Machine Learning in Science & Technology (Speech Recognition)

Nishant Ranjan Verma (*Author*)
Department of Big Data Analytics and Artificial Intelligence
Letterkenny Institute of Technology
Letterkenny, Donegal, Ireland

***Abstract—* In this paper we will discuss of the most widely used applications of machine learning that is known currently to us i.e. Speech Recognition. . Because of the high viability of speech signals, automated speech recognition, converting spoken words into text is still a difficult task. A recent field of machine learning is deep learning, also referred unsupervised function learning. Deep learning is becoming a popular speech recognition technology and has successfully replaced Gaussian speech recognition and function coding mixtures on an exponentially larger scale. In this paper we will try to discuss in detail the intricacies and how we have been benefited by machine learning/deep learning in betting our speech recognition systems.**

***Keywords*- speech recognition; GMM; HMM; recurrent neural network; RNN;DBM;RBM;**

Introduction

Machine learning is intended to render computers capable of learning problems on their own and solving problems. Using data and mathematical theories, computers can understand problems.

Deep learning has emerged as a fresh attraction over the last decade in the Machine learning area. It has been found that a variety of different research topics[1] have used deep learning. Input-fed algorithms in the form of multiple layered algorithms patterns. Typically, these models are neural networks consisting of Non-linear operations of various levels. Each learning function iteration adds one weights layer to a deep neural network. The subsequent layers it can eventually be loaded with learned weights to initialize a deeply tracked predictor[2], [3].

Research have demonstrated that fewer parameters are required to describe a certain non-linear function in a deep architecture then large number of parameters needed for representation for a shallower architecture with same feature. This demonstrates that from a statistical point, deeper architectures are more efficient [2],[3].

In particular, speaking, which is the key form of communication between human beings, has received a great deal of attention in the last five decades, right from the advent of artificial intelligence[4],[5]. A large number of research papers on the use of deep learning have been published to date for speech-related applications, specifically speech recognition[4],[5],[6],[7].

I. BRIEF HISTORY & RELATED WORKS

The traditional speech recognition systems are based on representing speech signals using Gaussian Mixture Models (GMMs) that are based on hidden Markov models (HMMs). This is because a speech signal may be viewed as a stationary signal in pieces or, in other words, a stationary signal for a short time. The speech signal can be approximated as a stationary phase in this short time scale, so it can be considered for several stochas-tic procedures as a Markov model. To model a spectral representation of the sound wave, each HMM utilizes a Gaussian mixture. In design and practical use, this type of scheme is called plain. However, for modeling non-linear or near non-linear functions, they are considered statistically inefficient[4],[6]. Neural networks, in comparison to HMMs, allow discriminatory training in a very efficient way. It works best for short time signals, (such as isolated words), and it is rarely effective when it comes to continuous speech signals. This is due to its inability to model continuous signals for temporal dependencies. Therefore, one approach uses neural networks as a pre-processing approach, e.g. feature transformation, reduction of dimensionality for HMM-based recognition[3].

There are several instances that illustrate that the use of deep neural networks delivers better results than conventional models. In 2012, Microsoft launched the latest version of its deep learning-based speech system, Microsoft Audio Video Indexing Service (MAVIS). Their final results clearly showed that the word error rate (WER) decreased by 30 percent on four major benchmarks relative to the state-of-the-art Gaussian mixture dependent models[2].

In the field of speech recognition, several surveys have been performed. Morgan[9], for example, carried out a speech recognition analysis helped by discriminatively qualified feed-forward networks.

Hinton [10] offers a summary of the usage of deep neural networks that include several hidden layers that use some of the latest techniques to train them. The description summarizes the results of four separate research groups that collaborated to reveal the value of a feed-forward neural network that has as an input quite a few frames of coefficients and generates subsequent probabities as an output over HMM states. The findings collected have shown that deep neural networks that contain several hidden layers and are trained by new techniques outperform GMMs-HMMs, often by a wide margin, on a range of speech recognition benchmarks.

An overview summary was given by Deng[11] at ICASSP-2013. The paper addressed the history of deep neural network creation for speech recognition acoustic models. The overview showed the rapid progress in acoustic models that use deep neural networks that, compared to those based on GMMs, can be seen on several fronts. The paper also revealed that in other signal processing applications, and not only speech recognition, these acoustic models can also be applicable and improve performance.

Deng[12] conducted a summary on the work performed by Microsoft since the year 2009 in the field of speech using deep learning. Speech-related applications included extraction features, language modeling, acoustic models, speech comprehension as well as estimation of dialogue. Experimental findings have clearly shown that the features of the speech spectrogram are more advanced compared to conventional speech spectrogram features in MFCC with deep neural networks in comparison to practice using GMMs - HMMs

Li[13] provided the basics of state-of-the-art solutions for both computational and phonological perspectives on automated spoken language recognition. In the field of spoken language recognition, which was primarily motivated by breakthroughs in related signal processing fields such as pattern recognition and cognitive science, tremendous progress has been made in recent years.

Li[14] gave an overview of current robust noise techniques built over the past three decades for automated speech recognition. This research helps the reader distinguish between the various noise-robust techniques and offers a detailed insight into the dynamic performance tradeoffs that should be taken into consideration when deciding between the techniques available.

II. CORE STUDY

A. SPEECH SIGNALS

Speech signals can provide us with different kinds of infor- mation, following are few noticeable:

- Speech recognition, provides us information on the speech signals.
- Speaker recognition that provides information about the speaker identity.
- Emotion recognition, which provides information about on emotional state of speaker.
- Language recognition,states information about spoken language.
- Gender recognition, helps with gender inforamtion

On the basis of the data embedded in his / her speech signal using the machine (computer), automatic speaker recognition can be identified as the process of recognizing the unknown speaker. Recognition of speakers is divided into two parts: recognition of speakers and verification of speakers (authentication). As the speaker identification portion, the method of determining which of the registered speaker a given utterance corresponds to is named. In public buildings or for the media, this component can be used. These cases include, but are not limited to, calls to radio stations, insurance companies or recorded tapes of district or other government authorities[15],[16]. The portion of speaker verification is described as the procedure for admitting or discarding the identity of the claimed speaker. In order to authorize the asserted speaker identity, the applications of this section include the use of voice as a focal factor.

Some of the application areas of this division are corporate relationships using telecommunications networks, data set access services, security control for private information areas, remote access to computers and intelligent healthcare systems [17], [18].

Emotion recognition by Computer can be defined as the task of recognizing unknown emotions based on information inserted into speech signals. The field of emotion recognition is subdivided into branches of emotion detection and emotion verification. Low speaker recogniton is considered to be the most challenging issues [19]–[22].

Recognition of language is the problem of determining the natural language in which a speech material is given. Differentiating between strongly associated languages is one of the enormous challenges of language recognition systems.

Automatic recognition of gender is the method of identifying whether a male or female is the speaker. Automatic recognition of gender usually results in high accuracy without much effort because the results of such recognition are binary (either male or female)[23],[24].

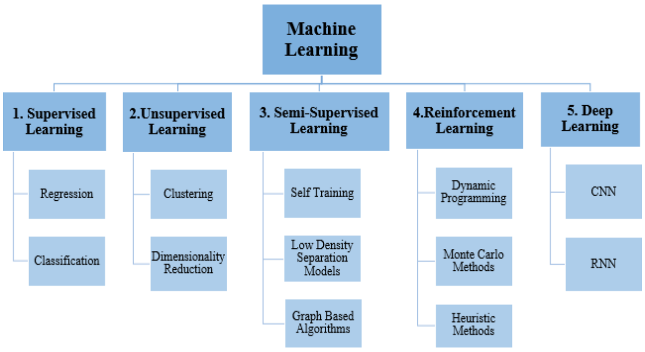

Fig 1 Machine Learning Classification

B. MACHINE LEARNING

Machine learning is characterized as the field of study that gives computers the ability to learn without being specifically programmed to do so from input data. From evaluated data and new input data, the learning process is conducted iteratively. This iterative feature helps computers, when exposed to new data, to recognise secret insights and recurring trends and to use these findings to adapt[25]. In this learning method, the various types of data used can vary from observations and examples to guidance and direct experience[25]. In producing consistent and repeated outcomes, the information acquired can benefit. Machine learning can thus be defined as a system that learns from past experiences and uses knowledge acquired to do better in the future[25],[26]. The main phases of machine learning are shown in Figure 2. In recent years , machine learning has gained a lot of popularity and can be used in many applications. Machine learning focuses on learning and adapting automatically when exposed to data without the need for human interaction. As stated earlier, the past influences the future, so machine learning is perceived as an example of programming. In order to solve a certain problem, instead of directly programming the computer to perform that problem, we let it create its own program on the basis of examples given from which the computer learns[26].

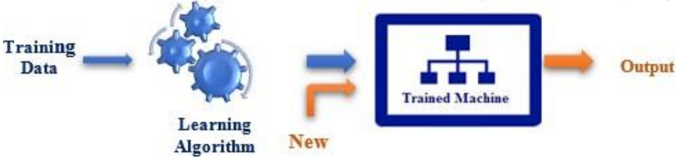

Fig 2 Phases of Machine Learning

There are five types of machine learning concepts/techniques:-

- Supervised Learning
- Unsupervised Learning
- Semi-Supervised Learning
- Reinforcement Learning
- Deep Learning

a. Supervised Learning

Supervised Learning uses labeled data to train the model/algorithm. It is called as labeled since it is present in pairs and its output which can be stated as supervisory signal. In here the correct output is know and the algorithm tries to predict the output and is corrects and learns from precdicted and already provided output. Through the analysis of training data by the supervised learning algorithm, we get a Classfier Function if the output was discrete and Regression Function if it is continuous. Regression (Linear,Multiple and Polynomial) and Classification Algorithms are two main categories of supervised learning.

b. Unsupervised Learning

Unlike Supervised Learning, unsupervised learning works without human supervision to be precise. Here the model algorithm works on its own to predict the outcome by discovering the pattern from the input data. There unsupervised learning is more subjective than supervised . There are basically three types of Unsupervised Learning: clustering, dimensionality reduction and anomaly detection.

c. Semi-Supervised Learning

This method falls between the above two methods, here we have huge data some of which is labeled and others are not. Both the previous algorithms can be utilised to train the learning algorithm in semi-supervised learning. Semi Supervised Learning make use of one of the assumptions: smoothness assumptions, cluster assumption and manifold assumption.

d. Reinforcement

It is basically learning achieved by interaction with the environment. It learns through its own practices rather then being told. It uses exploitation and exploration which is simply trial and error way of learning. In the form of a numerical reward value, the success of an action is determined by a signal obtained by the reinforcement learning agent. The goal of the agent is to learn to choose acts that maximize the numerical reward value. Actions can not only affect the current situation and the current value of the reward, but also affect successive situations and the values of the reward.

e. Deep Learning

Deep Learning is a sub field of machine learning, which is touces between neural networks, graphical modeling, optimization, artificial intelligence, identification of patterns and processing of signals. It algorithms uses higher level features from raw inputs. Earlier signal processing techniques uses only minimal structured architecture consisting only of one or two layers. Example being GMM and SVM which best suits constrained problems.

Deep belief networks (DBN) was introduced, which is a class of deep gene-erative models. A stack of restricted Boltzmann machines (RBMs) consists of the DBN. A greedy learning algorithm is at the heart of the DBN that optimizes DBN weights linear to the size and depth of the networks at time complexity[27]. Incorporation of hidden layers with large number of neurons of deep neural networks has increased modelling possibilities and thus closely required configurations[28].

Precisely there are three categories of Deep Learning:

a. Deep Network for Unsupervised Learning
b. Deep Network for Supervised Learning
c. Hybrid Deep Networks

i. Convolutional Neural Network (CNN)

This is a type of discriminative deep architecture where every model contains a convolutional layer and pooling layer and stacked on each other. In the convolutional layer, several weights are exchanged, while the pooling layer, on the other hand, sub-samples the output from the convolutionary layer and reduces the data rate of the lower layer. The sharing of weight along with the correctly selected pooling systems results in the invariance properties of the CNN. The CNN can also be used in speech recognition[29], with some necessary modifications in the CNN for image processing purposes so that it integrates speech properties.

ii. RECURRENT NEURAL NETWORKS (RNN)

In cases where the depth of the input data sequence can be as long as the length since RNNs allow parameters to be shared across the different layers of the network, recurrent neural networks (RNNs) are regarded as a class of deep networks for use in unsupervised learning. The RNN is used primarily for the purpose of using previous data samples to predict the future data series. When it comes to modeling sequence data, such as speech or text, the RNN is very prevalent. It has been shown that RNNs trained with Hessian free optimization are capable of generating sequential text characters[30].

III. AUTOMATIC SPEECH RECOGNITION MODEL

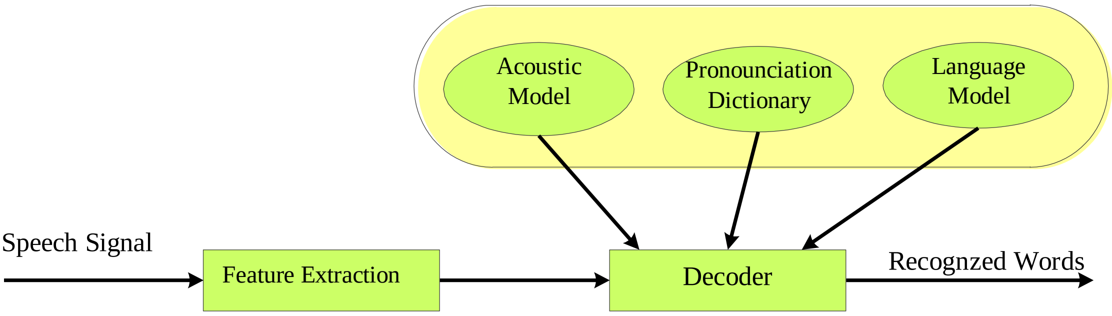

Fig 3 shows components of large vocalbulary continuous speech recognizer[31][32]. For any given W, by concatenating phone models, the are of importance in achieving the same.

The waveform inputted from a microphone is transformed into a sequence of fixed size accoustic vectors $\boldsymbol{Y} = [y1, \cdots, yT]$. This is called feature extraction. The sequence of word $\boldsymbol{W} = [w1, \cdots, wL]$, which has possibly generated Y, is then attempted to be found by the decoder.[6]

$$\hat{W} = \underbrace{argmax}_{w}\{P(\boldsymbol{W}|\boldsymbol{Y})\}$$

P(Y|W) is mentioned by acoustic model P(W) is determined using langurage model
Bayes's Rule to transform the above into equivalent problem.

$$\hat{W} = \underbrace{argmax}_{w}\{P(\boldsymbol{Y}|\boldsymbol{W})P(\boldsymbol{W})\}$$

The corresponding acoustic model is synthesized to make words as described by a dictionary of pronunciation. From training data consisting of speech waveforms and their orthographic transcriptions, the parameters of these phone models are calculated. Usually, the language model is an N-gram model in which only its N-1 predecessors are conditioned on the likelihood of each term. Parameters of the N-gram are estimated by counting N-tuples in the relevant corporate text.[6]

Applying DBN for Speech Recognition

Deep Belief Networks (DBNs) are neural networks consisting of a stack of small Boltzmann machine (RBM) layers trained one at a time to cause more and more abstract representations of the inputs in subsequent layers in an unsupervised manner.
A background window of n successive frames of feature vectors is used to set the states of the visible units of the lower layer of the DBN, which generates a distribution of probability over the possible labels of the central frame, to apply DBNs with defined input and output dimensionality to phone recognition. A series of probability distributions over the possible labels for each frame is fed into a standard decoder to produce speech sequences.[6]

IV. PERFORMANCE EVALUATION

Evaluation of the efficacy of machine learning can be achieved in several ways. Measuring accuracy and F1 calculation is the common way to calculate the accuracy of machine learning work. (1)[33]

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (1)$$

The potential outcomes are positive and negative by forecasting the binary classification outcomes. The outcomes of the model forecasts may be as follows: true positive (TP), true negative (TN), false negative (FN) and false positive (FP) predictions. F1 measurements can be determined as multiplying accuracy (recall), multiplying by recall and dividing by accuracy plus recall as multiplying accuracy
(recall).(2)

$$F1 = 2\left(\frac{precision \times recall}{precision + recall}\right) \qquad (2)$$

In order to quickly and efficiently construct a model, pre-processing[34] is performed before constructing a model. For instance, extraction of features[35] would identify the input feature and selection of features[36] and select features that are distinctive. The feature reduction[8] decreases the size of the feature to reduce the calculation time and improves the prediction speed.

$$precision = \frac{TP}{TP+FP} \text{ and } recall = \frac{TP}{TP+FN}$$

After the prediction, post-processing changes the data to increase or boost the performance quality of the model[33], such as a 1-to-k decoder to change the output of the model and to define the category in the desired format.

CONCLUSION

In this paper we tried to review and go through the time line and development of technologies in relation to Speech Recognition. Also we tried looking on various concepts that plays the basic fundamental of speech recognition development.
This paper also helps to depict how machine learning's latest field of deep learning has opend gateways for bettering results for automatic speech recognition systems.

REFERENCES

[1] A. H. Meftah, Y. A. Alotaibi, and S.-A. Selouani, ``Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis,'' *IEEE Access*, vol. 6, pp. 72845_72861, 2018.

[2] Y. Xie, L. Le, Y. Zhou, and V. V. Raghavan, ``Deep learning for natural language processing,'' in *Handbook of Statistics*. Amsterdam, The Netherlands: Elsevier, 2018.

[3] J. Padmanabhan and M. J. J. Premkumar, ``Machine learning in automatic speech recognition: A survey,'' *IETE Tech. Rev.*, vol. 32, no. 4, pp. 240_251, 2015.

[4]H. Singh and A. K. Bathla, ''A survey on speech recognition,'' *Int. J. Adv. Res. Comput. Eng. Technol.*, no. 2, no. 6, pp. 2186–2189, 2013.

[5]M. A. Anusuya and S. K. Katti, ''Speech recognition by machine: A review,'' *Int. J. Comput. Sci. Inf. Secur.*, vol. 6, no. 3, pp. 181–205, 2009.

[6]Y. Zhang, ''Speech recognition using deep learning algorithms,'' Stanford Univ., Stanford, CA, USA, Tech. Rep., 2013, pp. 1–5. [Online]. Available: https://scholar.google.com/scholar?as_q=Speech+Recognition+Using+Deep+Learning+Algorithms&as_occt=title&hl=en&as_sdt=0%2C31

[7]I. Shahin, A. B. Nassif, and S. Hamsa, ''Novel cascaded Gaussian mixture model-deep neural network classifier for speaker identification in emo- tional talking environments,'' *Neural Comput. Appl.*, to be published.

[8]B. Kitchenham and S. Charters, ''Guidelines for performing Systematic Literature reviews in software engineering version 2.3,'' *Engineering*, vol. 45, no. 4, p. 1051, 2007.

[9]N. Morgan, ''Deep and wide: Multiple layers in automatic speech recog- nition,'' *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 1, pp. 7–13, Jan. 2012.

[10]G. Hinton *et al.*, ''Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,'' *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[11]L. Deng, G. Hinton, and B. Kingsbury, ''New types of deep neu- ral network learning for speech recognition and related applications: An overview,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8599–8603.

[12]L. Deng *et al.*, ''Recent advances in deep learning for speech research at Microsoft,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 8604–8608.

[13]H. Li, B. Ma, and K. A. Lee, ''Spoken language recognition: From fundamentals to practice,'' *Proc. IEEE*, vol. 101, no. 5, pp. 1136–1159, May 2013.

[14]J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, ''An overview of noise- robust automatic speech recognition,'' *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.

[15]D. A. Reynolds, ''An overview of automatic speaker recognition technol- ogy,'' in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 4, May 2002, pp. 4072–4075.

[16]S. Furui, ''Speaker-dependent-feature extraction, recognition and pro- cessing techniques,'' *Speech Commun.*, vol. 10, nos. 5–6, pp. 505–520, Dec. 1991.

[17]G. Zhou, J. H. L. Hansen, and J. F. Kaiser, ''Nonlinear feature based clas- sification of speech under stress,'' *IEEE Trans. Speech Audio Process.*, vol. 9, no. 3, pp. 201–216, Mar. 2001.

[18]C. Fredouille, G. Pouchoulin, J.-F. Bonastre, M. Azzarello, A. Giovanni, and A. Ghio, ''Application of Automatic Speaker Recognition techniques to pathological voice assessment (dysphonia),'' in *Proc. Eur. Conf. Speech Commun. Technol. (Eurospeech)*, 2005, pp. 149–152.

[19]I. Shahin, ''Employing emotion cues to verify speakers in emotional talking environments,'' *J. Intell. Syst.*, vol. 25, no. 1, pp. 3–17, 2016.

[20]I. M. A. Shahin, ''Employing both gender and emotion cues to enhance speaker identification performance in emotional talking environments,'' *Int. J. Speech Technol.*, vol. 16, no. 3, pp. 341–351, Sep. 2013.

[21]I. Shahin, ''Speaker identification in emotional talking environments based on CSPHMM2s,'' *Eng. Appl. Artif. Intell.*, vol. 26, no. 7, pp. 1652–1659, Aug. 2013.

[22]I. Shahin, ''Identifying speakers using their emotion cues,'' *Int. J. Speech Technol.*, vol. 14, no. 2, pp. 89–98, 2011.

[23]T. Vogt and E. André, ''Improving automatic emotion recognition from speech via gender differentiation,'' in *Proc. Lang. Resour. Eval. Conf.*, Jan. 2006, pp. 1123–1126.

[24]I. M. A. Shahin, ''Gender-dependent emotion recognition based on HMMs and SPHMMs,'' *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 133–141, 2013.

[25]R. Schapire, *Theoretical Machine Learning* (Lecture). Princeton, NJ, USA: Princeton Univ., 2008, pp. 1–6.

[26]P. Domingos, ''A few useful things to know about machine learning,'' *Commun. ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[27]G. Hinton *et al.*, ''Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,'' *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[28]Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, no. 7553, p. 436, May 2015.

[29]O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, ''Convolutional neural networks for speech recognition,'' *IEEE/ACM Trans. Audio, Speech Language Process.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2015.

[30]J. Martens and I. Sutskever, ''Learning recurrent neural networks with Hessian-free optimization,'' in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 1033–1040.

[31]. S. Young, "Large Vocabulary Continuous Speech Recognition: A Review," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 45–57, 1996.

[32]. J. Baker, L. Deng, J. Glass, S. Khudanpur, Chin hui Lee, N. Morgan, and

[33] Michael Steinbach and Vipin Kumar, Introduction to Data Mining, Addison-Wesley, 2005.

[34].Feature extraction construction and selection: A data mining perspective, Springer Science & Business Media, vol. 453, 1998.

[35].I. Guyon and A. Elisseeff, "An introduction to feature extraction" in Feature extraction, Berlin, Heidelberg:Springer, pp. 1-25, 2006.

[36].Y. Yang and J.O. Pedersen, "A comparative study on feature selection in text categorization", Icml, vol. 97, pp. 412-420, July 1997.

[37]A. Phinyomark, P. Phukpattaranont and C. Limsakul, "Feature reduction and selection for EMG signal classification", Expert Systems with Applications, vol. 39, no. 8, pp. 7420-7431, 2012.