

# TOWARDS A NEW DEEP LEARNING ALGORITHM BASED ON GRU AND CNN: NGRU

Submitted: 26<sup>th</sup> June 2019; accepted: 25<sup>th</sup> March 2020

Abdelhamid Atassi, Ikram el Azami

DOI: 10.14313/JAMRIS/4-2020/44

**Abstract:** This paper describes our new deep learning system based on a comparison between GRU and CNN. Initially we start with the first system which uses Convolutional Neural Network (CNN) which we will compare with the second system which uses Gated Recurrent Unit (GRU). And through this comparison we propose a new system based on the positive points of the two previous systems. Therefore, this new system will take the right choice of hyper-parameters recommended by the authors of both systems. At the final stage we propose a method to apply this new system to the dataset of different languages (used especially in social networks).

**Keywords:** Convolutional Neural Network, CNN, Gated Recurrent Unit, GRU, SemEval, Twitter, word2vec, Keras, TensorFlow, Adadelta, Adam, soft-max, deep learning

## 1. Introduction

The vast use of the micro blog such as Twitter with more than 500 million tweets per day, make the sentiment analysis the one way to deal e-reputation and many problem should be taken into account, namely: 1) informal language used by the users; 2) users can switch between languages; 3) users can use multiple languages in one word or one sentence; 4) emoticons; 5) hashtags; 6) usernames used to call or notify other users. 7) URL; 8) Image; 9) Video. So this merger requires pre-treatment to normalize the text before starting the learning for both systems, the first use CNN model [1, 2] and the second use GRU model [3] for that we compare the two approaches to make sentiment analysis and we propose a new method "New GRU" (NGRU) to more efficient system applies to Arabic tweets.

The introduction of the Arabic language in SemEval-2017, because of the large number of tweets in recent years, but in our case we have the Moroccan dialectal and not the official Arabic language, which leads us to collect and assign polarity manually.

## 2. Data Preparation

Both systems use word embeddings but with difference, in first CNN system's [4], so the word embeddings with 50M unsupervised to training word2vec [5,

6], and 10M supervised to train CNN, but in the second, the GRU with the first word embeddings with 20.5M to train and the second set was obtained by training on supervised data using another GRU model, and add method for splitting hashtags and insert them in the body, before forwarding data in this step [7]. Therefore the amount of training data is very important without forgetting the quality with the treatment of tweets before the passes to the learning phase, CNN has taken a step forward with the amount of data on the other hand there is an equality with the pre-treatment of the text e.g. suppression of the words without value, and the treatment of the hashtags with tokenizing.

Fig. 1 shows the architecture of our deep learning model, our combined method the last two methods CNN and GRU, and we add three types of treatment on input tweets. Manual sentiment analysis, the user assigns the polarization manually to the data. Emoticons treatment, we apply a treatment on tweets contains emoticons for affect polarities, eg. :) , :-), :=) → Happy emotions (+), :( , :-), :=( → Sad emotions (-). The third type takes the results of the last two types and adds new tweets, to affect polarity by similarity.

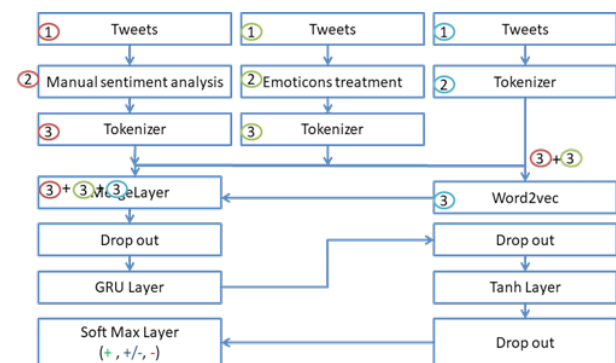


Fig. 1. The architecture of the NGRU deep Learning model

To collect more than 20k of tweet from different languages we need a lot of time and to assign polarity to these tweets we need more time. This is why we do not have the data today, but they are in preparation.

## 3. The Architecture of the GRU, CNN and NGRU Deep Learning Models

The GRU layer is the core of the GRU network, it is more computational efficient than CNN models [8]

Which it can capture long semantic patterns without tuning the model parameter, unlike CNN models where the model depends on the length of the convolutional feature maps for capturing long patterns, Which it achieved superior performance to CNNs [7]. And the network architecture is composed of a word embeddings layer, a merge layer, dropout layers, a GRU layer, a hyperbolic tangent tang layer and a soft-max classification layer.

The architecture of the CNN, it is fundamentally inspired by the architectures used in [1, 2] for performing various sentence classification tasks. This architecture requires large corpus. Therefore, different from [2] that presents an architecture with several layers of convolutional feature maps, as for the author [4] adopt a single level architecture, who has been shown in [1] to perform equally well.

Our architecture Fig. 1 is inspired by GRU [7], with changes in the input data, we use the method Adadel-ta to the learning rate, and the method Adam for optimization of the weights and the application of NGRU will be on the tweets from different languages.

#### 4. Network Parameters and Training

Before seeing the weights starts with the hyper parameters on the first position we find the CNN, the author use stochastic gradient descent to train the network and use back propagation algorithm to compute the gradients and then use Adadel-ta [9] to automatically tune learning rate, as for the convolutional feature map at 300, in the second system GRU use Adam [10] its new and computationally efficient for optimizing weights, for that all the experiments have been developed using Keras With the dimension of word embedding at 100.

We use TensorFlow, an open source software library, used for research and production at Google, we used to implement our system as soon as possible and to focus on architecture and performance tuning and parameter optimization.

The aim is not to implement both methods but to use it, because they already implement at the Tensor-Flow framework, by [9] Adadel-ta does not use a manual adjustment of a rate of learning and insensitive to hyperparameters, it's for adjust the hyperparameters without the intervention of the user.

Even though [11] demonstrates on his "An overview of gradient descent optimization", that the different gradient descent algorithms are the same, Adadel-ta proves that it is faster than SGD, Momentum, NAG, Adagrad and RMSProp, see Fig. 2.

Regarding the Adam (Adaptive Moment Estimation) method, we used this method to optimize weights, it combines the benefits of momentum with the benefits of RMSProp, and it differs mainly in two ways. First, the order moment moving average coefficient is decayed over time. Second, because the first and second order moment estimates are initialized to zero, some bias-correction is used to counteract the resulting bias towards zero.

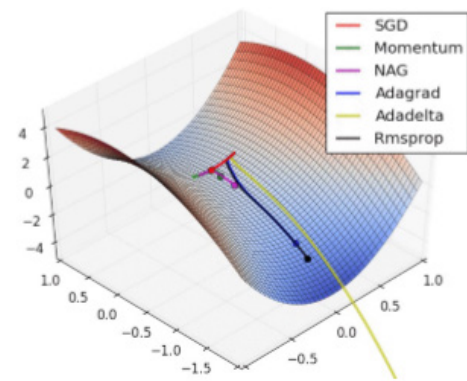


Fig. 2. A comparison of the different methods, which shows the speed of the Adadel-ta method

#### 5. Initializing the Model Parameters

The CNN system will be initialized from word embeddings already trained with unsupervised data, and the GRU system will be trained with two word embeddings, the first with unsupervised data and the second obtained from another GRU.

The NGRU requires three data sources with the use of both methods Adadel-ta to the learning rate, and the method Adam for optimization of the weights.

#### 6. Evaluation

The CNN system offers very best performance results in SemEval-2015 because he classification and demonstrates the amount of data the most important at adjustments at the level of hyper parameters and pre-treatment before starting training.

The GRU system is the newest type of Deep Learning, despite these results are not at the hiring point because the results were weak and the hyper parameter was insufficient, Even if we used the Adam method for optimizing weights.

Our model requires in a part a manual treatment, for this reason we have not yet results.

#### 7. Conclusion

We propose to use the best points of the two systems, with a feeding a large data for training our new model, we use the method Adadel-ta to the learning rate, and the method Adam for optimization of the weights and the application of this proposal will be on the tweets contains the mix of the Modern Standard Arabic Language and Dialectal Arabic Language.

#### AUTHORS

Abdelhamid Atassi\* – LaRI Laboratory, Faculty of Sciences, Ibn Tofail University, Kénitra, Morocco, e-mail: abdelhamid.atassi@gmail.com.

**Ikram el Azami** – LaRI Laboratory, Faculty of Sciences, Ibn Tofail University, Kénitra, Morocco, e-mail: akram\_elazami@yahoo.fr.

\*Corresponding author

---

## REFERENCES

- [1] Y. Kim, “Convolutional Neural Networks for Sentence Classification”, <http://arxiv.org/abs/1408.5882>. Accessed on: 2021-02-03.
- [2] N. Kalchbrenner, E. Grefenstette and P. Blunsom, “A Convolutional Neural Network for Modelling Sentences”. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 2014, 655–665, DOI: 10.3115/v1/P14-1062.
- [3] D. Bahdanau, K. Cho and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate”, <http://arxiv.org/abs/1409.0473>. Accessed on: 2021-02-03.
- [4] A. Severyn and A. Moschitti, “UNITN: Training Deep Convolutional Neural Network for Twitter Sentiment Classification”. In: *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, 464–469, DOI: 10.18653/v1/S15-2079.
- [5] R. Collobert and J. Weston, “A unified architecture for natural language processing: deep neural networks with multitask learning”. In: *Proceedings of the 25th international conference on Machine learning*, 2008, 160–167, DOI: 10.1145/1390156.1390177.
- [6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality”, *Advances in Neural Information Processing Systems*, vol. 26, 2013, 3111–3119.
- [7] M. Nabil, A. Atyia and M. Aly, “CUFE at SemEval-2016 Task 4: A Gated Recurrent Model for Sentiment Classification”. In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, 2016, 52–57, DOI: 10.18653/v1/S16-1005.
- [8] S. Lai, L. Xu, K. Liu and J. Zhao, “Recurrent convolutional neural networks for text classification”. In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, 2267–2273.
- [9] M. D. Zeiler, “ADADELTA: An Adaptive Learning Rate Method”, <http://arxiv.org/abs/1212.5701>. Accessed on: 2021-02-03.
- [10] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization”, <http://arxiv.org/abs/1412.6980>. Accessed on: 2021-02-03.
- [11] S. Ruder, “An overview of gradient descent optimization algorithms”. <http://arxiv.org/abs/1609.04747>. Accessed on: 2021-02-03.