information sources represented as independent WFST and at the same time provides a single and combined WFST for the model. The final WFST contains the entire speech process and is highly optimized, as the redundancy of the overall network is eliminated. In the context of speech recognition, the input labels represent the phonemes and the output labels the words. The WFST of an ASR is constructed offline and is used by the Viterbi search algorithm for the translations of phonemes to words.

The acoustic model is considered as a classifier that labels short fragments of audio into one of a number of phonemes, or sound units, in a given language. For example, the word ‘speech’ is comprised of four phonemes ‘s p iych’ . It simply defines the relationship between audio signals & linguistic units that make up speech. An acoustic model is a file that contains statistical representations of each of the distinct sounds that makes up a word. Each of these statistical representations is assigned a label called a phoneme. The English language has about 40 different sounds that are used for speech recognition, and therefore we have 40 different types of phonemes.

A Statistical Language Model is used to recognize speech which contains list of words with the probability of occurrence of each word.

CONCLUSION

Emergence of strong NLP algorithms, availability of powerful platforms like Azure for training and hosting chat-bot applications will enable chat-bots to penetrate our day to day life. They will become common and convenient means by which we would be interacting with our software systems and hardware systems. In fact, every home would be smart home having a voice enabled virtual assistant to interact with different smart devices. Customizing these services so as to meet personalized requirements of various organizations will boost the use of these services for regular tasks among local bodies.

REFERENCES

[1] Nils J. Nilsson, INTRODUCTION TO MACHINE LEARNING AN EARLY DRAFT OF A PROPOSED TEXTBOOK, Stanford University Stanford, CA 94305, Robotics Laboratory Department of Computer Science, Copyright © 2005 Nils J. Nilsson.

[2] Alex Smola and S.V.N. Vishwanathan, Introduction to Machine Learning. Yahoo! Labs Santa Clara –and– Departments of Statistics and Computer Science Purdue University –and– College of Engineering and Computer Science Australian National University, Copyright © Cambridge University Press 2008.

[3] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, Andrew Y. Ng, Reading Digits in Natural Images with Unsupervised Feature Learning, Google Inc., Mountain View, CA Stanford University, Stanford, CA.

[4] Milla T Mutiwokuziva, Melody W Chanda, Prudence Kadebu, A neural network based chat-bot , 2nd International Conference on Communication and Electronics Systems (ICCES 2017).

[5] Yun Kim, Convolutional Neural Networks for Sentence Classification, ACM Proceedings, 2008.

[6] Antony P.J, Santhanu P Mohan, SomanK.P,SVM Based Part of Speech Tagger for Malayalam, IEEE International Conference on Recent Trends in Information, Telecommunication and Computing, 2010.

[7] MNIST: http://yann.lecun.com/exdb/mnist/.

[8] J. J. Weinman. Unified Detection and Recognition for Reading Text in Scene Images. PhD thesis, University of Massachusetts Amherst, 2008.

[9] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition.In International Conference on Computer Vision, 2011.

[10] H. Lee, Y. Largman, P. Pham, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. In Advances in Neural Information Processing Systems 22, pages 1096–1104. 2009.

[11] G. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. Audio, Speech, and Language Processing, IEEE Transactions on, 2010.

[12] C. Bishop. Pattern recognition and machine learning.Springer, 2006.

[13] T. de Campos, B. Babu, and M. Varma.Character recognition in natural images.In VISAPP, Feb. 2009.

[14] B. Epshtein, E. Ofek, and Y. Wexler.Detecting text in natural scenes with stroke width transform.In CVPR, 2010.

[15] X. Chen and A. L. Yuille.Detecting and reading text in natural scenes.In CVPR, 2004.

[16] Azure Custom Vision Services: https://azure.microsoft.com/en-in/services/cognitive-services/custom-vision-service/.

[17] Azure Cognitive Services Documentation: https://docs.microsoft.com/en-in/azure/cognitive-services/.

[18] [Widrow& Lehr, 1990] Widrow, B., and Lehr, M. A., “30 Years of Adaptive Neural Networks: Perceptron, Madaline and Backpropagation,” Proc. IEEE, vol. 78, no. 9, pp. 1415-1442, September, 1990.

[19] Machine Learning Photo OCR: https://www.ritchieng.com/machine-learning-photo-ocr/.