Prediction of Werewolf Players by Sentiment Analysis of Game Dialogue in Japanese

Yingxue Sun[2,a)] Tomoyuki Kaneko[2,b)]

Abstract: The werewolf game is a communication game about trust and deception that is recently trending worldwide, and researchers have studied how to train AI agents to play this game. Compared to AI agents in other games, AI agents in the werewolf game need to speculate the intention of other players' through their conversations in natural language in addition to other discrete actions. There are different viewpoints for agents in the werewolf game because agents can be assigned to different roles, leading to different information sets. This paper starts from a villager player's viewpoint, and tries to analyze the only public information for them —— dialog using sentiment points, coming out information, and voting points. We implemented our models with Gated Recurrent Unit (GRU), and three tasks are evaluated via these data sets: werewolf prediction, prediction using CO information, and voting prediction. All of the tasks get accuracy better than the random model.

Keywords: Imperfect Information Games, The Werewolf Game, Sentiment Analysis, Gated Recurrent Unit

1. Introduction

After the impressive match of Go between Chinook and Deep Blue [4], game AI is suddenly well-known by the world. From the 20 century, many algorithms like Monte Carlo Tree Search [3] and Counterfactual Regret Minimization [13] are invented to create strong agents that could defeat even professional human players. Compared to perfect information games like Chess or Go, imperfect information games like poker are explored in different ways because players always receive asymmetric information. The werewolf game is one of those imperfect information games, where the communication between players is the only public information that could be truth or lies. Compared to other imperfect information games, the werewolf game needs the agents to understand their communication and deduce the werewolves that are hidden among the villagers which makes it even more challenging to design AI agents for this game. Most researches on the werewolf game can be divided into two categories, one is designing agents with a higher winning rate based on a protocol platform [10] which simplified the communication part but focuses more on the optimization of strategies, the other one focuses on studying the natural languages in the werewolf games, for instance, classification of the dialog [6].

Sentiment analysis is a newly trending research field that uses natural language processing, test analysis, linguistics to understand the sentiment conveyed in human languages. It is wildly used to analyze reviews, ratings, opinions written by people. Intuitively, the sentiment analysis can be used in studying the dialog in the werewolf game, and we are interested in how sentiment analysis can help the agents to distinguish the werewolves from the villagers. We conduct three tasks using sentiment analysis on the werewolf game dialog written by human players and deduce the werewolf on this basis. We give selected sentences sentiment points in time series and use Gate Recurrent Unit (GRU) model to learn the patterns.

2. Background

In order to predict werewolves from the dialog, we need the full game logs of the werewolf games played by real human players which should include dialog, character assignments, results of the game, etc. We choose *Werewolf BBS* [1] where 2000 game logs are recorded. Game logs on Werewolf BBS are written in Japanese and also used in researches [6], [9].

2.1 *The Werewolf Game*

The werewolf Game is a multiplayer, verbally driven strategy game that competes with persuasion skills and logical analysis ability which often involves four to eighteen players. In the beginning, players are divided into two opposing teams——villagers and werewolves. Werewolves know the identity of each other while villagers only know their own identity. Werewolves can then take their advantage of information to deceive villagers, and villagers need to analyze deliberately on other players' conversation to find out the players who are lying. The game lasts for several "days"(days in the game) until all the players of either the villager side or werewolves are out. On each "day", there is a daytime stage and a night stage. In the daytime, there

[1] Graduate School of Arts and Sciences, the University of Tokyo
[2] Graduate School of Interdisciplinary Information Studies, the University of Tokyo
[a)] yx-sun2020@g.ecc.u-tokyo.ac.jp
[b)] kaneko@acm.org

will be a conversation session at first where everyone can state their opinion. At the end of the daytime stage, there is a vote session where players can vote out the player they think is most suspicious during the conversation session (It's possible to vote out a villager side player). At night werewolves can kill a villager side player. Some players may be assigned special roles with special abilities. They will use their special abilities in the night stage. The winning condition for the villager side is to vote out all werewolves, while for the werewolf side, the winning condition is to eliminate all villager side players by voting in the daytime or killing at night. Table 1 shows the special roles used in werewolf BBS and their own abilities. The action of the seer and the spirit medium checking a persons' side is called *divine.* Although the werewolf game is played worldwide, players in different game groups only share these basic rules, sometimes the detailed rules are different. In this research, we adopt the rules specified in Werewolf BBS.

- There are several *villagers*, a *seer* and a *spirit medium* for each game. When there are fewer than 12 players, there will be 2 *werewolves*, otherwise, 3 *werewolves.* If there are over 11 players, there will be a *guardian* for the human side and a *possessed* for the werewolf side.
- Although possessed is in the werewolf side, the player doesn't know who are werewolves, also, werewolves don't know who are the possessed.
- Werewolves can chat during daytime in a private channel that only werewolves can see.
- The player killed by werewolves in the first night is set to be the designated AI player who only speaks once at the very beginning of the game, and the vote session will start on the second day, which means all human players can survive from the first night and the first voting session start on the second day.
- The information "Who votes for whom" won't be open public. Players only get the result of the players who received votes with the number of votes in the next morning.

Roles	Side	Abilities
Seer	Villager side	Check whether a player is a werewolf or not every night
Spirit Medium	Villager side	Check whether a dead player is a werewolf or not every night
Guardian	Villager side	Protect a player from being attacked by werewolf every night
Possessed	Werewolf side	Can't be recognized as werewolf while checked by Seer or Medium

Table 1 Special Roles and Abilities

Although some of the players have special abilities owning to their role to get information, they don't necessarily earn the trust of other players. Werewolf side players will use lies to confuse normal villager players. But sometimes their support for each other is too obvious that arouses suspicion from players on the villager side and they ended get voted out. Both sides of werewolves and villagers need to be careful about their speeches to avoid being eliminated during the daytime, also they need to analyze the conversation so that villagers can deduct the werewolves, and werewolves can deduct the players with special abilities and kill them at night. Hence we focus on the key point to win this game, that is, the conversation session in the daytime.

2.2 *Werewolf BBS*

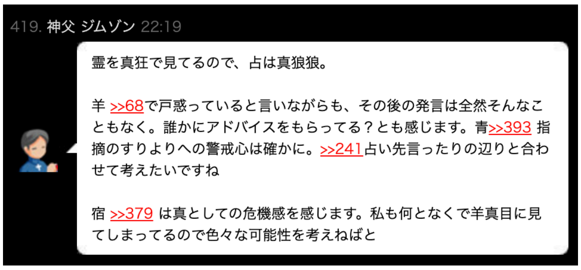
Our corpus is from Werewolf BBS, where the werewolf game lovers were able to play werewolf games online. The time in a single game is consistent with the time in the real world, which means a day in the werewolf game is a day in real life, so the communication is sufficient and in time series. Before the first day start, every player is assigned a role. To avoid players being eliminated too early so that they got a bad game experience, a computer player will die on the first day and there is no voting session on the first day either. In this paper, we use the dialog only from the first two days when every player is alive.

2.3 Gated Recurrent Unit

As the sentiment points are in time series, we choose Gated Recurrent Unit (GRU) which is an effective variant of long short-term memory (LSTM) and introduced in 2014 by Kyunghyun et al [5]. LSTM and GRU are RNN networks that can process series data. Another reason that we choose between GRU or LSTM is that the series of sentiment points between the two players are in different lengths which is hard for other models to handle. LSTM is formed of three gates: input gate, forgotten gate, output gate, and a memory cell while GRU combined forgotten gate and input gate into an update gate，merge cell status with hidden status and set up a new reset gate, which makes GRU much simpler than LSTM, and reach similar accuracy with less training time. So we choose GRU to process our sentiment points. The update gate is used to control the extent to which the state information from the previous moment is brought into the current state. The larger the value of the update gate, the more the state information from the previous moment is brought in. The reset gate controls how much information from the previous state is written to the current candidate set. The smaller the reset gate, the less information from the previous state is written.

3. Related Works

Recent studies on werewolf agents are mostly based on protocol platforms, for example, using protocol game logs to predict agent votes and werewolves [7]. LSTM model is used in this research to analyze the game logs in time series. Another research applies deep Q learning to create the werewolf game agents [11]. This research focuses more on the improvement of strategies and winning ratio against



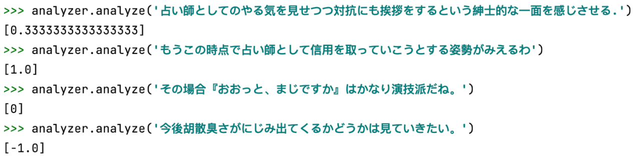
419. 神父 ジムゾン 22:19

霊を真狂で見てるので、占は真狼狼。

羊 >>68で戸惑っていると言いながらも、その後の発言は全然そんなこ
ともなく。誰かにアドバイスをもらってる？とも感じます。青>>393 指
摘のすりよりへの警戒心は確かに。>>241占い先言ったりの辺りと合わ
せて考えたいですね

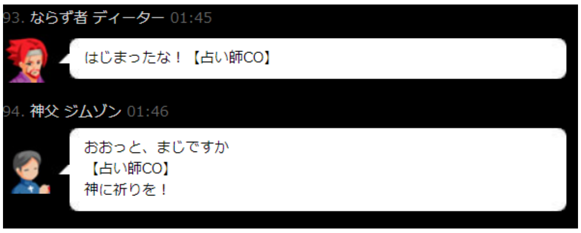
宿 >>379 は真としての危機感を感じます。私も何となくで羊真目に見
てしまってるので色々な可能性を考えねばと

Fig. 1 “>>” mark example



```
>>> analyzer.analyze('占い師としてのやる気を見せつつ対抗にも挨拶をするという紳士的な一面を感じさせる.')
[0.3333333333333333]
>>> analyzer.analyze('もうこの時点で占い師として信用を取っていこうとする姿勢がみえるわ')
[1.0]
>>> analyzer.analyze('その場合『おおっと、まじですか』はかなり演技派だね。')
[0]
>>> analyzer.analyze('今後胡散臭さがにじみ出てくるかどうかは見ていきたい。')
[-1.0]
```

Fig. 2 Estimated Sentiment Points



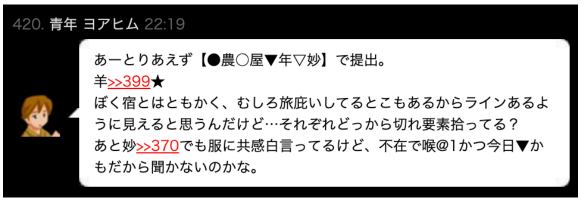
93. ならず者 ディーター 01:45

はじまったな！【占い師CO】

94. 神父 ジムゾン 01:46

おおっと、まじですか
【占い師CO】
神に祈りを！

Fig. 3 Coming Out Dialog



420. 青年 ヨアヒム 22:19

あーとりあえず【●農○屋▼年▽妙】で提出。
羊>>399★
ぼく宿とはともかく、むしろ旅庇いしてるとこもあるからラインあるよ
うに見えると思うんだけど…それぞれどっから切れ要素拾ってる？
あと妙>>370でも服に共感白言ってるけど、不在で喉@1かつ今日▼か
もだから聞かないのかな。

Fig. 4 Voting Symbols Dialog

other protocol-based agents. In the natural language processing area, researchers also try to generate human-like dialog through translating protocols using dialog from real human players [9].

4. Proposed Methods

4.1 Sentiment Points

Not only the sentiment classification but also the object of the sentiment tendencies are crucial problems that are explored in sentiment analysis. Normally, it’s hard to precisely identify the objects mentioned and relate them to the sentiment tendencies. But in Werewolf BBS, players often use the “>>” mark to write direct comments on what previous players said, which could convey the sentiment tendencies. An example is shown in Fig 1. We extract sentences with the mark “>>” and form a corpus at first. Then we use sentiment analysis tool *oseti* [12] which is based on Japanese Sentiment Polarity Dictionary [8] to assign sentiment points for each sentences in our corpus. Examples of sentiment points are shown in Fig 2. In the end, each player will have lists of sentiment points in time series for other players. For the reason that the sentiment point lists for players are in different lengths, we use padding which is often used in text learning tasks to pad it into the same length. Finally, we input the sentiment points between two players and let the machine learn the relationship between two players.

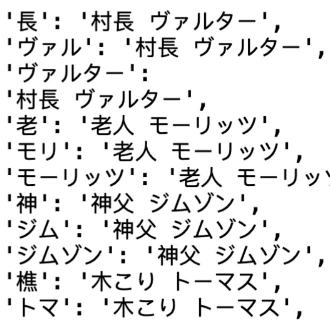
4.2 *Coming Out (CO) Information*

As in other games, the werewolf game has terms that are difficult for outsiders to understand. *Coming out* is one of them. Which means a player publicly announce his/her role. An example is shown in Fig 3. Although different people will using different strategies while playing the role seer, it gradually becomes a consensus that seer should better come out on the first day in order to offer more useful information for players. For the werewolf side players, if they want to play this game more actively, they should also come out as seers, otherwise, it’s hard to passively keep hiding in normal villager players and win. Furthermore, the player with the role possessed sometimes comes out as a seer too, for that the real seer can not identify whether this is a werewolf player or not. So they can maximize the benefit from this special ability by CO as a seer and confuse the players in the villager side. As there are always 2-3 players coming out as the seer on the first day and no real player will die on the first day, the topic for the first two days is always about who is the real seer. It’s intuitive to predict the seer among the first three players who coming out but the possibility that there is a werewolf or possessed in the first three persons that CO as a seer is much higher than random pick. Coming out information is useful for narrow down the candidates for the werewolf that can be the vote out on the first voting session.

4.3 Vote Points

There are many ways to express a player’s intention in natural languages but in werewolf BBS, players use several symbols to express their ideas simply and directly. As shown in Fig 4. The typical meaning of symbols is summarized in Table 2.

We extract those symbols from the dialog and assign different points for them as voting points. Every player is associated with a character name before the game start and they call each other by the name. As shown in Fig 3, but in oral languages, there are always several nicknames for one character name. In order to precisely identify the object of symbols, we generate a name dictionary using the werewolf dictionary [2] written by an expert werewolf game player. Part of the name dictionary is shown in Fig 5. In the end, we have a voting points matrix for each game, and for players, there will be a vector indicate other players’ voting intention towards them. We can then predict whether a player will be voted out or not.



```
'長': '村長 ヴァルター',
'ヴァル': '村長 ヴァルター',
'ヴァルター':
'村長 ヴァルター',
'老': '老人 モーリッツ',
'モリ': '老人 モーリッツ',
'モーリッツ': '老人 モーリッツ',
'神': '神父 ジムゾン',
'ジム': '神父 ジムゾン',
'ジムゾン': '神父 ジムゾン',
'樵': '木こり トーマス',
'トマ': '木こり トーマス',
```

Fig. 5 Character Name Dictionary

Symbol	Meaning
▼	First candidate for voting
▽	Second candidate for voting
●	First candidate for divining
○	Second candidate for divining

Table 2 Meanings for Special Symbols

5. Experiment

We extract 397 game logs from werewolf BBS which include games for around the last 3 years and create 39970 sentiment point pairs for tasks. We use 80% for training and left 20% for testing. Sentiment point pairs generated from one game won't be divided (All in training data set or testing data set). The proposed model used in tasks 5.1 and 5.2 is simply formed with a GRU layer and two fully connected layers. As the samples for the two classes are unbalanced, we choose cross-entropy loss with weight to balance the data.

5.1 Werewolf Prediction

Standing at the viewpoint of a villager player, we use the mutual sentiment points series between villager player A and another player B to predict the possibilities of player B being a werewolf which is a binary classification task. The result is shown in Fig 6. The training accuracy shows a small drop at first and then gradually increases. This may be caused by the unbalanced training data set. The line is the average accuracy of five training processes, the shading area between $100 - 200$ epochs shows a small fluctuation during training.

In real games, there would be other constraints that there are two or three players among over 10 players being a werewolf. So we use the prediction result as a possibility and calculate the likelihood of each combination and choose the one with maximum likelihood as the predicted result. The accuracy shown in Table 3 is calculated by comparing the most possible combination of werewolves with the real werewolves in a game. The test games used are from the test set. The model accuracy is calculated by the number of villager players getting matching results divides by total villager players tested. The number of candidate werewolves in the combinations of the highest 3 likelihoods is 4 or 5 because there are some repeated nominations of werewolves. The compared random model randomly chooses a player combination in which the number of werewolf candidates is the same as the proposed model. As shown in the table, the accuracy of the proposed model that at least one werewolf included in the combinations of the highest 3 likelihoods is 65.5% and at least one werewolf in the combination of highest likelihood is 48.3% which are better than the random model.

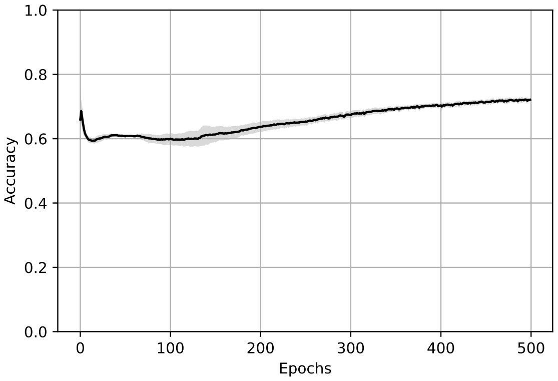


Fig. 6 Werewolf Prediction

	accuracy	
	Proposed	Random
Werewolf included in combinations of highest 3 likelihood	65.5%	55%
Werewolf included in the combination of highest likelihood	48.3%	40%

Table 3 Accuracy of Werewolf Prediction

	accuracy	
	Model	Random
Werewolf or possessed	46.7%	43.8%
Least suspicious player	61.7%	55.2%

Table 4 Accuracy of CO Prediction

5.2 Prediction Based on CO Information

In this task, We used the mutual sentiment points for a villager player A and three players who come out as seers. The target for this task is the combination of werewolf players or the possessed players in the three candidates. The output of this network should be the possibilities of the three players being on the werewolf side or not. So we use MSE loss in this task and the accuracy is calculated in two ways. One is to identify the most suspicious player among the three, another is to identify the least suspicious player among these three players. The result is shown in Table 4. The compared random model randomly chooses a player among the three candidates as a prediction. The accuracy is calculated by the number of games getting matching results divides by the total games tested. The result shows that the proposed model has a better performance than the random model.

5.3 Voting Prediction

The voting prediction network is formed with two fully connected layers. As the samples for the two classes are unbalanced, we choose cross-entropy loss with weight to balance the data. The result is shown in Fig 7. The same as the situation of the werewolf prediction task in 5.1, there is a constraint that there is only one player among over 10 players being voted out on the second day of each game. Similarly, we use the prediction result as a possibility and calculate the likelihood of each combination, and then choose the one with maximum likelihood as the predicted result. To increase the accuracy of prediction, we assume that decisions that appear later in the time series get higher weights, and enhanced the data by multiply the voting points in the latter half with an increasing factor, the factor increases 0.05 each time being used. Three kinds of accuracy are shown in Table 5. The random model randomly chooses the corresponding number of players as a prediction result. The accuracy is calculated by the number of games getting matching results divides by the total games tested. The result of voting prediction is much higher than the chance rate and the accuracy gets higher with the enhanced data.

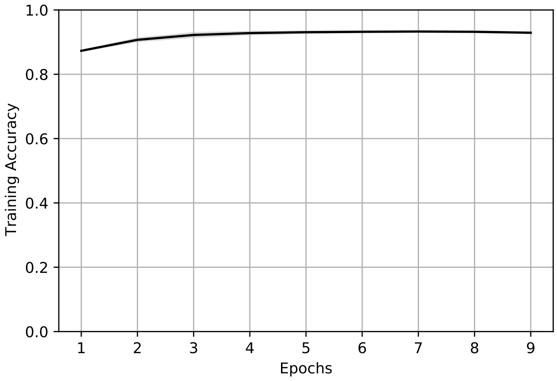


Fig. 7 Voting Prediction

	accuracy		
	Original	Enhanced	Random
Target included in combinations of highest 3 likelihood	96%	97.8%	40.6%
Target included in combinations of highest 2 likelihood	95.6%	97.8%	19.5%
Target matches the combination of highest likelihood	76%	77.8%	0.05%

Table 5 Accuracy of Voting Prediction

6. Conclusion and Future Work

We applied sentiment analysis on the dialog of werewolf games to generate information that is helpful for villager players to estimate the roles of other players. Three tasks, werewolf prediction, prediction on CO information, and voting prediction, are conducted to evaluating how sentiment analysis can help agents deal with the information in the dialog. The voting prediction task gets a consent accuracy while the other two tasks get accuracy better than the random model. The result can be used in further study to create an agent that can play the werewolf game with natural languages or served as an analyzer for human players to analyze the current situation of the game. Several possible factors influence performance. First, sentiment points are indirect features for predicting werewolves. The accuracy of voting prediction is high because the voting points generated from the voting tendencies are direct features for voting prediction. Using sentiment analysis to predict werewolves is a hard task even for humans. Another possible factor is that the time span of the whole game logs is long and the total number of games is small. Game strategies keep updating year by year which will reduce the consistency among all games. Moreover, we assume that all human players are experts that they play the werewolf game with logical consistency and have a certain consensus on game strategies, for example, seers should CO on the first day. But some players don't play by the conventional wisdom, which will cause the randomness of original data.

Despite all of these difficulties, there are still many methods that may increase the performance. The sentiment points now are based on the Japanese Sentiment Polarity Dictionary which is not specified in terms and language styles in werewolf games. Other terms in the werewolf game dictionary could be used in sentiment analysis such as "吊り" which also means voting intention and "白" which means the speaker believes the player is on the villager side. What's more, the divine result also reveals the possible combination of two sides because the werewolf seer may give a good identity to partners. Although the idea is primary, we may proceed with this research using a more advanced natural language processing method and in the end making an agent that can utilize these results.

7. Acknowledgement

A part of this work was supported by JSPS KAKENHI 21H03570.

References

[1] : Werewolf BBS G, http://ninjinix.x0.com/wolfg/index.rb.

[2] ancient: Werewolf Game Dictionary, https://w.atwiki.jp/wolfbbsdictionary/.

[3] Browne, C. B., Powley, E., Whitehouse, D., Lucas, S. M., Cowling, P. I., Rohlfshagen, P., Tavener, S., Perez, D., Samothrakis, S. and Colton, S.: A Survey of Monte Carlo

Tree Search Methods, *IEEE Transactions on Computational Intelligence and AI in Games*, Vol. 4, No. 1, pp. 1–43 (online), DOI: 10.1109/TCIAIG.2012.2186810 (2012).

[4] Campbell, M., Hoane Jr, A. J. and Hsu, F.-h.: Deep blue, *Artificial intelligence*, Vol. 134, No. 1-2, pp. 57–83 (2002).

[5] Cho, K., Van Merriënboer, B., Gulcehre, C. and et al.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, *arXiv preprint arXiv:1406.1078* (2014).

[6] Fukui, T., Ando, K., Murakami, T., Ito, N. and Iwata, K.: Automatic classification of remarks in Werewolf BBS, *2017 5th Intl Conf on Applied Computing and Information Technology/4th Intl Conf on Computational Science/Intelligence and Applied Informatics/2nd Intl Conf on Big Data, Cloud Computing, Data Science (ACIT-CSII-BCD)*, IEEE, pp. 210–215 (2017).

[7] Kondoh, M., Matsumoto, K. and Mori, N.: Development of Agent Predicting Werewolf with Deep Learning, *International Symposium on Distributed Computing and Artificial Intelligence*, Springer, pp. 18–26 (2018).

[8] Lab, T. N.: Japanese Sentiment Polarity Dictionary, `http://www.cl.ecei.tohoku.ac.jp/Open_Resources-Japanese_Sentiment_Polarity_Dictionary.html`.

[9] Nakamura, H., Katagami, D., Toriumi, F. and et al.: Generating human-like discussion by paraphrasing a translation by the AIWolf protocol using werewolf BBS logs, *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, pp. 1–6 (2017).

[10] Toriumi, F., Osawa, H. and Inaba, Michimasa, a. e. a.: AI wolf contest—development of game AI using collective intelligence—, *Computer Games*, Springer, pp. 101–115 (2016).

[11] Wang, T. and Kaneko, T.: Application of Deep Reinforcement Learning in Werewolf Game Agents, *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*, IEEE, pp. 28–33 (2018).

[12] Yukino, I.: oseti:Dictionary based Sentiment Analysis for Japanese, `https://github.com/ikegami-yukino/oseti`.

[13] Zinkevich, M., Johanson, M., Bowling, M. and Piccione, C.: Regret minimization in games with incomplete information, *Advances in neural information processing systems*, Vol. 20, pp. 1729–1736 (2007).

Appendix

A.1 Hyper parameters for training

Parameters	Value
Optimizer	Adam
Loss Function	Cross-entropy Loss
Weight	[1., 3.53]
Learning Rate	0.0005
Epochs	500

Table A·1 Parameters for Task 5.1

Parameters	Value
Optimizer	Adagrad
Loss Function	MSE Loss
Learning Rate	0.0001
Epochs	500

Table A·2 Parameters for Task 5.2

Parameters	Value
Optimizer	Adam
Loss Function	Cross-entropy Loss
Weight	[19., 1.]
Learning Rate	0.0001
Epochs	500

Table A·3 Parameters for Task 5.2