Hypernet based AGI Assessment

Yixin Huang, Shanghai Jiao Tong University

***Abstract*—AI-generated images (AGI) is a promising application of AI technology in computer vision field. However, a universal evaluation criteria hasn't been formulated. In this paper we introduce a well-rounded hypernet based AGI assessment model, which can evaluate AGI on both solely image-based features like image quality and image authenticity and features related to texts like text alignment. The proposed method is tested on 2 datasets, i.e. AGIQA-3K[2] and AIGCIQA2023[3]. The model outperforms the evaluation models included in the papers.**

***Index Terms*—AIG, Hypernet.**

I. Introduction

AI-Generated images (AGI) have recently been popular due to the rapid technological advancement of visual computing and networking. However, there still lacks an effective way to evaluate AGI's quality, bringing difficulty to the future development of both the image generator and the AGI identifier. Scores given by naive methods such as Inception Score or Frechet Inception Distance (FID) diverge significantly from humans' subjective perception. In recent years, deep learning based methods thrive and show better performance. In this paper, a hypernet based method is proposed to evaluate AGI on their quality and text alignment. Multiple other possible methods that aren't finally implemented due to GPU memory limitation are also listed to show author's thoughts on ways to evaluate AGIs.

II. Backbone Model Description

The backbone of the proposed model is the hypernet[1]. Traditional deep learning based AGI quality assessment models receive image related data and directly map it to a score. The procedure can be described as $\phi(\text{data}, \theta) = \text{score}$. However, from human experience, the criteria on image quality vary with the content and multiple other properties of the images and hence the parameter θ should be dependent on images and possible auxiliary data (like the prompts used to generate the image). The procedure is reformulated as follows:

$$\phi(\text{data}, \theta_{\text{data}}) = \text{score}$$

Specifically, the model could be separated into 2 neural networks, a hypernet and a target network. The hypernet uses image related data to generate weights and bias for the target network. The target network maps some specific AGI features to a quality-related scores. Following this scheme, a model able to evaluate AGI on their quality and text alignment could be built.

III. Image Quality Evaluation

Since [1]'s primary intuition is to evaluate image quality, experiments on it will firstly be carried out on the given AGIQA-3K and AIGCIQA2023 dataset. The best results on the 2 datasets listed in the paper is provided in table I for reference:

TABLE I
Best results on AGIQA-3K[2] and AIGCIQA2023[3] from papers

	quality(3K)	quality(2023)	authenticity(2023)
SRCC	0.8355	0.7961	0.6701
PLCC	0.8903	0.8402	0.6807

Since AIGCIQA2023 evaluates images on both their quality and authenticity while AGIQA-3K only focuses on images' quality, they will be discussed separately in following subsections. Generally, the hypernet takes the image as the input. A pretrained Resnet extracts image features from the images, which is used to generate parameters for the target network. The target network fits the features perceived by Resnet as input to a specific image quality related score.

A. AGIQA-3K

80% of the data is set as train data and the left 20% as test data. The hypernet model has a performance of SRCC 0.8603 and PLCC 0.9100 on test data, which is quite satisfying.

B. AIGCIQA2023

Given that both quality scores and authenticity scores are provided and that quality is highly correlated with human perceived authenticity, 3 sets of experiments are carried out on this dataset. One for quality score assessment, one for authenticity score assessment and the other evaluate both properties at the same time.

The output dimension of the last layer of the target network, which takes the weights and bias generated by the hypernet and outputs the estimated score of a given image, is changed to 2 to enable double outputs. The experiment result is given as follows, where T_qua stands for trained for quality, T_auth stands for trained for authenticity, T_both (qua) stands for the quality score generated by the model trained for both and T_both (auth) stands for the authenticity score generated by the model trained for both.

TABLE II
AIGCIQA2023 quality and authenticity evaluation result

	T_qua	T_auth	T_both (qua)	T_both (auth)
SRCC	0.8227	0.7423	0.8412	0.7827
PLCC	0.8312	0.7561	0.8576	0.7778

As we can conclude from table I, quality score and authenticity score are to some extent related to each other and training for the two scores at the same time boosts model performance.

C. Discussion

As we may conclude from Table II, the model performs worse on image authenticity evaluation. From my perspective, this could be attributed to the fact that the image quality depends solely on the image itself while the image authenticity is more correlated to human experience and human knowledge. Resnet could effectively image features like resolution and blurring by observing the image array distribution. However, whether this image is based on real scenario or not is more of subjective judgement. Extra data like some commonsense about real-world images should be inserted to make the model better aware of the criteria for authenticity. Nonetheless, the authenticity evaluation is enhanced by the company of quality evaluation. This is probably due to the fact that people tend to believe in the authenticity of the images if they are of high image quality. Hence, we can improve the performance of the model by training on both quality score and authenticity score at the same time on absence of the required commonsense dataset

IV. Text Alignment Evaluation

Text alignment refers to how the images match the texts used to generated those AGI and is a key factor when evaluating an image-generating AI. The best alignment scores from the papers are provided for future reference:

TABLE III
BEST ALIGNMENT SCORES ON AGIQA-3K[2] AND AIGCIQA2023[3] FROM PAPERS

	AGIQA-3K	AIGCIQA2023
SRCC	0.7472	0.7961
PLCC	0.7058	0.7153

A. Image based hypernet

As is stated before, different kinds of images correspond to different types of evaluation criteria. In this section we implement a hypernet that generates weights and bias based on images' characteristic.

As in [1], the hypernet first uses a pretrained Resnet to extract images' semantic features. However, distinct from [1], local characteristics extracted for quality evaluation are not used as input to the target network. Instead, the provided prompts are encoded to serve as the network's input. A model with a target network that takes both the extracted features and the prompts is implemented in appendix A for comparison.

I choose to implement the text encoder in the same way as in [4]. A pretrained BLIP model is utilized to encode prompts in an image-aware fashion. From another perspective, originally in [4], a simple MLP is implemented to fit the encoded text to text alignment score. Here I replace the MLP by a hypernet, which takes each image's individual characteristics into consideration. The experiment result is shown in Table IV.

We can draw the conclusion that this model evaluates AGI more accurately on both of the dataset. Also, the hypernet model outperforms the imagereward method significantly on AIGCIQA2023, which shows that the text alignment score is strongly correlated with the images' content in this dataset.

TABLE IV
IMAGE BASED HYPERNET BASED TEXT ALIGNMENT EVALUATION RESULT

	hyper(3K)	imgreward(3K)	hyper(2023)	imgreward(2023)
SRCC	0.8629	0.8521	0.7952	0.3562
PLCC	0.8458	0.8216	0.7203	0.3413

B. Text based hypernet

On the other hand, a more natural perspective on hypernet based evaluation metrics is to apply different criteria on images according to the prompts used to generate the AGI. However, this model should be trained on datasets with a large number of pictures with the same prompt. In AIGCIQA2023, we only can find 24 pictures under the same prompt and no more than 10 pictures of the same prompt in AGIQA-3K. Thus, the performance will not be satisfying on these 2 datasets. However, experiments are done to test the executability of the proposed model.

A key component of this network is an encoder to prompts. This encoder should be trained for encoding descriptive phrases for images, diverging from the common text embedders trained for queries or conversations. For simplicity, I adopt a pretrained phrase encoder introduced in [5].

As regards the network structure of the hypernet, now the input to the hypernet directly goes through the network to generate weights and bias for the target networks, skipping the Resnet. The input to the target network is still the image-aware encoded prompts generated by BLIP. The experiment result is listed as below.

TABLE V
TEXT BASED HYPERNET BASED TEXT ALIGNMENT EVALUATION RESULT

	AGIQA-3K	AIGCIQA2023
SRCC	0.1087	0.2025
PLCC	0.0923	0.1781

Though the model performs badly on these two datasets, considering the scarce data, the performance seems satisfying. A possible extension to this method is to leverage the GPT-model to predict the criteria for certain prompts. But due to time limitation, this idea is left for future work.

C. Discussion

The quality of the text encoder makes a great difference to the final performance of the model. I've also tested the model using encoders provided by [6]. [6] supports text embedding scenario like Q&A, long conversation, tool searching and so on. The model performs even worse than the model using [5], scoring images with negative SRCC and PLCC. The great performance gap between these 2 text embedders unveils the possibility of improving accuracy by developing tailored text embedders. The encoder should precisely extracts the subject,

subject's features, and many other characteristics from the prompts and evaluate prompts' similarity and difference in an appropriate manner.

V. Conclusion

In this paper, we engage the hypernet[1] to evaluate multiple quality metrics relevant to AGI. The proposed method outperforms the metrics presented in [2] and [3]. Further thoughts and complementary experiments are presented in the appendix.

Appendix-A hypernet with image semantics input

I've also attempted to make use of the image features extracted as input to aid text alignment evaluation. The results are as follows:

TABLE VI
IMAGE BASED HYPERNET BASED TEXT ALIGNMENT EVALUATION RESULT WITH IMAGE SEMANTICS INPUT

	AGIQA-3K	AIGCIQA2023
SRCC	0.6442	0.6026
PLCC	0.8038	0.5937

Apparently, the extracted image semantic features hinder the evaluation of text alignment. This is because that cases are that the extracted subject of the images may even be contrary to the real prompts, which makes the model even harder to evaluate the actual extent of text alignment.

Appendix-B Hybrid Evaluation

The last attempt is to take all image quality characteristics into consideration and evaluate for all. This brings a slight change to the model. The input of the hypernet is still the images but the input of the target net is a concatenation of image features extracted using Resnet and the encoded text. The structure of the target network is also changed to enable multiple outputs. The results are shown as follows:

TABLE VII
HYBRID HYPERNET EVALUATION RESULT ON AGIQA-3K

	Alignment	Quality
SRCC	0.6131	0.7942
PLCC	0.7998	0.8464

TABLE VIII
HYBRID HYPERNET EVALUATION RESULT ON AIGCIQA2023

	Quality	Authenticity	Alignment
SRCC	0.7118	0.6512	0.7734
PLCC	0.7146	0.6383	0.7668

Basically this hybrid model is inferior to the separate ones as regards performance. Not the more the data included, the better the performance. Actually, an image may score starkingly different on quality, authenticity and text alignment. If we can disentangle the image evaluation metrics more intricately and cluster those image characters that are highly correlated, we can train the model better by letting the model focus on a more concentrated area.

Appendix-C Previous Attempts

SRCC and PLCC are 2 commonly applied indicator of the AGI quality assessment accuracy. SRCC focuses on the rank of the scores given while PLCC also pays attention to the exact score given and the relationship between scores. In [4], it proposes a method that operates on paired data. It aims to maximize the score difference according to its original score rank. Namely, it serves to increase the model's SRCC. However, regarding PLCC, the case is slightly different. Scoring 2 images 0.8, 0.9 or o.1, 0.9 is different in PLCC since it also focuses on the exact values. Hence my intuition is simple, just to also take the value difference into the loss function. This is implementable because instead of the dataset used in [4], both AGIQA-3K and AIGCIQA2023 gives not only the performance rank, but also the performance scores.

The original loss used in image reward is:

$$\text{loss\_rank}(\theta) = -\mathbb{E}_{T,x_i,x_j}[\log(\sigma(f_\theta(T,x_i) - f_\theta(T,x_j)))]$$

where x_i is the image ranked higher than x_j and T is the prompt. Then we take the score into comparison and introduces the following loss:

$$\text{loss\_score}(\theta) = -\mathbb{E}_{T,x_i}[\log(\sigma(f_\theta(T,x_i) - s_{x_i}))]$$

where s_{x_i} represents the provided score for the image x_i.

By introducing extra loss which takes score into consideration, we can finally make full use of the data provided. Moreover, for calculation simplicity, this method could be used to finetune the model trained on merely rank data. This attempt failed because of limited GPU memory(only 3090 is available to me, but it could not even load the original ImageReward model).

References

[1] S. Su, et al., "Blindly Assess Image Quality in the Wild Guided by a Self-Adaptive Hyper Network," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020 pp. 3664-3673.

[2] Li, Chunyi & Zhang, Zicheng & Wu, Haoning & Sun, Wei & Min, Xiongkuo & Liu, Xiaohong & Zhai, Guangtao & Lin, Weisi. (2023). AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. IEEE Transactions on Circuits and Systems for Video Technology. PP. 1-1. 10.1109/TCSVT.2023.3319020.

[3] Wang, Jiarui & Duan, Huiyu & Liu, Jing & Chen, Shi & Min, Xiongkuo & Zhai, Guangtao. (2023). AIGCIQA2023: A Large-scale Image Quality Assessment Database for AI Generated Images: from the Perspectives of Quality, Authenticity and Correspondence.

[4] Xu, Jiazheng & Liu, Xiao & Wu, Yuchen & Tong, Yuxuan & Li, Qinkai & Ding, Ming & Tang, Jie & Dong, Yuxiao. (2023). ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation.

[5] Wu, Di & Yin, Da & Chang, Kai-Wei. (2023). KPEval: Towards Fine-grained Semantic-based Evaluation of Keyphrase Extraction and Generation Systems.

[6] Peitian Zhang and Shitao Xiao and Zheng Liu and Zhicheng Dou and Jian-Yun Nie. (2023). Retrieve Anything To Augment Large Language Models.