ENHANCING LEARNING IN SPOKEN LANGUAGE UNDERSTANDING BY MITIGATING INTERNAL COVARIANT SHIFT IN LEARNING MODELS

Sheetal Jagdale
Shree L.R Tiwari engineering college, kanakia park,
Mira Road (East), Thane, Maharashtra, 401107, India.
sheetal.jagdale2@gmail.com

Milind Shah
Electronics and telecommunication engineering department,
Fr. C. Rodrigues institute of technology, sector 9A, Vashi, Navi Mumbai,400703, India.
milind.shah@fcrit.ac.in

Abstract - Spoken Language Understanding (SLU) is a part of the Spoken dialogue system (SDS) and it assists to achieve the user's goal. SLU maps user utterance to a logical structure that can be easily understood by the computer. SLU accomplish this task by using deep learning models such as Convolution Neural Network (CNN) from machine learning. These machine-learning models suffer from an Internal Covariant Shift (ICS). Due to ICS, the learning efficiency of the deep model has reduced which in turn reduces the learning efficiency of SLU. An ICS in SLU can be reduced by techniques proposed in machine learning. The techniques are batch normalization, cosine normalization, group normalization, layer normalization, and instance normalization. The work in the paper extends these techniques to a deep model in SLU to mitigate the ICS and to enhance the performance of SLU. The results show improvement in SLU performance with normalization. The speed of training is also improved with normalization. The evaluation parameters are F-score, accuracy, and detection rate. When the SLU is trained with less training data, SLU with the instance normalization displayed good results. SLU with instance normalization displayed minimum variation in the learning curve. SLU with batch normalization displayed better accuracy, detection rate, and F-score for both area and price slot. In terms of training time, SLU with cosine normalization and batch normalization required the least time.

***Keywords*:** Keyword1; keyword2; keyword3. Internal covariant shift; Spoken language understanding; learning models; machine learning.

1. Introduction

Spoken language understanding (SLU) tasks require logical representations of user utterances inputs. This logical representation can be directly learned from data. Learning models have quickly become standard learning methods and have yielded improvements in SLU. Peng et al. in paper [1] presented a unified framework for multi-tasking and multi-domain. In the paper, the learning model is used for multi-tasking and multi-domain task. The work in paper [2] presents a single recurrent neural network that integrates three task domain classification, intent determination, and slot filling over multiple domains for a single natural language understanding model. In another paper [3] a learning model architecture is for multi-domain task-oriented of spoken language understanding. Liu et al. in paper [4] presented a system that can perform better when tested on incomplete or partial queries using learning models. The paper research [5][6] investigates different classifiers from machine learning for SLU. Thus, learning models plays important role in accompanying SLU task. This learning model suffers from the ICS. Hence learning performance of the SLU is reduced. Thus, ICS reduces the performance of SLU. The work in this paper aims to improve learning in SLU by reducing internal covariant shift in learning models in SLU. The techniques from machine learning are investigated to reduce internal covariant shift. The work in the paper investigates batch normalization, cosine normalization, group normalization, layer normalization, and instance normalization to reduce ICS.

Machine learning models are widely used in improving the performance of many applications. In intrusion detection, machine-learning models are used for optimizing feature selection [7]. The feature selection and deep models are used to enhance the performance of detection of breast cancer in magnetic resonance imaging [8]. Classification model from machine learning is used for breast cancer classification [9]. Deep CNN model is also used for the diagnosis of COVID 19 [10]. The convolution regression model is also used for crop production prediction [11]. Another work [12]. RNN is used for disease prediction. Prakash Annamalai in paper [13] used a deep belief network for face recognition.

In paper [14], batch normalization is investigated for SLU. This work investigates group normalization, cosine normalization, and instance normalization to mitigate the ICS in SLU. The objective of work in the paper is as follow:

1) To improve learning in SLU by reducing internal covariant shift.

2) To investigate techniques from machine learning to reduce the internal covariant shift in SLU.

3) To do a comparative study of different techniques and analyze the impact of these techniques on SLU performance.

2. Background

2.1. *Spoken language understanding*

SLU framework is a field emerging from Natural Language Processing (NLP) and it's growing by utilizing advances from machine learning. The paper [15] presented a method is to enriching word embedding with semantic technique. Future work presented in the paper is to implement word embedding to complex networks such as bidirectional long short-term memory with memory networks. In this paper, the machine learning model is used for intent detection. Variation Bayesian model was used for intent detection from user queries [16]. The experiments were carried out with a large volume of searched queries and results showed significant improvement over the state-of-the-art system. In this paper, the machine learning model is used for intent detection. Word embedding or word representation use numbers to represent words in machine learning. The author in the paper [17] uses a large-scale word representation machine learning model to solve the intent and domain classification problem when a lot of unlabeled data is available. In this work, a machine learning model is used to extract features from unlabeled data to solve intent classification and domain classification problems.

The research in paper [18] presented an architecture for SLU which has a pre-training method for training deep learning models. The future work presented in the paper is to build an end-to-end framework where the learning of the latent class labels from unlabeled data must be handled by the same network. Here the machine learning model is used for the slot filling task without large manually trained data. The task of the SLU of a dialogue system is to identify semantic components in user utterances [19]. Bhargava et al. in paper [20] presented a technique for slot filling and intent detection. In a human-computer interface, there are turns of conversation. The future scope of this paper is to implement joint modeling of intent and slot determination. SLU task involves slot filling, intent detection, and semantic frame parsing. Nature language provides meaningful full properties that give organized logical structure for understanding [21]. Chen et al. presented in paper [22] proposes to apply information guided primary networks which furthermore fuse topologies by earlier information, for joint semantic parsing. Words or phrases from the vocabulary are converted to real numbers which constitute a vector. Word embedding is now the most widely used techniques in speech and language processing [23]. Celikyilmaz et al. in paper [24] used word embedding with CRF and CRF-CNN network for sequence tagging task. The experiment results demonstrated improvement in F-score by using the CRF-CNN technique then only the CRF model. The domain used in the paper is the movie domain. Future work in the paper is to obtain rich embedding for other domains such as restaurant and transportation. In this paper, machine learning is used for sequence tagging.

2.2. *Techniques for reducing Internal covariant shift:*

Deep learning models suffer from the ICS. It is caused due to the change in the distribution of input to each layer. This reduces the learning speed and decreases the performance of the model. The techniques for reducing internal covariant shift are batch normalization, cosine normalization, group normalization, layer normalization, and instance normalization.

2.2.1 Batch-normalization

Batch normalization reduces the ICS. It uniforms the input distribution across the entire batch [25]. The statistical parameter such as mean and variance are calculated and are used to standardize the input across the batch.

2.2.2 Cosine normalization:

Cosine similarity can be used instead of the dot product in the neural network [26]. The results in the paper were obtained on a text dataset. Cosine normalization was compared with layer normalization, weight normalization, and batch normalization. Cosine normalization displayed good results.

2.2.3 Switchable normalization:

Switchable Normalization utilizes statistics across batch, channel, and layer [27][28]. There is no fixed normalization technique used. It utilizes layer normalization, Instance normalization, and batch normalization. Depending upon the application algorithm decides which technique to use more or less. The sparse switchable normalization in which ratios are sparse [29].

2.2.4 Group normalization:

Group normalization alternative to batch normalization is proposed in paper [30]. Batch normalization is not effective if the batch size is small. Group normalization is independent of batch and it divides the channel into groups.

2.2.5 Instance normalization:

The instance normalization calculates means and variance. It standardizes the input across each channel of training [31]. Instance normalization applied to style transfer displayed better results than batch-normalization. Adaptive instance normalization is proposed in literature, which is an extension instance normalization [32].

3. Proposed Methodology:

ICS reduction methods are the most widely used technique in machine learning. These methods have improved enhanced learning and model performance. The work in this paper investigates ICS reduction methods for SLU. The proposed methodology is shown in figure 1. SLU maps the user utterance into a logical structure that the computer can understand. The three inputs to SLU are candidate pair, context, and user utterance. The context is the earlier interaction of the user. The candidate is predefined slot value pairs. The third inputs are user utterance. All three inputs are given to the convolution layer. The deep representation from all the three inputs are extracted. The extracted deep features are standardized to reduce internal covariant shift. This task is performed at the normalization layer. The next block is the model is of the ReLU activation function. The output of ReLU is again given to another layer of convolution, normalization, and ReLU. The last layer of the deep network is max pooling. Deep features of three inputs are obtained. The Neural belief tracker (NBT) [31] model for SLU is used for investigation. Next candidate slot value pair and user utterance representation are fed to semantic decoder. The semantic decoder will decide if the user utterance match to candidate slot value pair. The output of the semantic decoder and context are fed to the summary generator. The output of the summary generator is given to the softmax classifier. Then the final decision for the slot value pair is done.The experiment on the proposed model is done with different ICT reduction methods. The methods used for investigation are batch-normalization, layer normalization, group normalization, switchable normalization, and instance normalization. The dataset for evaluation is the WOZ dataset. The experimentation was done for three slots area, food, and price range. Each slot experiment was repeated for a varying quantity of training data. The training data was selected 20 %, 40%, 60%, 80%and 100%. The learning curve was obtained to evaluate the performance of the model with different normalization techniques. The evaluation metric used is the detection rate, F-score, and accuracy. The other parameters for comparison used are time and speed of learning.

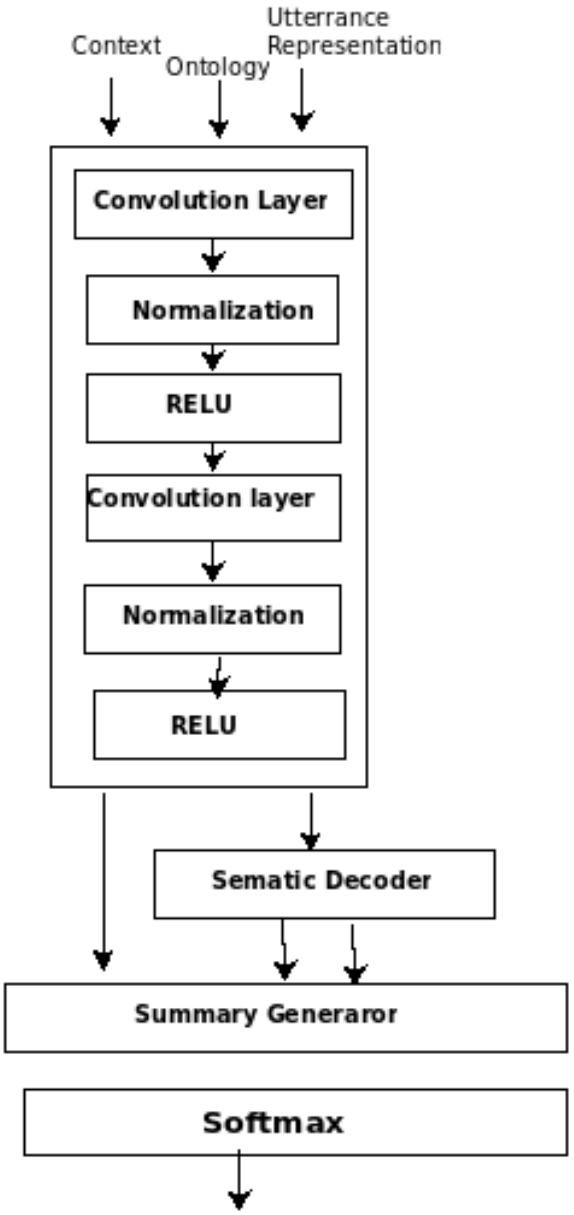

Figure 1 Proposed methodology for SLU

4. Results and Discussion:

4.1 Dataset:

The experimentation is done on WOZ 2.0 datasets. The domain for the dataset is a restaurant. The total number of dialogues is 1200. The total dataset is split into 600 training data, 200 validation data, and 400 testing data.

4.2 Experimental Setting:

NBT model [33] of SLU is used for experimentation. NBT model incorporating batch normalization, cosine normalization, group normalization, Layer normalization, and instance normalization is investigated. Experimentation was done for the two slots area and price range. The model is trained on varying quantities of training data. The effect of a model trained with different training data on the performance of the model was observed. The amount of training data is 120, 240, 360, 480, and 600.

4.3 Evaluation Parameter:

Evaluation parameters used for experimentation are f–score, accuracy, and detection rate.

4.4 Results:

Results were obtained for the area and price range slot.

4.4.1 Results for Area Slot:

Figure 2 displays the F score learning curve for the area slot for NBT with and without normalization. The graph shows that the improvement in the model with normalization then without normalization. The model with batch normalization learning curve shows steady improvement in model performance. The highest f–score was obtained by batch normalization. The first learning curve indicates that the model with normalization has less variation in F–score as compared to without normalization. The learning curve with a group and instance normalization has shown the least variation. The second observation is that the generalization of the model with normalization is better than without normalization. The model with normalization has a better f-score both in the case of minimum and maximum training data. In the case of minimum training data, the best result was obtained with model incorporated instance normalization. In the case of the highest training data, the best results were obtained with batch normalization.

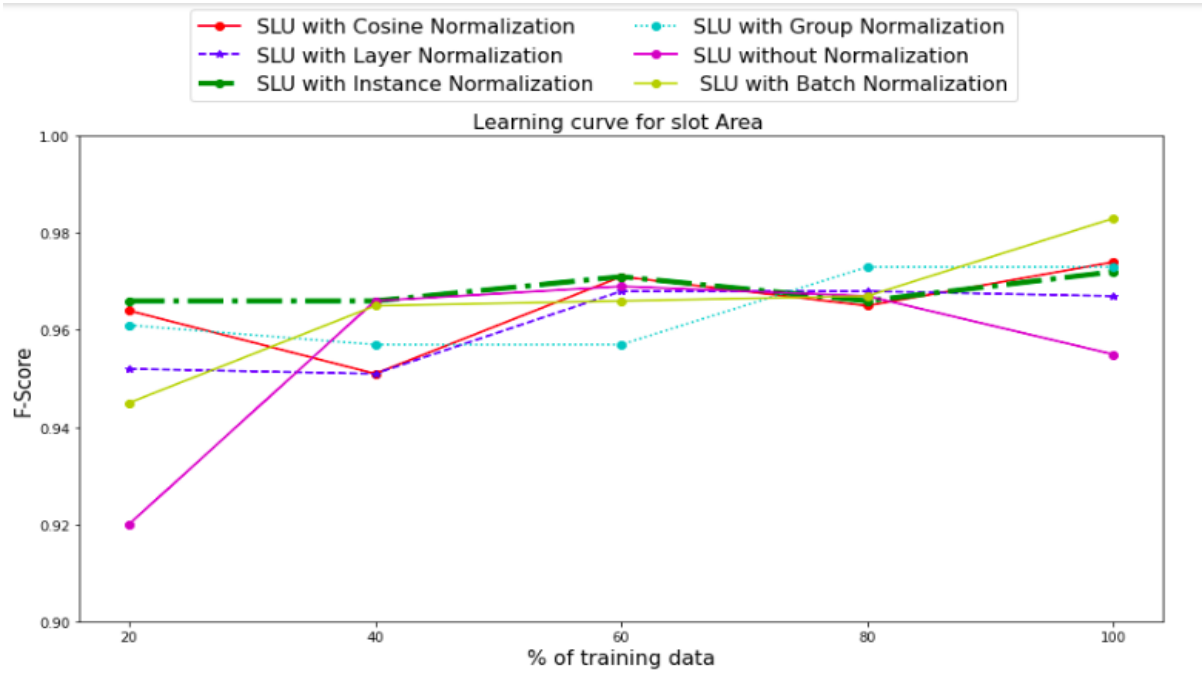

Figure. 2 F–score learning curve for the slot area

Table 1 indicates the accuracy of the models with normalization and without normalization. There is improvement in accuracy with normalization. The highest accuracy is obtained by model with batch normalization. The least accuracy is obtained by layer normalization.

Table 1. Accuracy

Sr.No	Method	Accuracy
1	NBT without Normalization	91
2	NBT with Batch Normalization	96
3	NBT with Cosine Normalization	94
4	NBT with Group Normalization	94
5	NBT with Layer Normalization	93
6	NBT with Instance Normalization	94

Table 2 indicates the F–score for the models with normalization and without normalization. The highest F–score is obtained by model batch normalization. The least F–score is obtained by layer normalization.

Table 2. F–score

Sr.No	Method	F–score
1	NBT without Normalization	95.5
2	NBT with Batch Normalization	98.3
3	NBT with Cosine Normalization	97.4
4	NBT with Group Normalization	97.3
5	NBT with Layer Normalization	96.7
6	NBT with Instance Normalization	97.2

Table 3 indicates the time required for the models with normalization and without normalization. normalization. The table indicates the training time is reduced for a model with normalization. The training time required for normalization with cos normalization is the least. The maximum training time was required for most models with instance normalization.

Table 3. Time for training

Sr.No	Method	Time
1	NBT without Normalization	2645
2	NBT with Batch Normalization	2232
3	NBT with Cosine Normalization	2048
4	NBT with Group Normalization	2520
5	NBT with Layer Normalization	2607
6	NBT with Instance Normalization	2613

Table 4 indicates the detection rate for the models with normalization and without normalization. The detection rate of the model is improved by normalization. The highest detection rate is obtained by the model with batch normalization and cos normalization. The least detection rate is obtained by layer normalization.

Table 4. Detection Rate

Sr.No	Method	Detection rate
1	NBT without Normalization	91.38
2	NBT with Batch Normalization	96.8
3	NBT with Cosine Normalization	96.8
4	NBT with Group Normalization	95
5	NBT with Layer Normalization	93.8
6	NBT with Instance Normalization	94.5

4.4.2 Results for Price range Slot:

Figure 3 displays the learning curve for the Price range slot. The graph shows an improvement in the model with normalization. The highest improvement was observed with the model with group and batch normalization and group normalization. The learning curve for batch normalization is smooth and increasing. This displays a good improvement in model performance. The first observation is there is an improvement in model performance with normalization. The learning curve for the group and instance normalization has displayed the least variation. The next observation is that model with normalization has better generalization. In the case of minimum training data, good performance is displayed by instance normalization. In the case of maximum training data, good performance is displayed by group and batch normalization.

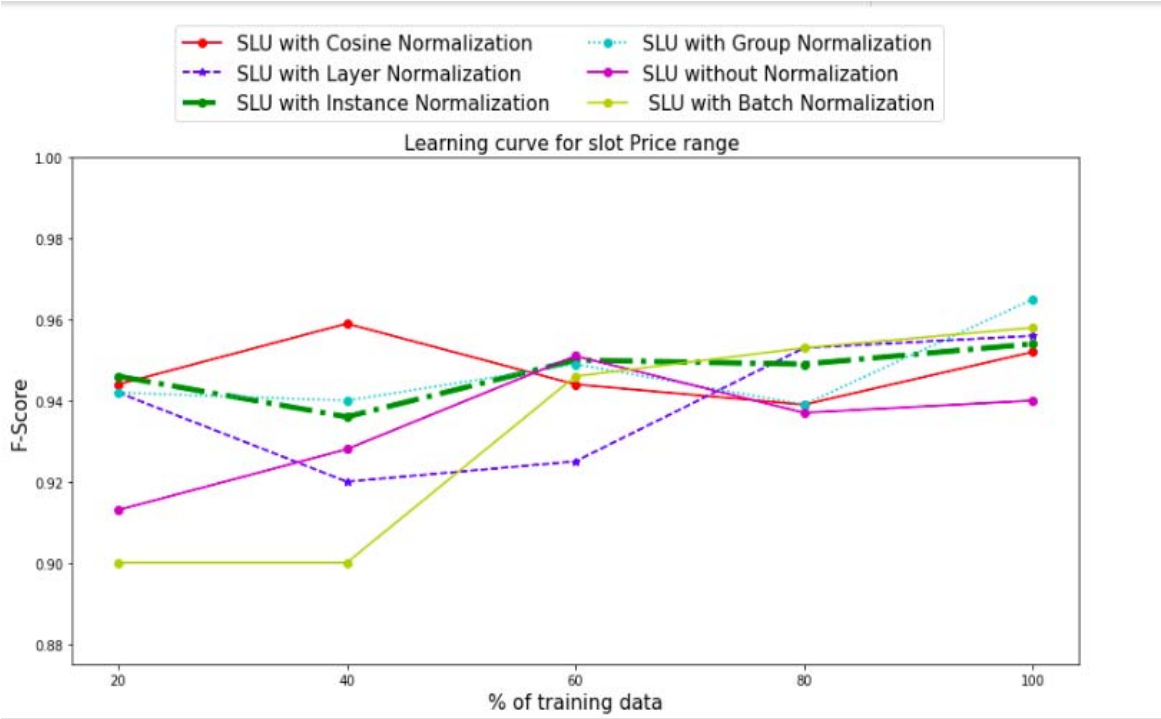

Figure. 3 F–score learning curve for the Price rage slot

Table 5 indicates the accuracy for the slot price range with normalization and without normalization. The results show that accuracy was improved by normalization. The accuracy was highest for batch normalization. The least accuracy was for layer normalization.

Table 5. Accuracy

Sr.No	Method	Accuracy
1	NBT without Normalization	91.8
2	NBT with Batch Normalization	93.8
3	NBT with Cosine Normalization	93
4	NBT with Group Normalization	93.1
5	NBT with Layer Normalization	93
6	NBT with Instance Normalization	93.2

Table 6 shows the comparative study of model performance incorporating different normalization. The maximum f–score is obtained by group normalization and batch normalization. The least f–score is obtained for cosine normalization.

Table 6. F–score

Sr.No	Method	F–score
1	NBT without Normalization	94
2	NBT with Batch Normalization	95.8
3	NBT with Cosine Normalization	95.2
4	NBT with Group Normalization	96.5
5	NBT with Layer Normalization	95.6
6	NBT with Instance Normalization	95.4

Table 7 shows the impact of normalization on training time. Training time is reduced by normalization. The least training time is obtained by batch normalization and maximum training is required for training with Instance normalization.

Table 7. Time for training

Sr.No	Method	Time
1	NBT without Normalization	2536
2	NBT with Batch Normalization	2172
3	NBT with Cosine Normalization	2215
4	NBT with Group Normalization	2183
5	NBT with Layer Normalization	2216
6	NBT with Instance Normalization	2289

Table 8 shows the impact of normalization on the detection rate. The detection rate is improved by normalization. The detection rate for batch normalization is maximum and minimum for layer normalization.

Table 8. Detection Rate

Sr.No	Method	Detection rate
1	NBT without Normalization	93.5
2	NBT with Batch Normalization	95.8
3	NBT with Cosine Normalization	95.2
4	NBT with Group Normalization	95.1
5	NBT with Layer Normalization	89.6
6	NBT with Instance Normalization	95.7

5. Conclusion:

SLU takes part in a significant role in understanding user objectives. To understand user goals, SLU utilizes deep learning models from machine learning. These machine learning models are affected by the internal covariant shift. There are techniques in machine learning to reduce the internal covariant shift and improve model performance. There is a significant improvement in model performance by applying these techniques.

The work in this paper extends ICS reducing normalization techniques from machine learning to deep models in SLU. Overall performance of SLU was improved by incorporating ICS reducing techniques. The results were obtained for the area and price range slot. The results were similar for both slots. The F–score learning graph for both the indicates there is an improvement in SLU performance. The least change in value of F–score for varying training data was reported with the model with instance normalization for both the slots. The next observation from the learning curve of both slots is the generalization ability of the model is improved with normalization. At minimum training data for both slots best results were obtained by instance normalization and for maximum training, data was obtained by batch normalization and group normalization. In experiments conducted for both slots, the detection rate was maximum for the model incorporating batch normalization and minimum for layer normalization. The results show that reducing covariant shifts in SLU results in enhanced learning and improvement in performance.

References

[1] Peng, N., Dredze, M., and Processing, S., 2016, “Multi-Task Multi-Domain Representation Learning for Sequence Tagging.” arxiv preprocessing 2016(online).

[2] Hakkani-T, D., Tur, G., Celikyilmaz, A., Chen, Y. N., Gao, J., Deng, L., and Wang, Y. Y., 2016, “Multi-Domain Joint Semantic Frame Parsing Using Bi-Directional RNN-LSTM,” Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 715–719

[3] Crook, P. A., Marin, A., Agarwal, V., Aggarwal, K., Anastasakos, T., Bikkula, R., Boies, D., Celikyilmaz, A., Chandramohan, S., Feizollahi, Z., Holenstein, R., Jeong, M., Khan, O. Z., Kim, Y., Krawczyk, E., Liu, X., Panic, D., Radostev, V., Ramesh, N., Robichaud, J., Rochette, A., Stromberg, L., Sarikaya, R., Corporation, M., and Way, O. M., 2016, “Task Completion Platform : A Self-Serve Multi-Domain Goal Oriented Dialogue Platform,” Naacl-Hlt-2016, 2016, pp. 47–51.

[4] Liu, X., Celikyilmaz, A., Sarikaya, R., and Corporation, M., “Natural Language Understanding for Partial Queries.” IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU).

[5] Jagdale, S., & Shah, M. (2019). Evaluation of Stochastic Gradient Descent and Spoken Language Understanding. 2019 International Conference on Nascent Technologies in Engineering (ICNTE), Icnte, 1–4

[6] Jagdale, S., & Shah, M, “Extending the Classifier Algorithms in Machine Learning to Improve the Performance in Spoken Language Understanding Systems Under Deficient Training Data”,Volume 5, Issue 6, Page No 464-471, 2020

[7] Iman, A. N., & Ahmad, T. (2020). Data Reduction for Optimizing Feature Selection in Modeling Intrusion Detection System. International Journal of Intelligent Engineering and Systems, 13(6), 199–207.

[8] Pullaiah, N., Venkatasekhar, D., Venkatramana, P., & Sudhakar, B. (2020). Detection of Breast Cancer on Magnetic Resonance Imaging Using Hybrid Feature Extraction and Deep Neural Network Techniques. International Journal of Intelligent Engineering and Systems, 13(6), 229–240.

[9] Kusuma, E. J., Shidik, G. F., & Pramunendar, R. A. (2020). Optimization of Neural Network using Nelder Mead in Breast Cancer Classification. International Journal of Intelligent Engineering and Systems, 13(6), 330–337.
[10] Fibriani, I., Widjonarko, W., Prasetyo, A., Raharjo, A., & Irawan, D. (2020). Multi Deep Learning to Diagnose COVID-19 in Lung X-Ray Images with Majority Vote Technique. International Journal of Intelligent Engineering and Systems, 13(6), 560–568.
[11] Talasila, V., Madhubabu, K., Mahadasyam, M. C., Atchala, N. J., & Kande, L. S. (2020). The prediction of diseases using rough set theory with recurrent neural network in big data analytics. International Journal of Intelligent Engineering and Systems, 13(5), 10–18.
[12] Annamalai, P. (2020). Automatic face recognition using enhanced firefly optimization algorithm and deep belief network. International Journal of Intelligent Engineering and Systems, 13(5), 19–28.
[13] Sumpavakup, C., Mongkoldee, K., Ratniyomchai, T., & Kulworawanichpong, T. (2020). Touch and step voltage evaluation based on computer simulation for a mass rapid transit system in Thailand. International Journal of Intelligent Engineering and Systems, 13(5), 159–169.
[14] Jagdale, S., & Shah, M,R. Ruskone, "Investigating batch normalization in Spoken Language Understanding", In: Proc. of International Conf. On International Conference on Electronics, Communications and Information Technolog, Shenzhen, China , 2020.
[15] Kim, J., Tur, G., Celikyilmaz, A., Cao, B., and Wang, Y., "Intent detection using semantically enriched word embeddings ," IEEE Workshop on spoken dialogue technology (California).
[16] Ji, Y., Celikyilmaz, A., and Heck, L., 2014, "A Variation Bayesian model for user intent detection ,"School of Interactive Computing , Georgia Institute of Technology," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),pp. 4072–4076.
[17] Zhang, J., Yang, T. Z., and Hazen, T. J., 2015, "Large-Scaleword Representation Features for Improved Spoken Language Understanding," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., 2015–Augus, pp. 5306–5310.
[18] Celikyilmaz, A., Sarikaya, R., Hakkani-tur, D., Liu, X., Ramesh, N., and Tur, G., 2016, "A New Pre-Training Method for Training Deep Learning Models with Application to Spoken Language Understanding." Proceedings of The 17th Annual Meeting of the International Speech Communication Association ,INTERSPEECH
[19] Chen, Y. N., Hakkani-Tur, D., & He, X. (2016). Zero-shot learning of intent embeddings for expansion by convolutional deep structured semantic models. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2016-May(January), 6045–6049. https://doi.org/10.1109/ICASSP.2016.7472838
[20] Bhargava, A., Celikyilmaz, A., and Sarikaya, R., 2012, "Easy contextual intent prediction and slot detection," Microsoft Research."
[21] Cohen, K. B., Varile, G., Zampolli, A., Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., & Zue, V. (2000). Survey of the State of the Art in Human Language Technology. Language, 76(1), 214. Cohen, K. B., Varile, G., Zampolli, A., Cole, R., Mariani, J., Uszkoreit, H., Zaenen, A., & Zue, V. (2000). Survey of the State of the Art in Human Language Technology. Language, 76(1), 214. https://doi.org/10.2307/417436
[22] Peng, N., Dredze, M., Hakkani-tur, D., 2016, "Syntax or Semantics? Knowledge-Guided Joint Semantic Frame Parsing." in Proceedings of 6th IEEE Workshop on Spoken Language Technology.
[23] Wang, P., Xu, B., Xu, J., Tian, G., Liu, C. L., & Hao, H. (2016). Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. Neurocomputing, 174, 806–814.
[24] Celikyilmaz, A., 2010, "Convolutional Neural Network Based Semantic Tagging with Entity Embeddings," Microsoft research , pp. 1–7.
[25] Duan, J., Zhang, R., Huang, J., & Zhu, Q. (2018). The Speed Improvement by Merging Batch Normalization into Previously Linear Layer in CNN. ICALIP 2018 6th International Conference on Audio, Language and Image Processing, 67–72.
[26] Luo, C., Zhan, J., Xue, X., Wang, L., Ren, R., & Yang, Q. (2018). Cosine normalization: Using cosine similarity instead of dot product in neural networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11139 LNCS, 382–391.
[27] Luo, P., Ren, J., Peng, Z., Zhang, R., & Li, J. (2018). Differentiable learning-to-normalize via switchable normalization. ArXiv, 1–19
[28] Luo, Phang, R., Ren, J., Peng, Z., & Li, J. (2019). Switchable normalization for learning-to-normalize deep representation. ArXiv,
[29] Shao, W., Li, J., Ren, J., Zhang, R., Wang, X., & Luo, P. (2020). SSN: Learning Sparse Switchable Normalization via SparsestMax. International Journal of Computer Vision, 128(8–9), 2107–2125
[30] Wu, Y., & He, K. (2020). Group Normalization. International Journal of Computer Vision, 128(3), 742–755. https://doi.org/10.1007/s11263-019-01198-w
[31] Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance Normalization: The Missing Ingredient for Fast Stylization. 2016.
[32] Huang, X., & Belongie, S. (2017). Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization. Proceedings of the IEEE International Conference on Computer Vision, 2017-October, 1510–1519.
[33] Mrkšić, N., & Vulić, I. (2018). Fully Statistical Neural Belief Tracking. 108–113.
[34] Duan, J., Zhang, R., Huang, J., & Zhu, Q. (2018). The Speed Improvement by Merging Batch Normalization into Previously Linear Layer in CNN. ICALIP 2018 - 6th International Conference on Audio, Language and Image Processing, 67–72. https://doi.org/10.1109/ICALIP.2018.8455587