A Study on the Influence of Architecture Complexity of RNNs for Intent Classification in E-Commerce Chats in Bahasa Indonesia

Renny Pradina Kusumawardani
Department of Information Systems
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
renny.pradina@gmail.com

Muhammad Azzam
Department of Information Systems
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
muhammadazzam1602@gmail.com

Abstract

We present our work in the intent classification of chat utterances. We use several recurrent neural network (RNN) architectures of different complexity levels; basic RNN, GRU, LSTM, and BiLSTM. Experiments are performed on e-commerce smartphone sales chat in Bahasa Indonesia. We found that GRU gives the best performance, with 87.10% accuracy and 86.67% F1-measure. In comparison to the other architectures, GRU is also fast to train.

1 Introduction

Despite its sequential nature, RNN has been shown to perform well on classification tasks (Tang et al., 2015; Yin et al., 2017). In this paper, we explore the use of several architectures of RNNs: GRU (Cho et al., 2014), LSTM (Hochreiter and Schmidhuber, 1997), and BiLSTM (Graves and Schmidhuber, 2005) for the classification of chat intent in a smartphone online sale in Bahasa Indonesia. Indonesia is the highest growing market for online shopping with 78% per year (PPRO Group, 2018), while chat service is the service most used by Indonesian internet users (89.35%) (APJII, 2017). As Indonesia is the 4th most populous country in the world, this presents an economic potential for the use of automated chat services in e-commerce. However, in most cases, computing resources are limited. It is, therefore, necessary to take complexity into account when deciding on an architecture on which an intent classification model is to be trained.

2 Method

We tagged 1,806 customer utterances based on their intents into sixteen classes with the following labels: availability, price, delivery, specification, after-sales, greeting, payment, type, features, condition, accessory, originality, store, transaction, and return. These utterances are part of dialogues generated in the style of conversations that appear in Indonesian popular e-commerce sites, e.g., Tokopedia and Bukalapak. For each product, a small number of such conversation is available on the product page as discussions, since customers usually contact sellers in private chats. This tagged data will be made available with the publication of the paper of this extended abstract. Table 1 shows a sample of the data.

Original Text	English Translation	Intent
Misi gan, hp Oppo F1 ready?	Excuse me, is Oppo F1 in stock?	Availability
yang silver metallic ada gan?	Do you have the metallic silver variant?	Availability
garansi apa mas?	What type is the guarantee?	After-sales

Table 1: Sample chat data.

The data set was split randomly into train, dev, and test sets with a ratio of 80:10:10. We maintained the proportion of each class in the splits. Table 2 shows the distribution of utterances of each class in these splits, ordered based on frequency.

	Train Size	Dev Size	Test Size
Availability	390	49	49
Price	174	22	22
Delivery	167	21	21
Specification	144	18	19
After-sales	127	16	16
Greetings	93	12	12
Payment	66	8	9
Type	62	8	8
Features	61	8	8
Condition	52	6	7
Accessory	30	4	4
Originality	22	3	3
Store	20	2	3
Transaction	17	2	3
Return	14	2	2
TOTAL	1439	181	186

Table 2: Data distribution of each class for the train, dev, and test sets.

We varied the parameters optimizer (SGD, Momentum, ADAM), learning rate (0.001, 0.1, and 0.015), and the number of layers (1, 2, 3). We used 150 epochs to train the model, saving the model with the best accuracy on the development set. We repeated the experiments five times and reported the average test performance. The pre-trained embeddings used was by Afnandika and Kusumawardani (2018), trained using Word2Vec (Mikolov et al., 2013) on 62,834,464 tweets in Bahasa Indonesia. These tweets have a similar level of language informality to the text in this work.

3 Result

Table 3 shows the best accuracy and F1-score of each architecture and the parameters at which they occur. GRU gave the highest accuracy of 87.78% and F1-score of 86.67%. Additionally, this result was achieved using the ADAM optimizer, with a learning rate of 0.001.

Arch. Type	Optimizer	Num. Layer	Learn. Rate	Acc. (%)	F1-score (%)	Time (s)
RNN	SGD	3	0.001	78.46	76.83	1787
GRU	ADAM	1	0.001	87.78	86.67	1581
LSTM	ADAM	1	0.001	85.40	83.44	1546
BiLSTM	Momentum	1	0.01	84.56	83.21	1836

Table 3: The best model for each RNN architectures.

The last column of Table 3 shows the time it took to train 150 epochs of each setting, within which we can reasonably expect the model to have converged, using an Nvidia GeForce GTX1060. With 1581 seconds, a single-layer GRU is relatively fast to train, taking only slightly longer than LSTM and significantly faster than RNNs, which requires three layers. Table 4 gives the detailed measurement of training times, showing that for our intent classification task, using Bi-LSTM, which is usually the favored architecture, could lead to training times of twice as long as taken by the other architectures. The brevity of the utterances, and thus the lack of long-range dependencies, might be the reason why GRU works better in this case than a bidirectional architecture such as Bi-LSTM. These results highlight the importance of the wise selection of model architecture, especially when computing resources are limited.

	Num. Layer	SGD			Momentum			ADAM		
Learning Rate		**0.001**	**0.01**	**0.015**	**0.001**	**0.01**	**0.015**	**0.001**	**0.01**	**0.015**
RNN	1	1140	1114	1116	1255	1247	1248	1561	1560	1539
	2	1447	1444	1440	1679	1675	1676	2176	2178	2144
	3	1787	1793	1792	2141	2125	2131	2820	2823	2783
GRU	1	1139	1138	1139	1269	1264	1267	1581	1580	1579
	2	1501	1500	1503	1737	1736	1740	2229	2227	2230
	3	1881	1880	1903	2216	2217	2219	2897	2898	2893
LSTM	1	1173	1171	1171	1279	1279	1287	1546	1556	1544
	2	1548	1549	1550	1765	1767	1769	2217	2222	2213
	3	1960	1963	1958	2270	2271	2274	2900	2901	2890
Bi-LSTM	1	1612	1602	1598	1836	1837	1842	2278	2280	2275
	2	2419	2425	2422	2833	2834	2837	3660	3720	3659
	3	3211	3209	3211	3786	3782	3793	5040	4980	5012

Table 4: The time for training models for each setting (in seconds).

4 Conclusion

In this extended abstract, we explore the use of RNN architectures, to find the best architecture given the task of intent classification of e-commerce smartphone sales chat. We find that the single-layer GRU, which is of medium-low complexity when compared to the other architectures explored, gives the best accuracy. This finding signifies that models of higher complexity are not necessarily the best, a result that is important to note when the computing resources available are limited.

Reference

Adrian Afnandika & Renny Pradina Kusumawardani. 2018. Sentiment analysis for social media text in Bahasa Indonesia using convolutional neural networks (study case: telecommunication operator). Undergraduate thesis, Institut Teknologi Sepuluh Nopember (ITS). Surabaya, Indonesia.

Asosiasi Penyelenggara Jasa Internet Indonesia. 2017. *Penetration and Behavior of Internet Users in Indonesia*.

Kyunghun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *The Eighth Workshop on Syntax, Semantics, and Structure in Statistical Translation*.

Alex Graves and Jürgen Schmidhuber, 2003. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks, 18(5-6),* pages 602-610.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735-1780.

Thomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *ICLR 2013.*

PPRO Group. 2018. *Payments and e-commerce report: high-growth markets*.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of EMNLP*, pages 1422-1432.

Wepeng Yin, Katharina Kann, Mo Yu, and Hinrich Schütze. 2017. Comparative Study of CNN and RNN for Natural Language Processing. arXiv:1702.01923 [cs].