DLAI Project 12: Bot Playing ”Dixit” Board Game

January 17, 2025

Cristiana Di Tullio 1803880

Abstract

”Dixit” is a creative board game where players describe images in imaginative ways. In this project I implemented an agent capable of playing Dixit in both the roles of who gives the hint (storyteller) and who guesses (guesser). To reach this purpose I employed a CLIP-like model for multimodal processing and an LLM to filter and refine the textual outputs. Experiments to assess the model's performance were conducted against GPT-4o multimodal and humans in 15 games. The Github repository of this project is available at https://github.com/laacri/DLAI-MultimodalDixitBot.

1. Introduction

This project proposed a unique challenge: teach an agent to play a game that is inherently abstract and imaginative. Naturally, this ambitious goal didn't come alone but was accompanied by a number of technical and structural issues that started to arise as I started to explore models and tools at my disposal. Let's start with some background research.

Text input	Prob.
”a dog”	0.00963724
”a cat”	0.96125025
”a turtle”	0.00261977
”a bird”	0.00573432
”a rabbit”	0.02075837

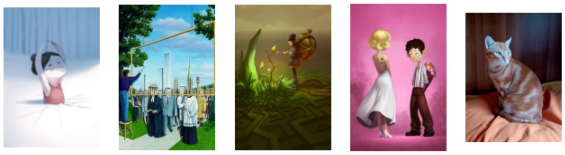
Figure 1. Meet my cat Ipa. Example of Image-to-Text retrieval given an input image and a set of possible captions.

The CLIP model. CLIP (Contrastive Language-Image Pre-Training) is a multimodal model trained on a variety of (image, text) pairs. By design, it learns to align input image and text in a shared embedding space; it leverages a contrastive loss function that brings close together the similar (image, text) pairs and pushes far apart the dissimilar ones. This way, the model efficiently learns visual concepts. The model outputs a similarity score based on the inputs features that can be turned into a probability through a *softmax* function. Thus, CLIP can be used to find the most relevant caption for an image, or vice-versa the most relevant image for a caption. Examples of CLIP's capabilities are shown in Figure 1 and Figure 2 respectively.

Email: Cristiana Di Tullio <ditullio.1803880@studenti.uniroma1.it>.

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

Input query: ”an image of a cute cat”



Card	n. 1	n. 2	n. 3	n. 4	n. 5
Prob	3.608e-2	3.041e-5	1.2058e-3	2.152e-4	9.6246e-1

Figure 2. Example of Text-to-Image retrieval given an input text caption and a set of possible images.

The LLM. Meta's LLama 3.2 with 3B parameters is a powerful and lightweight text-only Large Language Model freely available on HuggingFace. This model fits very well on-device and has demonstrated good generation capabilities that meet the needs of this project.

2. Related work

CLIP (Radford et al., 2021) has set a milestone in the increasingly popular context of multimodality. Since its release in 2021, this model has been fine-tuned and employed in various fields up to Remote Sensing, Medical Imaging and Autonomous Vehicles. Recently, improved general-purpose versions of the model were represented by EVA-CLIP (Sun et al., 2023) and SigLIP (Zhai et al., 2023), with the latter holding currently state-of-the-art performances. Despite the excitement, few steps have been taken in the field of creative captioning so far. Nevertheless, as (Kunda & Rabkina, 2020) propose it could become a valuable class of tasks for training powerful models in the future and Dixit itself might then become an important benchmark.

3. Methodology

The fundamental problems to deal with were two: first, CLIP-like models are not generative. To create a clue for a card, we need to feed the model some textual reference too. Second, Dixit is an imaginative game, so we need the model to be able to reason in abstract terms. The presented solutions to these problems are: first, craft a suitable text corpus that can be used to retrieve keywords associated to the card's content. Second, make the CLIP-like model interact with an LLM to handle the abstraction factor.

The Text Corpus. A suitable vocabulary of keywords is created combining a NLTK base corpus, CIFAR labels and preprocessed descriptions from a paintings dataset for a total of ≈ 1000 words.

Figure 3. Is an image worth just 1000 words? Wordcloud of the crafted text corpus based on a paintings dataset with descriptions.

Word	Prob.
cat	13.68%
striped	5.98%
pose	2.20%
copper	2.07%
brown	1.85%
orange	1.71%
sitting	1.56%
tiger	1.38%
fat	1.35%
tail	1.21%

Figure 4. Example of Zero-Shot classification with a sample image and class labels coming from our crafted text corpus.

Storyteller. Given an image, outputs a clue (Image-to-Text). It was implemented in steps: first, retrieve the most relevant 10 keywords with SigLIP from the text corpus; however, a clue made out just of these keywords will be very obvious, so use the properly prompted LLM to generate three creative candidate clues; lastly, select the best clue using again SigLIP to find the one with highest matching probability.

Guesser. Given a clue, guess the image (Text-to-Image). It was also implemented in steps: first, extract 10 possible keywords from the clue with the LLM to simplify the abstraction level; then, use SigLIP to select from a pool the card that best matches the keywords.

4. Results

In Figure 5 and Figure 6 are shown two example tests, for the Storyteller and the Guesser roles.

Storyteller's clue: "The last knight of the kingdom dies of old age."

Figure 5. Example of storyteller clues based on input images. In this case, the first image was the right one.

Human clue: "Ash to ash, dust to dust."

Figure 6. Example of human clue from which the model had to guess the right image. In this case, it correctly indicated the fourth one.

In general, the agent's performances can be defined poor: although it is able to guess the right image (if the clue isn't too ambiguous) and it can craft interesting clues (after a few trials), as soon as the level of abstraction gets higher it starts making mistakes and in most cases it will output complex - yet too clear clues. The experimental results about the agent's performance are summarized in Table 1 and Table 2. Consider that the optimal value for the proposed metrics is 50%. After all, the sense of playing Dixit is to make mistakes every now and then.

Table 1. Storyteller performance comparison.

vs.	Human	GPT-4o
Storyteller	0.27	0.07

Table 2. Guesser performance comparison.

vs.	Human	Storyteller
Guesser	0.20	0.60

5. Discussion and Conclusions

It's interesting that in the Storyteller tests GPT-4o scored an accuracy higher than humans. However, this can be explained with the human mind being much more imaginative, so much so, that sometimes it thinks of a connection even where there isn't one. That being said, there is a lot of room for improvement, starting from refining the text corpus to experimenting with different LLM prompts and models to trying different strategies for the clue generation, for example involving image segmentation.
For more details about the implementation and future work directions, please check the notebook.

References

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

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Sun, Q., Fang, Y., Wu, L., Wang, X., and Cao, Y. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11975–11986, 2023.