Comment generation using a large language model for fashion item recommendation

Yugen Sato*
Sayaka Aoba*
yugen_sato@cs.meiji.ac.jp
aoba_@cs.meiji.ac.jp
Department of Computer Science,
Meiji University
Kanagawa, Japan

Ryozo Masukawa
Makoto Sato††
Haji Shosuke‡‡
Tomohiro Takagi
Department of Computer Science,
Meiji University
Kanagawa, Japan

Taiga Matsui
Keita Ishikawa
airCloset,Inc.
Tokyo, Japan

ABSTRACT

In personal styling, the stylist selects fashion items considering various conditions such as the client's characteristics, purpose of use, and season, and comments specifically on the reasons for the selection, which is sent to the client along with the recommended fashion items. Simple phrases are effective in facilitating the understanding of fashion coordination, and characteristics of each item to be combined and compatibility of items are used to explain fashion coordination in a straightforward and correct manner. In order to provide a clear and correct explanation of fashion coordination, it is necessary to capture both the characteristics of each item to be combined and the compatibility of the items. Therefore, we define two tasks: generating a description of the coordination concept and generating a description of the dressing advice. For these tasks, we use MAGMA, a method that supports multimodal input of language models by means of adapter-based fine tuning, to build a system that generates fashion phrases and comments from a combination of item images and prompts. Quantitative and qualitative evaluations of our models show that they are more accurate than conventional baseline methods and comparable to an actual fashion expert.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Image representations**.

KEYWORDS

Vision and Language, Natural language generation, Multimodal processing

*Both authors contributed equally to this research.
†Currently with DeNA Co., Ltd.
‡Currently with Recruit Co., Ltd.

ACM Multimedia Asia (MMAsia'23), December 6–8, 2023, Tainan,Taiwan, 15.00

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Yugen Sato, Sayaka Aoba, Ryozo Masukawa, Makoto Sato, Haji Shosuke, Tomohiro Takagi, Taiga Matsui, and Keita Ishikawa. 2023. Comment generation using a large language model for fashion item recommendation. In *Proceedings of ACM Multimedia Asia (MMAsia'23), December 6–8, 2023, Tainan,Taiwan.* ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Large-scale models, as well as multimodal processing that handles both images and language, have been the focus of increasing attention and research. The purpose of our study is to generate comments in the fashion domain using multimodal large-scale models. In personal styling, a stylist selects fashion items considering various conditions such as customer characteristics, purpose of use, and season, and specifically comments on the reasons for the selection, which is sent to the customer along with the recommended fashion items. Simple phrases are effective in facilitating the understanding of fashion coordination. In order to explain fashion coordination in a straightforward and correct manner, it is necessary to capture both the characteristics of the individual items to be combined and the compatibility of the items. Therefore, we define two tasks: generating a description of the coordination concept and generating dressing advice. For these tasks, we used MAGMA, a method that supports multimodal input of language models by means of adapter-based fine tuning. We developed two systems, one for generating a description of the coordination concept from the combination of item images and prompts, and one for generating suggestions on how to wear an item from the combination of item images and prompts. Fine tuning a large-scale model is usually expensive, but by using adapters (described below in 3.4), the cost can be reduced and high accuracy can be achieved. In addition, since tuning is performed on image-text prompt pairs, it is possible to control the generation of prompts depending on their content. In particular, in generating dressing advice (4.2), our model generates comments with respect to the category of the point of interest specified in the prompt. Our model has been evaluated quantitatively and qualitatively, and we have verified that it is more accurate than conventional baseline methods and comparable to an actual fashion expert.

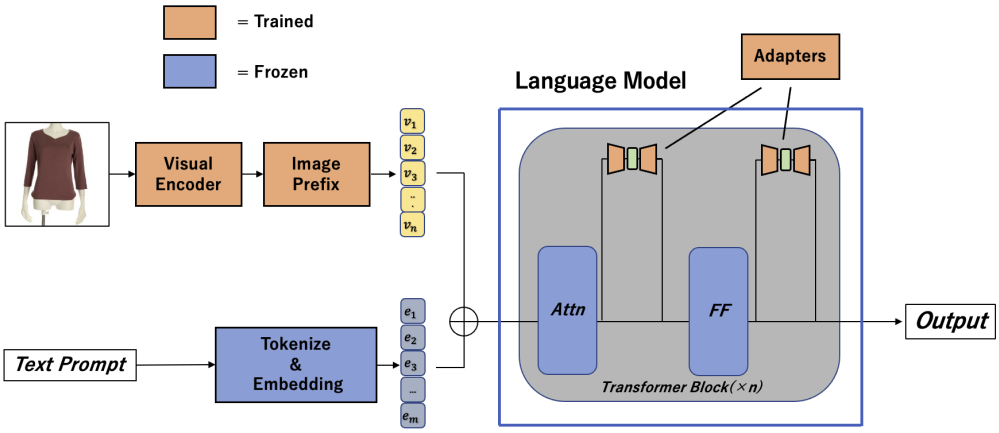

Figure 1: MAGMA's architecture

2 RELATED WORKS

2.1 Fashion intelligence system: An outfit interpretation utilizing images and rich abstract tags

Fashion style is difficult to evaluate quantitatively. For example, when judging the quality of coordination or the direction of fashion (taste), ambiguous expressions unique to fashion are often used in the absence of indicators. The Fashion Intelligence System [9] maps a full-body coordination image and tag information attached to the image in the same space, and by utilizing the coordinates of the image and tags in this space (=embedded representation), the system can understand ambiguous fashion expressions and accurately responds to user inquiries.

2.2 Transfer Learning Analysis of Fashion Image Captioning Systems

Filippo's study [1] deals with caption generation for fashion images using deep learning. We propose a novel Transformer-based approach to generating text from images and metadata, and as part of this approach, we combine the vision transformer [3] and BERT [2] encoders to leverage the generative performance of GPT models. Specifically, we utilize a contrast-trained image encoder and text encoder for fashion image data and metadata pairs, i.e., vision transformer and BERT, to encode the image and metadata of the fashion item for which the caption is to be generated. The image and text features are input as hidden states to the GPT-based decoder multi-head attention to inject fashion domain knowledge into GPT. Fine-tuning was performed on the Fashion-gen dataset, showing its superiority over existing methods.

2.3 Masked Vision-language Transformer in Fashion

Ji et al. [5] proposes a Masked Vision Language Transformer (MVLT) for fashion-specific multimodal representations. By replacing BERT in the pre-trained model with the vision transformer, MVLT became the first end-to-end framework in the fashion field. MVLT can accept raw multimodal input without additional preprocessing models such as ResNet, and is an extensible and convenient architecture for implicitly modeling Vision Language alignments. It is an extensible and convenient architecture for implicitly modeling Vision Language alignments. In addition, MVLT can be easily generalized to a variety of matching and generation tasks.

3 MAGMA ARCHITECTURE

Multimodal Augmentation of Generative Models through adapter-based finetuning (MAGMA) [4] is a language generation model for multimodal input that achieves high accuracy in open-ended generation tasks and OKVQA. The architecture of MAGMA consists of four main modules (Figure 1). Our study briefly introduces its components.

3.1 Visual Encoder

The visual encoder is a module that processes input images. It extracts semantic information about images by using the encoder part of CLIP [6], a model that has been pre-trained to bring image and text features close together in the same space. The encoder part of CLIP is used to extract semantic information about the image.

3.2 Image Prefix

The image prefix is responsible for connecting the output of visual encoder to the input of the language model (described below in 3.3), which maps the image features to a series of embedded vectors. And linearly transforms embedded vectors to the hidden dimension d_h of the language model, which is the output of the image features.

3.3 Language Model

As with GPT [7], the language model module uses an autoregressive model based on the Transformer. Therefore, it is possible to transfer the weights from the trained GPT. [4] uses weights from GPT-J [11], which has 6 billion parameters, but GPT-J is essentially an English model and is not expected to have much processing power for Japanese. Therefore, we used weights from a Japanese GPT model [8] with 336M parameters.

This time, we have combined a simple design front tucked blouse with wide tucked pants with a D-can belt for a beautiful casual coordination. How do you feel?

It also looks great as an inner layer of a jacket or cardigan. It can be worn with wide pants for an adult casual look, or with a flared skirt for a more feminine look. It might be nice to coordinate your accessories in silver, or enjoy wearing flat shoes made of suede.

Figure 2: Comments and target item images

3.4 Adapters

Adapters are a series of modules placed between the elements of Transformer that can be used to fine-tune the model weights instead of the model weights as an efficient method of parameter fine-tuning. The GPT weights of the language model are not changed during fine tuning, and the parameters of the adapters module are trained to perform the costly task of fine tuning a large model with fewer resources.

4 TASK DEFINITION

As described in Chapter 1, our study deals with the task of generating comments for fashion recommendations. The comment consists of two elements: an explanation of the coordination concept shown on the left of Figure 2 and an outfit advisory comment shown on the right. The explanation of the coordination concept is text data created when recommending a combination of two fashion items (one top and one bottom) selected by the stylist to the user as a single coordinated outfit. It serves to deepen the user's understanding of the recommended coordination by using fashion-specific expressions that express the atmosphere of the coordination. Meanwhile, the dressing advice contains advice information that helps the user to find items that go well with the recommended items and provides more detailed information about the recommended items. To achieve the purpose of this system, we define two tasks: generating a description of the coordination concept (the blue part in Figure 2) and generating dressing advice (the green part in Figure 2).

4.1 Task 1. Generating a description of the coordination concept

The goal of this task is to generate a description of the coordination concept, i.e., a combination of two fashion items, which represents a mood. The input features are images and text prompts of the two fashion items that make up the coordination. However, since only one image can be input to MAGMA, it is necessary to be creative when dealing with two images as in this case. Therefore, we took advantage of the fact that the two images are of a top and a bottom and combined them vertically to treat them as a single coordinated image.

4.2 Task 2. Generating dressing advice

The goal of this task is to generate advice that will help the user find items that go well with the recommended items. Unlike Task 1, the input features are a single top image and a text prompt. The content of the advisory comments varies and can be classified into four categories: tops, bottoms, shoes, and accessories. We aim to control the target category of advice comments generated by the text prompts and to generate comments flexibly with a single model. Specific prompts are discussed in 6.2.

5 OUR MODELS

We constructed two models, one for each task; examples of their operation are shown in Figure 3. The left side of Figure 3 shows a model that handles Task 1, inputting an image of a coordination concept to be sent to a customer and Generating a description of the coordination concept. The right side of Figure 3 shows the model for Task 2, which asks the customer what accessories go with the top shown in the image and generates dressing advice.

6 FINE-TUNING

6.1 Dataset

The dataset used in this study is part of the past comment data sent by actual stylists to their clients. We analyzed this comment data in detail and extracted and used only the parts that mentioned items that matched the recommended items. In addition, personalized comments based on customer information and comments related to seasons and trends were removed. Finally, about 1 million comments were used for both generating a description of the coordination concept and generating dressing advisory comments, which were then used for fine tuning.

6.2 Text prompt

MAGMA is finetuned using the dataset described in the previous section. as described in 4.2, in the Generating advisory comments task, the content of the output comments changes depending on the category specified in the text prompt along with the image of the item. For example, if the category is bottoms, a comment such as "bottoms like __ go well with the item in the input image" is output. In Task 1, the image data used for finetuning is the item image, the caption is a description of the coordination concept for the item, and the prompt is a template-based "Q:{top description} and {bottom description} together, what kind of coordination do you get?". In Task 2, the image data used for tuning is the item image, the caption is the comment data for the item, and the prompt is the template-based "Q: What {category} goes with {item name}?"

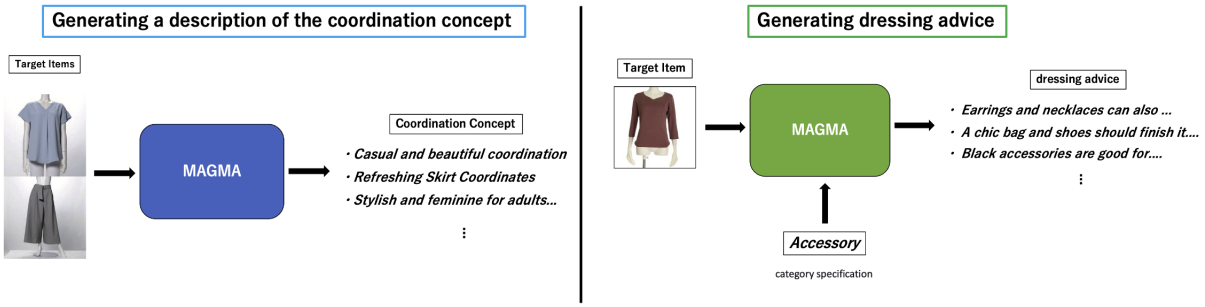

Figure 3: Example of our model in operation: generating a description of the coordination concept (Task 1) on the left and generating dressing advice (Task 2) on the right.

Fine-tuning was performed with these image, caption, and prompt combinations.

6.3 Learning and Loss

In training, we do not update the weights of the Language Model but optimize the weights of the Visual Encoder, Image Prefix, and adapters modules. The Language Model module is initialized with the weights of the pre-trained Japanese GPT model, and the Visual Encoder uses the weights of the pre-trained CLIP. The image prefixes and adapters are always trained from 0. In the following, the trainable parameters are denoted by subscript θ. The training task is image captioning, which is the task of generating plausible captions for input images, and next word prediction based on image features. Given an image-caption pair (x, y), each embedded representation is obtained as follows.

$$v_{1,\theta}, ..., v_{n,\theta} = {V^p}_\theta \otimes {V^e}_\theta(x) \tag{1}$$

$$e_1, ..., e_m = E(t_1), ..., E(t_m) \tag{2}$$

In Equations (1) and (2), V^p : ImagePrefix vector, V^e : Visual Encoder vector, t_k : tokenize caption. The length of the image sequence n is fixed, but the length of the caption m is variable. The vectors embedded in this way are concatenated and passed to the Language Model module, which calculates the loss according to Equation (3).

$$L_\theta(x, y) = -\sum_{i=1}^{m} l_\theta(v_{1,\theta}, ..., v_{n,\theta}, e_1, ..., e_m) \tag{3}$$

In Equation (3), l_θ is computed as $H \otimes \tilde{T}_\theta$ using the vector H, which maps the Transformer output of the Language Model module into token space, and the vector $\tilde{T}$ of the Transformer model with embedded adapters.

7 EXPERIMENTS

In this chapter, we examine the accuracy of our models and the baseline from both quantitative and qualitative perspectives. Experiments are conducted in the context of a problem in which a stylist recommends items to a customer in a personal styling session of a fashion rental service, and the customer is asked to create a recommendation comment to be added to the item.

7.1 Generation example

In this section, we specifically discuss how our two models generate a description of the coordination concepts and dressing advice for Task 1 and Task 2, respectively. As for Task 1, Table 1 shows an example of generating coordinating concepts for the input shown in Figure 4. From this table, we can verify that we were able to generate phrases of similar form but with different semantic orientations. As for Task 2, Table 2 shows two generated dressing advice sentences for each category for a fitted heart-neck knit (Figure 5) as an example. The generated examples show that the advice is accurately separated by category, and the wording is stable. The evaluation is based on these generated results.

7.2 Evaluation of a description of the coordination concept

7.2.1 *Baseline.*

As a baseline for comparison with our model, we use a 188-label classification model that combines the trained vision transformer and Efficient-Net [10] with only two images (one top image and one bottom image) as input. The labels are created from the comment data attached to the delivery of the two items, and are specifically words that describe the coordination, such as "beautiful" and "feminine". The actual phrases included in the comments are detailed phrases taken from the original comment data, such as "beautiful casual coordination" and "ladylike coordination", which are detailed phrases taken from the original comment data.

7.2.2 *Quantitative evaluation.*

In evaluating the proposed model, we used ROUGE, which is used to evaluate summarization tasks, BLEU, which is used to evaluate translation tasks, and BERTScore, which utilizes BERT. The results are shown in Table 3. The boldface letters indicate the highest accuracy in each evaluation index.

7.2.3 *Qualitative evaluation.*

It is difficult to evaluate the experimental results by quantitative evaluation alone. Due to the subjective nature of fashion and the inherent problem of the generation task that there is no single correct answer. Therefore, in this study, we conducted a qualitative evaluation with actual stylists. We randomly arranged the fashion item images and candidates from our model and baseline for each category in three sentences and scored them from 0 to 3 points

Figure 4: Target items

Table 1: Example of Fashion phrase generation

Prompt	Example
Q: How would a simple front pleated blouse and tuck wide pants with a sash belt coordinate? A: (ja. Q: シンプルなデザインのフロントタックブラウスとDカンベルト付きのタックワイドパンツ合わせると、 どんなコーディネートになりますか? A:)	neat casual style (ja.きれいめカジュアルなコーディネート)
	Fresh looking skirt style (ja.爽やかなスカートコーディネート)
	Relaxed outfit (ja.気負わない雰囲気のコーディネート)
	Stylish and not-too-swat look (ja.スタイリッシュな大人フェミニンコーディネート)
	neat skirt style (ja.きれいめなスカートコーディネート)

Figure 5: Target item

Table 2: Generation example of our model

Category	Examples
Top	It looks great as an inner layer for jackets and cardigans. (ja.ジャケットやカーディガンのインナーにも映えますね!)
	Wear the knit tucked in a flared skirt in navy, black, or gray for a feminine look! (ja.ニットはウエストインしてネイビーやブラック，グレーなどのフレアスカートを合わせると，女性らしい着こなしに!)
Bottom	For the bottoms, you can wear wide pants for an adult casual look or a flared skirt for a more feminine look. (ja. ボトムスには，ワイドパンツを合わせて大人カジュアルに, フレアスカートを合わせて女性らしい着こなしも素敵ですよ^^)
	We also recommend putting it together with a plaid skirt or lace skirt for a glamorous look! (ja. チェック柄のスカートやレースのスカートに合わせて華やかに まとめるのもオススメです!)
Shoes	Round toe pumps are recommended to soften the look of the feet. (ja. 靴はラウンドトゥパンプスで足元に抜け感を出すと◎)
	It is easy to match with short boots in griege or dark brown. (ja. 足元には，グレージュ，こげ茶のショートブーツにも合わせやすいですよ．)
Accessory	It looks great worn as it is, but it would also look great with a delicate necklace to add a touch of class. (ja. そのままサラリと着てもサマになりますが，デコルテに華奢な デザインのネックレスを合わせてクラスアップしても素敵ですね．)
	You can also pair it with silver accessories and flat suede shoes on your feet. (ja. アクセサリーはシルバーで統一して，足元はスエード素材のフラットシューズで， お楽しみいただくのも良いですね．)

in each category from two perspectives: fashion evaluation and grammar evaluation. These two perspectives enable us to determine the appropriateness of the generated content in the domain of fashion, as well as the evaluation of the structure, grammar, and wording. In addition, the most suitable sentence in each category is given 1 point as the optimal description of the coordination concept. Using this method, 15 coordinates were evaluated by three stylists, the average scores of which are shown in Table 4. In Table 4, Fashion means Fashion rating average, Grammer means Grammer rating average and Suitable means average of cumulative the most suitable points.

7.3 Evaluation of dressing advice

7.3.1 Baseline.

As a baseline for comparison with our model, we utilize a multi-label classification model based on contrastive learning using trained BERT and vision transformer. The model takes an item image and its comments as input and outputs a score for each label. Based on the scores, the closest comments are taken from the dataset in Chapter 6 and used as candidate comments.

7.3.2 Quantitative evaluation.

As with the generation of coordinating expressions, the evaluation is performed with ROUGE, BERTScore, and BLEU. The results are shown in Table 3.

Table 3: Quantitative evaluation results

Task	Model	ROUGE-1	BLEU	BERTScore
Generating a description of the coordination concept	baseline	0.00	0.265	0.654
	ours	**0.192**	**1.037**	**0.786**
Generating dressing advice	baseline	0.433	12.57	0.759
	ours	**0.483**	**17.91**	**0.821**

Table 4: Qualitative evaluation results

Task	Model	Fashion	Grammar	Suitable
Generating a description of the coordination concept	baseline	1.256	**2.422**	3.54
	ours	**1.991**	2.404	**10.78**
Generating dressing advice	baseline	2.197	**2.164**	**12.0**
	ours	2.245	2.116	**12.0**
	human	**2.311**	2.147	9.67

7.3.3 Qualitative evaluation.
As with generating a description of the coordination concept, an evaluation is performed by a stylist to measure the appropriateness of the content of the dressing advice. In this task, the stylist evaluates three kinds of results: our model and the baseline, plus the correct data. The format of the evaluation is roughly equivalent to 7.2, differing only in that the size of the test data is 10. The results are shown in Table 4.

7.4 Results and Consideration

In Table 3 and 4, boldface type indicates the highest value in that column. Table 3 shows that our model quantitatively outperforms baseline in all indices for both Task 1 and Task 2. Our model outperforms both ROUGE and BLEU, which focus on common words without considering the meaning of the words, and BERTScore, which considers the meaning of the words. The qualitative evaluation results in Table 4 show that our model outperforms baseline in both tasks, although the correct answer data is the highest in Task 2 in terms of fashion evaluation. Considering these results and the results of the quantitative evaluation, our model is able to generate comments with slightly different wording while targeting words and vocabulary similar to those in the correct answers. However, the baseline method produced the highest results for both Task 1 and Task 2 in terms of Grammar evaluation. The baseline method does not generate output comments but quotes them from the dataset, so the comments are manually generated. Therefore, the baseline outperforms our model, which generates sentences from scratch, in terms of sentence structure, wording, etc. The cumulative average of the most suitable points for Task 2 in Table 4 shows that our model and the baseline both outperformed the correct answer data. Taking into account the fact that the correct answer data is the highest in the fashion evaluation, the baseline and our model have a range of examples with high and low evaluations.

8 CONCLUSION

We have developed models for generating fashion comments. We succeeded in generating a description of the coordination concept and dressing advice using adapter-based fine-tuning of a large language model and constructed a system with sufficient capability for practical use. Generating a description of the coordination concept using the proposed model demonstrated its effectiveness over the baseline method, and our model was shown to be comparable to comments given by an actual stylist. We believe that the system's value will be further enhanced if it is able to learn user information and time series information.

REFERENCES

[1] Filippo Colombo. 2022. Transfer Learning Analysis of Fashion Image Captioning Systems. https://www.politesi.polimi.it/bitstream/10589/187559/5/2022_04_Colombo_02.pdf

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* (2021).

[4] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. 2022. MAGMA – Multimodal Augmentation of Generative Models through Adapter-based Finetuning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2416–2428. https://aclanthology.org/2022.findings-emnlp.179

[5] Ge-Peng Ji, Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Christos Sakaridis, and Luc Van Gool. 2023. Masked Vision-language Transformer in Fashion. *Machine Intelligence Research* 20, 3 (2023), 421–434. https://doi.org/10.1007/s11633-022-1394-4

[6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

[7] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. (2018).

[8] rinnaCo.Ltd. 2022. rinna/japanese-gpt2-medium. https://huggingface.co/rinna/japanese-gpt2-medium.

[9] Ryotaro Shimizu, Yuki Saito, Megumi Matsutani, and Masayuki Goto. 2023. Fashion intelligence system: An outfit interpretation utilizing images and rich abstract tags. *Expert Systems with Applications* 213 (2023), 119167. https://doi.org/10.1016/j.eswa.2022.119167

[10] Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 6105–6114. https://proceedings.mlr.press/v97/tan19a.html

[11] Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model. https://github.com/kingoflolz/mesh-transformer-jax.