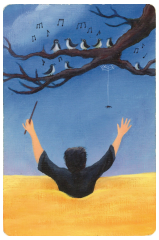
Dixit AI

September 18, 2024

Stefano Saravalle

Abstract

The Dixit board game present a major challenge in AI because of the creativity and abstraction needed to play it. Spanning from image captioning to image-text associations, a bot able to play the game must reason ingeniously to trick both humans and other AIs. The following work present a fully functioning artificial player made by an image captioning model and a CLIP model, along with a custom dataset for Dixit cards created using an open source LLM. All the code can be found at the following link: https://github.com/4idrossifenil-etanammide/Dixit



(a) Twitter

(b) Whimsical, fierce, captivating.

Figure 1. (a) Card caption from the dataset found online. (b) Card caption given by the agent

1. Introduction

Assuming that the reader has a general knowledge about the game, we can identify three main phases(Kunda & Rabkina, 2020), with their own respective challenges and methods used to solve them:

- **Narrator phase** In this phase the narrator has to select a card and give a description for it. The given sentence must not be too obvious neither too vague, otherwise the card may be voted by everyone or no one. This task can be also rephrased as a *creative captioning* problem. Because of this, a fine tuned image captioning BLIP model was used.
- **Selecting card phase** Given the narrator caption, the player now has to select a card that best matches the description. For this task the obvious choice was to use a CLIP model, explicitly created to match image-text pairs.
- **Voting card phase** Given that also this phase involve an image-text pairs matching, the same CLIP model was used. The fine tuning details for this architecture will be later discussed.

Email: Stefano Saravalle <saravalle.1948684@studenti.uniroma1.it>.

Deep Learning and Applied AI 2024, Sapienza University of Rome, 2nd semester a.y. 2023/2024.

2. Related work

The first idea of using Dixit to introduce the concept of creative captioning in AI was presented by Kunda et al.(Kunda & Rabkina, 2020). Since then, a very small amount of articles have been produced regarding this task. One of the most notable was from Vatsakis et al.(Vatsakis et al., 2022): in their work they collected a dataset of Dixit games played on an online platform from July 2012 to September 2021. Then they used classical NLP algorithms, like TF-IDF, to produce an image-text match model, obtaining promising results. To improve this baseline, Wei(Wei, 2023) used a CLIP model on the Dixit games dataset to obtain better result than the classical NLP methods involved in previous works. Even if those articles present good attempts in many subtasks of Dixit, there are still some other to solve, like for example the *creative captioning* task per se. Other than that, all those models were never tested on real or artificial players, compromising the actual quality of the produced results.

3. Method

Obtained the dataset from the article of Vatsakis et al. the idea was then to fine tune an image captioning model and a CLIP model on it. The image captioning model was choosen to be BLIP, a transformer based architecture that achieved state of the art performance on various image-text tasks(Li et al., 2022). After an initial fine tuning

	1 GPT vs 4 Bots		2 GPT vs 3 Bots		3 GPT vs 2 Bots		4 GPT vs 1 Bot		Humans vs Bot	
	GPT	Bot	GPT	Bot	GPT	Bot	GPT	Bot	Humans	Bot
Voted by all	53.33	**10.00**	65.00	**10.56**	64.44	**12.50**	52.50	**16.67**	27.50	**15.00**
Voted by none	33.33	50.42	3.33	37.78	2.22	43.33	9.17	66.67	30.28	**15.67**
Voted by someone	13.33	**39.58**	31.67	**51.67**	33.33	**44.17**	38.33	16.67	42.22	**69.33**

Table 1. Average percentages of votes when that type of player was a narrator

of BLIP on the online Dixit games, the agent was then tested to check the actual performances. Unfortunatly, the model was constantly outputting gibberish or empty captions. This behaviour can be explained due to how humans think about a caption. Given that the population of people in the dataset is extremely heterogeneous (the website from which the games were collected was publicly accessible without any registration), the same images could led to thousand of different sentences wihtout any correlations between them. Also, given the particular time in which the game was played, there could be references to events in that specific time frame, regarding for example politics or pop culture, completely unrelated to other sentences produced in the past, like it can be seen in Figure 1. Because of the tweeting bird in the image, the user selected as caption "Twitter", the famous social network created in 2006, that changed name in "X" in 2023. For this reason, a new strategy was adopted: the idea was to extract descriptive captions with a simple BLIP model, pass those captions to an LLM to rephrase them in a vague and mysterious way, and finally fine tune both the BLIP and CLIP model on this new dataset. The LLM used for this task was Zephyr(Tunstall et al., 2023), chosen because open-source and pretty good in rephrasing the captions in a very indistinct way. Two different prompts were tested, and even if in this way two different dataset were produced, just the second one was used in the actual fine tuning because of the better quality of the rephrased captions.

3.1. BLIP fine tuning

The dataset obtained through the rephrasing was very imbalanced, having a thousand captions for image, for a total of 84 cards. For this reason, the actual fine tuning of BLIP was performed with the vision model frozen, to avoid overfitting the few pictures available.

3.2. CLIP fine tuning

For the CLIP model different strategies were adopted (frozen visual encoder, fine tuning only projection layers, weight decay, etc...) but the model was consistently overfitting. Because of this, the strategy was then to use the already fine tuned BLIP model to extract creative and vague captions from the COCO dataset, and fine tune CLIP on them. The idea was that in this way the model could learn to associate images and creative texts on more general photos, which could then be applied to Dixit. It's worth noting that the model was fine-tuned by freezing all parts except the projection layers to prevent overfitting. These layers were adjusted to better map the extracted image and text features into a space where similar pairs are closer and different pairs are more distant.

4. Results

To check the actual performances of the agent it was tested in 30 games: 25 against GPT and 5 against humans. The GPT games have been played always with 5 players with a different balance between GPT bots and custom agents, while instead the human games have been played with 3 to 4 players with very different skills in the game. As it can be seen from Table 1, the bot is able in almost every setting against GPT to be voted by someone more than GPT itselfs. Instead, against human, the narrator part of the bot is significantly stronger, being able to be voted by someone and less by no one or by everyone.

5. Discussion and conclusions

Given those results, the bot wasn't able to win consistently when the GPT bots where too many and achieved a second place in 4 out of the 5 games against humans. This can be explained by the weaker CLIP model: the majority of the game is played when not the narrator, and a consistent number of points are made by picking the right narrator card, while trying to trick everyone by selecting a good enough card. Other than that, Dixit is a game in which a little bit of randomness is involved, and not always the player has a good card to place for the caption given by the narrator. Also a lot of meta-mechanics are involved: in the only game in which the agent placed last, it was playing against expert players that knew the game and each other very well. Humans reason in a very different way, and choosing the best caption or best card is not a deterministic process, where there is always a best move, but it depends on the players you are playing with. In conclusion, the bot show good performances, but much work can be done to better the agent, in particular the CLIP model, for example, by collecting a good and cleaned dataset of real life games, or train the image captioning model with RLHF.

References

Kunda, M. and Rabkina, I. Creative captioning: An ai grand challenge based on the dixit board game, 2020. URL https://arxiv.org/abs/2010.00048.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. Zephyr: Direct distillation of lm alignment, 2023. URL https://arxiv.org/abs/2310.16944.

Vatsakis, D., Mavromoustakos-Blom, P., and Spronck, P. An internet-assisted dixit-playing ai. In *Proceedings of the 17th International Conference on the Foundations of Digital Games*, FDG '22, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450397957. doi: 10.1145/3555858.3555863. URL https://doi.org/10.1145/3555858.3555863.

Wei, R. Dixit player with open clip, 2023. URL https://www.scirp.org/pdf/jdaip_2023112814012413.pdf.