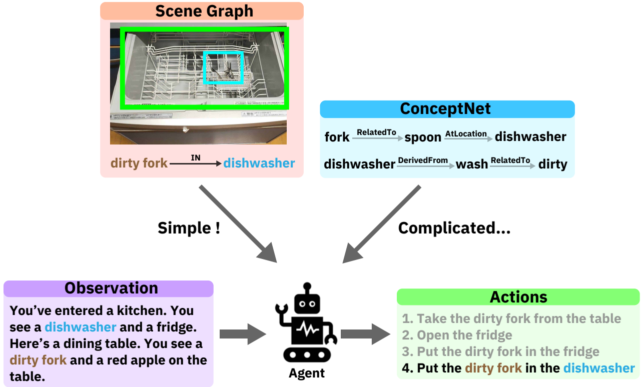
Commonsense Knowledge from Scene Graphs for Textual Environments

Tsunehiko Tanaka, [1,2*] **Daiki Kimura,** [1] **Michiaki Tatsubori** [1]

[1] IBM Research [2] Waseda University
tsunehiko@fuji.waseda.jp, daki@jp.ibm.com, mich@jp.ibm.com

Abstract

Text-based games are becoming commonly used in reinforcement learning as real-world simulation environments. They are usually imperfect information games, and their interactions are only in the textual modality. To challenge these games, it is effective to complement the missing information by providing knowledge outside the game, such as human common sense. However, such knowledge has only been available from textual information in previous works. In this paper, we investigate the advantage of employing commonsense reasoning obtained from visual datasets such as scene graph datasets. In general, images convey more comprehensive information compared with text for humans. This property enables to extract commonsense relationship knowledge more useful for acting effectively in a game. We compare the statistics of spatial relationships available in Visual Genome (a scene graph dataset) and ConceptNet (a text-based knowledge) to analyze the effectiveness of introducing scene graph datasets. We also conducted experiments on a text-based game task that requires commonsense reasoning. Our experimental results demonstrated that our proposed methods have higher and competitive performance than existing state-of-the-art methods.



You've entered a kitchen. You see a dishwasher and a fridge. Here's a dining table. You see a dirty fork and a red apple on the table.

1. Take the dirty fork from the table
2. Open the fridge
3. Put the dirty fork in the fridge
4. Put the dirty fork in the dishwasher

Figure 1: Illustration of our commonsense acquisition from scene graphs. To provide commonsense: dirty fork → IN → dishwasher to an agent, a single image is sufficient for scene graphs (top left), but ConceptNet requires several graphs to be combined, which is redundant.

Introduction

Reinforcement learning (RL) is a type of machine learning method that has a great advantage of not requiring labeled data and has been used in various simulation environments (Mnih et al. 2015; Silver, Huang, and et al. 2016; Kimura 2018; Kimura et al. 2018). Since textual conversation agents are commonly used in our daily lives in the real world, text-based environments, where both observation and action spaces are restricted to the modality of text, have been attracting attention. RL in such environments requires developing an agent to have language comprehension skills by natural language process and sequential decision-making in the complex environment. This means the textual observation contains a lot of noisy information and the problem of partial observability.

Text-based games are a partially observable Markov decision process (POMDP) (Kaelbling, Littman, and Cassandra 1998) where the agent cannot observe the entire information from the text given by the environment. TextWorld (Côté et al. 2018) is a textual game generator and extensible sandbox learning environment for RL agents, and various methods have been proposed for this game to compensate for the missing information (Kimura et al. 2021b; Murugesan et al. 2021; Carta et al. 2020; Murugesan, Chaudhury, and Talamadupula 2021; Shridhar et al. 2020; Kimura et al. 2021a,c; Chaudhury et al. 2021). There are three types of extensions: external knowledge, new modality, and logical rule extraction. External knowledge that is useful for training agents from humans or other domain sources. A study reports commonsense knowledge is an important aspect of human intelligence (Murugesan et al. 2021). In this study, TextWorld Commonsense (TWC), which requires commonsense as external knowledge, is proposed as an extension of TextWorld. The task of the TWC game is cleaning up a room, and the commonsense in this game is mainly place information for each object. The same study also includes a baseline agent for TWC games that uses a commonsense subgraph extracted from external knowledge (we call this model *TWC agent* and the environment *TWC games* to distinguish them). Another study reported that introducing ex-

[*] This work was done during an internship at IBM Research.

ternal knowledge from humans as logical functions helps the training of the agent (Kimura et al. 2021b). New modality information extracted from observations or action text can be introduced to make decisions (Carta et al. 2020; Murugesan, Chaudhury, and Talamadupula 2021; Shridhar et al. 2020). In these methods, visual information from images or videos is commonly used since it has been used in many other studies (Tanaka and Simo-Serra 2021; Kimura et al. 2020) to understand attention and sequential information in decision making. Logical rule extraction can be exploited to improve the speed of training and interpretability of the agent (Kimura et al. 2021c; Chaudhury et al. 2021). Since commonsense knowledge is normally represented by a graph structure, the logical rule representation is compatible with commonsense knowledge.

However, at the time of writing, there has been no research that utilizes the benefits of these multiple extensions to compensate for missing information. In particular, we hypothesize that the commonsense knowledge of object place relationships that are used in TWC games can be easily obtained from visual information. For example, instead of stating the place name of each object, operators can display a picture of a tidy room, which is a quicker explanation for humans.

In this paper, we propose a novel agent that challenges a TWC game by leveraging visual scene graph datasets to obtain commonsense. The original TWC agent (Murugesan et al. 2021) constructs a commonsense subgraph from ConceptNet (Speer, Chin, and Havasi 2017a), which is textual knowledge, but it is necessary to combine many graphs to obtain one commonsense and to create a complicated subgraph. In fact, Murugesan et al. prepared a 'manual' commonsense subgraph from ConceptNet to tackle this complexity of graphs in their study. However, since scene graph recognition achieves high accuracy from complex images, visual information can deliver various detailed and organized graph information all at once. Figure 1 shows an example for the acquisition of commonsense knowledge from scene graphs in an image. In this example, despite ConceptNet having redundant information for extracting a commonsense subgraph, the proposed extraction from scene graphs has necessary and sufficient information for the cleaning-up task. Furthermore, relationships from scene graphs also contain direct spatial relationships such as "on" or "in" (Figure 2) between objects because agents need to determine an object's place in the TWC game. Therefore, we use scene graph datasets as visual external knowledge. A scene graph dataset contains a large number of graphs that represent the relationships between entities in images. We use Visual Genome (VG) (Krishna et al. 2017) as a scene graph dataset, which is the most commonly used, and compare its statistics with ConceptNet. We also conduct experiments to evaluate the performance of agents with commonsense knowledge from a scene graph dataset in RL on text-based games.

Related Work

Text-based RL Games

Text-based interactive RL games has been gaining the focus of many researchers due to the development of environments such as TextWorld (Côté et al. 2018) and Jericho (Hausknecht et al. 2019). In these games, RL agents are required to understand the high-level context information from only textual observation. To overcome this difficulty, a number of prior works on these environments have extracted new information from textual observations: knowledge graphs, visual information, and logical rule.

Knowledge graphs represent relationships between entities like real-world objects and events, or abstract concepts. A new text-based environment, called "TextWorld Commonsense", was proposed in (Murugesan et al. 2021) to infuse RL agents with commonsense knowledge and developed baseline agents using a commonsense subgraph constructed from ConceptNet(Liu and Singh 2004; Speer, Chin, and Havasi 2017a) as an external knowledge. We use this work as a baseline method, and introduce a new type of commonsense from visual datasets. Worldformer (Ammanabrolu and Riedl 2021) represents environment status as a knowledge graph and uses a world model to predict changes caused by an agent's actions and generates a set of contextually relevant actions.

While knowledge graphs are useful for organizing abstract information from only text descriptions, visual information enables the agent to obtain a detailed locational situation like human imagination and visualization. The most important issue in using visual information is how to obtain it from only textual observation in text-based games. VisualHints (Carta et al. 2020) proposed an environment that can automatically generate various hints about game states from textual observation and changes the difficulty level depending on their type. The main sources of images in (Murugesan, Chaudhury, and Talamadupula 2021) are retrieved from the Internet and generated from a text-to-image pre-trained model, AttnGAN (Xu et al. 2018) with given text descriptions. ALFWorld (Shridhar et al. 2021) combines TextWorld and an embodied simulator called ALFRED (Shridhar et al. 2020) to obtain information on two modalities. Shridhar et al. proposed an agent that first learns to solve abstract tasks in TextWorld, then transfers the learned high-level policies to low-level embodied tasks in ALFRED.

In addition, even if we use the aforementioned methods, improvements in the speed of training are few and the interpretability of the trained network is still missing. A number of studies (Kimura et al. 2021c; Chaudhury et al. 2021) proposed novel approaches to extract symbolic first-order logics from text observations, and select actions by using neuro-symbolic Logical Neural Networks (Riegel et al. 2020). These logical representations are compatible with commonsense graph structures.

As previously described, there have been various approaches using knowledge graphs, visual information, and logical rules. However, at the time of writing, there has been no method that combines any of them. Therefore, we pro-

	entity	relationship	triplet
VG	63,686	36,550	662,934
ConceptNet	38,556	46	298,394
Manual	111	2	132

Table 1: Statistics of the three types of external knowledge. Each item denotes the number of species. Manual is manual commonsense knowledge used in (Murugesan et al. 2021).

pose an approach to extract and utilize knowledge graphs from visual information.

Scene Graph Dataset

A scene graph is a structured representation of the relationships between objects in a scene. To train a scene graph generation model, a number of datasets have been created. VG (Krishna et al. 2017) is a large-scale scene graph dataset that is most commonly used these days because it contains various elements such as objects, attributes, relationships, QA descriptions, and so on. Since scene graphs can provide a large number of visual relationships in a single image, we use VG datasets as external knowledge for training agents.

ConceptNet vs Scene Graph Datasets

In this section, to show scene graph datasets are effective as external knowledge for solving TWC games, we compare ConceptNet and scene graph datasets. We first show the statistics of ConceptNet, VG (Krishna et al. 2017), and manual commonsense knowledge designed in (Murugesan et al. 2021). Next, we compare ConceptNet and VG in terms of similarity to the manual commonsense knowledge. VG is the most commonly used scene graph dataset. The manual commonsense knowledge is manually extracted from ConceptNet to include only the pairs of an object in TWC games and goal location for each object. Since the entities are directly related to actions in the games, the agent with this manually-crafted information is more effective for solving the games. Therefore, external knowledge that is similar to the manual commonsense knowledge are comfortable with this task.

Knowledge Statistics

We summarize the statistics of the three types of external knowledge in Table 1. In general for all external knowledge, each graph is represented as a triplet $\langle e_1, r_{12}, e_2 \rangle$: e_1, e_2 denote entities in an image, and r_{12} denotes a relationship between e_1 and e_2. In Table 1, 'entity', 'relationship', and 'triplet' indicate the number of species of e, r, $\langle e_1, r_{12}, e_2 \rangle$, respectively. The huge difference between ConceptNet and VG is the number of species of relationships, and this indicates that VG has more detailed information in relationships than ConceptNet. Figure 2 shows an example of r in these datasets. In ConceptNet, the spatial relationship is only 'at location', which is the second most common, but its ratio to the total is low. In contrast, spatial relationships such as 'on', 'in', and 'under' dominate VG. In addition, 'has' and 'with' can also express spatial relationships, such as $\langle \text{building}, \text{has}, \text{window} \rangle$ and $\langle \text{window}, \text{with}, \text{building} \rangle$. Thus, we can see that VG has a larger number of spatial relationships than ConceptNet. However, the manual commonsense knowledge has two species of relationships: 186 'at location' and 6 'related to'. This means that spatial relationships are important for solving TWC games, and VG has an advantage in this respect.

Similarity to Manual Commonsense Knowledge

To examine whether VG contains knowledge graphs useful for solving TWC games, we compare ConceptNet and VG in terms of similarity to the manual commonsense knowledge. We calculate the similarities for both entity e and entity pairs $\{e_1, e_2\}$ as described in the following.

Entity e Given an entity e_i^V from VG, we use GloVe (Pennington, Socher, and Manning 2014) embeddings to represent e_i^V as a d-dimensional vector $\boldsymbol{z}_i^V$, where $\boldsymbol{z}_i^V \in \mathbb{R}^d$ is the word embedding of the entity. Similarly, the embedding of each entity in ConceptNet e_j^C and in manual commonsense knowledge e_k^M are denoted as $\boldsymbol{z}_j^C, \boldsymbol{z}_k^M$, respectively. We calculate the similarity $s_i^{eV} k$ between an entity in VG e_i^V and in manual commonsense knowledge e_k^M by Eq. 1.

$$s_{ik}^{eV} = \text{cos\_similarity}(\boldsymbol{z}_i^V, \boldsymbol{z}_k^M) \tag{1}$$

Similarly, Eq. 1 is executed for entities in ConceptNet. We count the number of entities whose similarity is above the threshold 0.7.

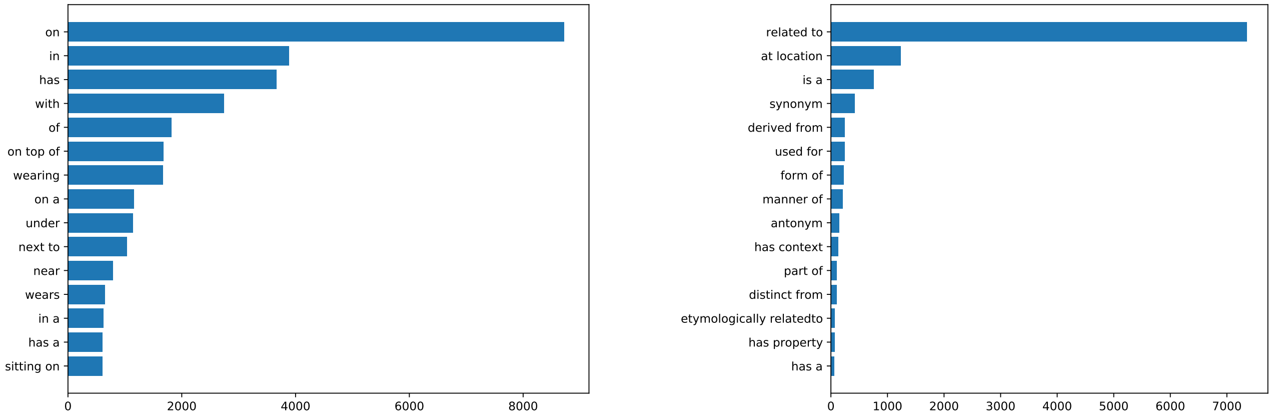
Pair of entities $\{e_1, e_2\}$ We also compare sets of pairs of e_1 and e_2 from these datasets in terms of similarity to the 132 pairs in the manual commonsense knowledge. In the same way as the entities previously described, we use GloVe to represent e_{i1}^V and e_{i2}^V from triplet $t_i^V = \langle e_{i1}^V, r_{i1i2}^V, e_{i2}^V \rangle$ in VG as d-dimensional vectors $\boldsymbol{z}_{i1}^V, \boldsymbol{z}_{i2}^V$, respectively. Similarly, $\boldsymbol{z}_{j1}^C, \boldsymbol{z}_{j2}^C$ and $\boldsymbol{z}_{k1}^M, \boldsymbol{z}_{k2}^M$ are the embeddings of entities from each triplet in ConceptNet $t_j^C = \langle e_{j1}^C, r_{j1j2}^C, e_{j2}^C \rangle$ and manual commonsense knowledge $t_k^M = \langle e_{k1}^M, r_{k1k2}^M, e_{k2}^M \rangle$, respectively. We calculate the similarity using the sum of both embeddings of entities in a triplet. Thus, the similarity s_{ik}^{pV} is given by Eq. 2.

$$s_{ik}^{pV} = \text{cos\_similarity}(\boldsymbol{z}_{i1}^V + \boldsymbol{z}_{i2}^V, \boldsymbol{z}_{k1}^M + \boldsymbol{z}_{k2}^M) \tag{2}$$

For ConceptNet, we use Eq. 2 similarity. We set a threshold to 0.65 and count pairs over the threshold.

The results of the entity and pair counts are summarized in Table 2. Although there is no significant difference in the number of types of entities, VG has more pairs associated with the manual commonsense knowledge, which indicates that VG has more game-related relationships than ConceptNet.

The aforementioned comparisons show that TWC games need more spatial relationships in the external knowledge, and VG is more effective for solving TWC games than ConceptNet.



(a) VG

(b) ConceptNet

Figure 2: Histogram of relationships (top 15) included in VG and ConceptNet. VG has much more spatial relationships than ConceptNet.

	entity	pair
VG	147	58,005
ConceptNet	156	11,550

Table 2: Comparison of the number of entities and pairs similar to manual commonsense knowledge in external knowledge. A pair is a combination of e_1 and e_2 from a triplet $\langle e_1, r_{12}, e_2 \rangle$. The number of entities is not very different, but the number of pairs is much higher in VG than in ConceptNet.

Proposed Method

Previous TWC agent

The proposed methods extend the TWC agent (Murugesan et al. 2021), which is a baseline model for TextWorld Commonsense. We briefly explain the network architecture as follows.

The TWC agent consists of the six components: (a) *action encoder*, which encodes all admissible actions a, (b) *observation encoder*, which encodes the observation o_t, (c) *context encoder*, which encodes the dynamic context C_t, (d) *dynamic commonsense subgraph*, which is commonsense information G_C^t extracted by the agent, (e) *knowledge integration*, which combines the information from textual observation and the extracted commonsense subgraph, and (f) *action selection*, which selects an action from given action candidates. We subsequently describe *dynamic commonsense subgraph* and *knowledge integration*, which are important for this paper.

For *dynamic commonsense subgraph*, the TWC agent retrieves commonsense from external knowledge like ConceptNet(Speer, Chin, and Havasi 2017a) and updates a subgraph by combining it with the graph at a previous time step. At time t, the agent first extracts entities involving game status from textual observation and then obtains a set of cumulative entities E_t by combining it with the entities from the previous graph G_C^{t-1}. The commonsense subgraph G_C^t is constructed automatically from E_t and *Context Direct Connections (CDC)*, which is another algorithm of external knowledge. For CDC, the entities are split into two groups in accordance with their attributes, and then links between the groups are added.

For *knowledge integration*, the TWC agent encodes the commonsense subgraph and integrates the graph embedding vector with the observation context feature. In the encoding phase, the node embedding is first extracted from the commonsense subgraph using a pre-trained knowledge graph embedding called Numberbatch (Speer, Chin, and Havasi 2017b) and a *sentinel* vector (Lu et al. 2017) is added to enable the attention to not attend to any specific nodes in the commonsense subgraph. These embeddings are updated by messages passing between the nodes of graph attention networks (GAT) (Veličković et al. 2018). In the integration phase, *Co-Attention* is used, which is a bidirectional attention flow layer between the observational context and the commonsense subgraph.

TWC agent is an attractive design for RL agents on text-based games, and it accesses a commonsense and uses it while selecting actions. However, the source format of external knowledge is limited to text. Since textual knowledge such as ConceptNet is very useful, it is redundant because multiple concepts need to be concatenated to express more detailed information. Therefore, they prepared manually retrieved graphs in the paper (Murugesan et al. 2021). The manual graphs contain direct connections for the objects and their goal locations.

Level	Objects	Objects to find	Rooms
Easy	1	1	1
Medium	2, 3	1, 2, 3	1
Hard	6, 7	5, 6, 7	1, 2

Table 3: Specification of TWC games from (Murugesan et al. 2021).

Proposed Method

From the comparison in the previous section, it is revealed that scene graph datasets are effective for TWC games because they have more spatial relationships. This suggests that the issue of the TWC agent is that the external knowledge is limited to being text-based. To address this, we propose an approach to use scene graph datasets as external knowledge to build a commonsense subgraph for agents. Our proposal has two types depending on the training method and external knowledge.

Scene Graph The simplest model is to replace the external knowledge of the TWC agent with scene graph datasets from ConceptNet. As shown in Table 2, a scene graph dataset holds many triplets that are effective for solving TWC games, so we expect to obtain a higher score.

ConceptNet + SG We also propose a method to complement the weak point of textual knowledge with scene graph datasets. Inspired by curricular learning (Bengio et al. 2009), we first provide the agent with textual knowledge and train it, then provide the same agent with external knowledge from scene graph datasets and continue training. We hypothesize that this method is effective in training agents that have commonsense knowledge balanced between abstract and concrete knowledge. In the first step, the overall commonsense is given by the textual knowledge. In the second step, the specific commonsense focused on location is given from the scene graph dataset.

Experiments

Experimental Setup

We conduct our experiments on TWC games (Murugesan et al. 2021). A TWC game is a text-based game where the goal is to tidy up a house by putting objects where they should be. The connection between objects and the locations where they should be is not given by the game, so the agent needs to depend on commonsense knowledge. This domain has three difficulty levels (easy, medium, and hard) depending on the total number of objects in the game, the number of objects in which the agent needs to find their locations, and the number of rooms to explore. The numbers are randomly sampled from the list in Table 3. In our evaluation, we consider all difficulty levels. We also use two types of test sets: *IN* and *OUT*. The games in the *IN* were built on the same entities as the training set, and the entities in the *OUT* do not appear in the training set. We can evaluate the ability to generalize unseen entities from these test sets.

Our experimental setup is based on the evaluation system in (Murugesan et al. 2021); we use the Advantage Actor-Critic algorithm (Mnih et al. 2016). The most significant difference from the previous system is that all agents use GloVe for graph embedding. Numberbatch (Speer, Chin, and Havasi 2017b) used in the previous system is a combination of existing pre-trained embeddings such as word2vec (Mikolov et al. 2013) and GloVe retrofitted with ConceptNet's graph. The evaluation experiments in (Murugesan et al. 2021) have shown that a TWC agent with Numberbatch achieved a better performance than with GloVe because of a high affinity with ConceptNet. Since we focus on the impact of external knowledge, we use only GloVe for both graph and observation embeddings in all agents.

Metrics

We measure the performance of agents with the various external knowledge on TextWorld using two metrics: the normalized score and the number of steps taken. The normalized score is calculated by dividing the actual score by the maximum possible score. Steps indicate time spent to reach the goal and the lower the value, the higher the performance.

Results

We compare four types of agents: ConceptNet, ConceptNet + Manual, Scene Graph, and ConceptNet + SG. Both Scene Graph and ConceptNet + SG are our proposed methods that use VG. The other agents are baselines proposed in (Murugesan et al. 2021). ConceptNet + Manual uses manually-prepared knowledge directly related to the game from ConceptNet. ConceptNet + Manual should show human-like performance and is regarded as the upper bound of the performance of the proposed method (especially for IN). ConceptNet + SG is first trained for 100 episodes using ConceptNet, followed by 100 episodes using VG. All agents except ConceptNet + SG are trained for 100 episodes each and all results are the average of five runs. The results are summarized in Table 4. We also show the training curves in Figure 3.

We can see three overall trends in these results. First, our proposed methods outperform the baselines in both steps and scores in the easy and medium levels. The commonsense knowledge obtained from scene graph datasets is proved to be effective in solving TWC games. Second, all agents struggle with the hard level. It is necessary to have the ability to deal with complicated situations where the location to pick up objects is different from the location to place them. The development of agents with this ability is future work. Finally, ConceptNet + Manual shows high performance for the IN set regardless of difficulty level. ConceptNet + Manual has crucial information for solving TWC games, so both efficiency and scores are higher when the same species of entity is given as the training set. However, its performance in the *OUT* is lower than that of other agents because it is given only a minimum number of necessary graphs in the training set, so it overfits to those graphs.

We describe the performance of our proposed models in detail. We propose two models, Scene Graph and Concept-

	Method	Easy		Medium		Hard	
		#Steps↓	Score↑	#Steps↓	Score↑	#Steps↓	Score↑
IN	ConceptNet [*]	20.34 ± 0.96	0.82 ± 0.04	41.80 ± 1.10	0.65 ± 0.06	50.00 ± 0.00	0.25 ± 0.07
	ConceptNet + Manual [*]	18.34 ± 2.78	0.85 ± 0.05	**37.27 ± 2.83**	**0.77 ± 0.06**	**49.59 ± 0.62**	**0.35 ± 0.03**
	Scene Graph	**17.23 ± 3.16**	**0.90 ± 0.05**	42.18 ± 2.66	0.58 ± 0.08	50.00 ± 0.00	0.32 ± 0.08
	ConceptNet + SG	20.43 ± 1.28	0.78 ± 0.04	38.30 ± 6.08	0.71 ± 0.12	50.00 ± 0.00	0.28 ± 0.08
OUT	ConceptNet [*]	18.05 ± 4.64	0.88 ± 0.08	44.30 ± 4.42	0.50 ± 0.09	**50.00 ± 0.00**	**0.19 ± 0.04**
	ConceptNet + Manual [*]	27.01 ± 2.72	0.69 ± 0.05	46.44 ± 1.15	0.55 ± 0.05	**50.00 ± 0.00**	**0.19 ± 0.02**
	Scene Graph	**17.24 ± 3.84**	0.91 ± 0.05	**41.44 ± 5.45**	0.55 ± 0.14	**50.00 ± 0.00**	0.15 ± 0.06
	ConceptNet + SG	17.47 ± 3.13	**0.92 ± 0.04**	42.57 ± 2.69	**0.63 ± 0.04**	**50.00 ± 0.00**	0.13 ± 0.05

[*] Baseline text-based RL agents with commonsense from external knowledge (Murugesan et al. 2021)

Table 4: Generalization results for two test sets, *IN* and *OUT*, on games with three difficulty levels. Scene Graph and ConceptNet + SG are our proposed methods, ConceptNet and ConceptNet + Manual are baseline agents with commonsense from external knowledge (Murugesan et al. 2021). *IN* is built using the same entities as the training set, and *OUT* is built using different entities. **#Steps** (lower is better) denotes the steps needed to accomplish the goals and **Score** (higher is better) denotes the normalized score by maximum possible score. Each value is a pair (average) ± (standard deviation).

Net + SG, depending on the type of external knowledge and training method. Scene Graph is a simple agent that uses only a scene graph dataset. From Table 4, we found that this agent is very efficient in its graph search. For easy-level games, Scene Graph is superior to even ConceptNet + Manual in the *IN* set. In medium-level games, Scene Graph has the largest number of steps in the *IN* set, but it has the smallest number in the OUT set. This indicates that Scene Graph is robust against unseen objects. The efficiency can be seen from the fastest convergence of the training curves in the medium and hard levels in Figure 3. In terms of performance, although it is sometimes inferior to our other proposed method, it is better than the baselines in *OUT*. The high efficiency and performance of Scene Graph can be attributed to scene graph datasets having many graphs relevant to the TWC game with spatial relationships. ConceptNet + SG is an agent that is trained with ConceptNet, followed by scene graph datasets. Table 4 shows that the performance of this agent is very high. It has the best performance in the *OUT* set for both easy and medium levels and the second-best performance in the *IN* set after ConceptNet + Manual. This indicates that it also has robustness against unseen objects. The reason for the performance improvement is considered to be the wide range that can be handled by both commonsense knowledge from ConceptNet and VG. In easy-level games, the score increases by 0.14 from *IN* to *OUT*, which is an uncommon improvement. The reason for this could be that VG has a small number of graphs related to the easy-*IN* games, which decrease the results of ConceptNet + SG for easy-IN games. As you can see in Table 3, easy games has only require one object to be explored, so the score changes dramatically depending on whether or not external knowledge contains graphs related to the object and the goal location. However, since the number of graphs increases, the efficiency of the search decreases, resulting in inferiority with scene graphs in steps. The training curve in Figure 3 shows that VG enhances the performance of ConceptNet alone from 100 episodes (in the easy level, the training curve converges during the ConceptNet phase, so ConceptNet and ConceptNet + SG overlap).

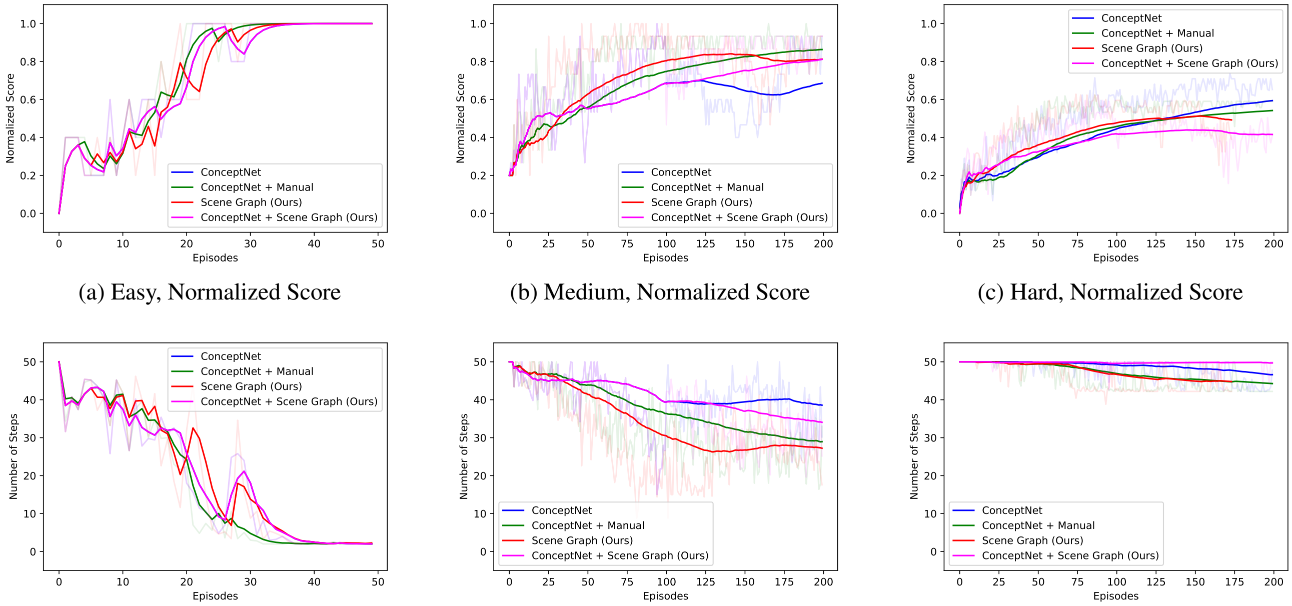
In addition, we use GloVe for graph embedding in these experiments for a fair comparison, but agents using ConceptNet can be improved by replacing GloVe with Numberbatch.

In summary, our proposed method achieves both efficiency and performance improvements, and it is robust to unseen objects. However, it is still a challenge to deal with complex situations such as hard-level games. We discuss how to address this issue in the next section.

Conclusion and Future Work

We have presented new approaches to leverage commonsense subgraphs constructed from scene graph datasets for text-based games. We conducted experiments on a TWC game, which is a benchmark to evaluate how well an agent learns with commonsense knowledge. Experimental results showed that our proposed approaches using a VG dataset demonstrate highly competitive performances compared with existing state-of-the-art approaches with textual knowledge. We also illustrated that the performance can be further improved by using ConceptNet and scene graph datasets sequentially.

Although this work is the first step to utilize both visual information and commonsense, we still have a few challenges as future work. One topic is to deal with more complex situations like hard difficulty level games. We expect that this can be improved by exploiting the relationships among information available from external knowledge. The current model adopts GAT (Veličković et al. 2018) as a graph encoder, but GAT only takes the node features as input and ignores the edge features except for whether they exist or not. In complex tasks, the key to solving this game is more specific information than simply the link between an object and location. In particular, scene graph datasets provide more detailed information on relationships than textual knowledge as shown in Fig. 1. We consider applying networks such as Edge feature enhanced Graph Neural Networks (EGNN) (Gong and Cheng 2019) that can take advan-



(a) Easy, Normalized Score

(b) Medium, Normalized Score

(c) Hard, Normalized Score

(d) Easy, Steps

(e) Medium, Steps

(f) Hard, Steps

Figure 3: Performance evaluation for the three levels of the training set games (Smoothing is performed to clarify the difference in the results of a single run).

tage of the edge features of graphs. Another topic is introducing logical rule training into our proposed method. Since graph information can be easily converted to logical rules, we hope the commonsense graph can directly contribute to logical rule training for action policies in RL.

References

Ammanabrolu, P.; and Riedl, M. O. 2021. Learning Knowledge Graph-based World Models of Textual Environments. *arXiv preprint arXiv:2106.09608*.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Carta, T.; Chaudhury, S.; Talamadupula, K.; and Tatsubori, M. 2020. VisualHints: A Visual-Lingual Environment for Multimodal Reinforcement Learning. In *arxiv*.

Chaudhury, S.; Sen, P.; Ono, M.; Kimura, D.; Tatsubori, M.; and Munawar, A. 2021. Neuro-Symbolic Approaches for Text-Based Policy Learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3073–3078. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Côté, M.-A.; Kádár, A.; Yuan, X.; Kybartas, B.; Barnes, T.; Fine, E.; Moore, J.; Tao, R. Y.; Hausknecht, M.; Asri, L. E.; Adada, M.; Tay, W.; and Trischler, A. 2018. TextWorld: A Learning Environment for Text-based Games. *CoRR*, abs/1806.11532.

Gong, L.; and Cheng, Q. 2019. Exploiting edge features for graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9211–9219.

Hausknecht, M.; Ammanabrolu, P.; Côté, M.-A.; and Yuan, X. E. 2019. Interactive Fiction Games: A Colossal Adventure. In *AAAI 2020*.

Kaelbling, L. P.; Littman, M. L.; and Cassandra, A. R. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1): 99–134.

Kimura, D. 2018. DAQN: Deep Auto-encoder and Q-Network. *arXiv:1806.00630*.

Kimura, D.; Chaudhury, S.; Narita, M.; Munawar, A.; and Tachibana, R. 2020. Adversarial Discriminative Attention for Robust Anomaly Detection. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2161–2170.

Kimura, D.; Chaudhury, S.; Ono, M.; Tatsubori, M.; Agravante, D. J.; Munawar, A.; Wachi, A.; Kohita, R.; and Gray, A. 2021a. LOA: Logical Optimal Actions for Text-based Interaction Games. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, 227–231. Online: Association for Computational Linguistics.

Kimura, D.; Chaudhury, S.; Tachibana, R.; and Dasgupta, S. 2018. Internal Model from Observations for Reward Shaping. In *ICML workshop*.

Kimura, D.; Chaudhury, S.; Wachi, A.; Kohita, R.; Munawar, A.; Tatsubori, M.; and Gray, A. 2021b. Reinforcement Learning with External Knowledge by using Logical Neural Networks. *KBRL Workshop at IJCAI-PRICAI 2020*.

Kimura, D.; Ono, M.; Chaudhury, S.; Kohita, R.; Wachi, A.; Agravante, D. J.; Tatsubori, M.; Munawar, A.; and Gray, A. 2021c. Neuro-Symbolic Reinforcement Learning with First-Order Logic. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3505–3511. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Liu, H.; and Singh, P. 2004. ConceptNet — A Practical Commonsense Reasoning Tool-Kit. *BT Technology Journal*, 22: 211–226.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 375–383.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 1928–1937. PMLR.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; and et al. 2015. Human-level control through deep reinforcement learning. *Nature*.

Murugesan, K.; Atzeni, M.; Kapanipathi, P.; Shukla, P.; Kumaravel, S.; Tesauro, G.; Talamadupula, K.; Sachan, M.; and Campbell, M. 2021. Text-based RL Agents with Commonsense Knowledge: New Challenges, Environments and Baselines. In *Thirty Fifth AAAI Conference on Artificial Intelligence*.

Murugesan, K.; Chaudhury, S.; and Talamadupula, K. 2021. Eye of the Beholder: Improved Relation Generalization for Text-based Reinforcement Learning Agents. *arXiv preprint arXiv:2106.05387*.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Riegel, R.; Gray, A.; Luus, F.; Khan, N.; Makondo, N.; Akhalwaya, I. Y.; Qian, H.; Fagin, R.; Barahona, F.; Sharma, U.; et al. 2020. Logical neural networks. *arXiv preprint arXiv:2006.13155*.

Shridhar, M.; Thomason, J.; Gordon, D.; Bisk, Y.; Han, W.; Mottaghi, R.; Zettlemoyer, L.; and Fox, D. 2020. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shridhar, M.; Yuan, X.; Côté, M.-A.; Bisk, Y.; Trischler, A.; and Hausknecht, M. 2021. ALFWorld: Aligning Text and Embodied Environments for Interactive Learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Silver, D.; Huang, A.; and et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529: 484–503.

Speer, R.; Chin, J.; and Havasi, C. 2017a. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4444–4451. AAAI Press.

Speer, R.; Chin, J.; and Havasi, C. 2017b. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. AAAI'17, 4444–4451. AAAI Press.

Tanaka, T.; and Simo-Serra, E. 2021. LoL-V2T: Large-Scale Esports Video Description Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 4557–4566.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. *International Conference on Learning Representations*.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks.