Visual Metaphor Generation based on Similarities in Appearance and Concept

Dan WANG[†], Ryota MIBAYASHI[†], and Hiroaki OHSHIMA[†]

† Graduate School of Information Science, University of Hyogo
8–2–1 Gakuen-nishimachi, Nishi-ku, Kobe, Hyogo 651–2197, Japan
E-mail: †ad21h036@gsis.u-hyogo.ac.jp, ††threeforest8@gmail.com, †††ohshima@ai.u-hyogo.ac.jp

Abstract In this paper we propose a framework for automatically generating visual metaphors based on similarities on visual and conceptual features. We put the idea into practice for generation of visual metaphors for fashion goods. For example, for a red jacket with material with good heat retention, the visual metaphor generated would be *flame*. *Flame* and product have same visual features of color and are consistent in the concept of "giving a warm feeling". In order to explore the patterns of establishment of visual metaphors based on human views, we build visual metaphor dataset for fashion goods manually, which contains pairs of fashion goods images and metaphor images. Based on this dataset, we propose visual metaphor generation methods based on color, shape, and textural similarity. Finally, we evaluate visual metaphors in terms of both appearance and concept, then discuss the uses of generating catchphrases for fashion goods based on visual metaphors.

Key words feature extraction, image recognition, clustering, search model

1. Introduction

Visual metaphor, which is mainly discussed in this study, is a metaphor that is established based on visual similarity. *The girl's face is as red as an apple* is an example of visual metaphor. This visual metaphor is valid because the girl's reddened face in shyness looks like a round, red *apple* from the appearance. In addition to the similarity in appearance, the girl's face and *apple* also have conceptual similarity. We often say that a girl's face is cute, and the adjective "cute" is also used to describe an *apple*. This is the conceptual similarity between a girl's face and an *apple*.

This paper attempts to challenge the generation of visual metaphors based on visual similarity and conceptual similarity. We try to propose a methodology on visual metaphor generation that can be applied to anything, so that relevant researchers can refer to our framework and implement visual metaphor generation in various fields. In this paper, we take fashion goods as the research object. We aim to automatically construct visual metaphors based on product images and descriptions. Figure 1 shows examples of input and output of this study. A bright and shiny red handbag with a product description text that emphasizes a premium feel. The visual metaphor generated for this item is the word *ruby*. *Ruby* is a very appropriate visual metaphor for this product, because the shape, color, texture and other characteristics of the product are consistent with the general impression of *ruby*. When imagining a *ruby*, each person may have a different image in mind, but the color and texture of the *ruby* are common to all. Therefore, when this product is compared to a *ruby*, it resonates with anyone. The purpose of this study is to automatically generate such visual metaphors based on the images and descriptions of the goods that can be understood by the public.

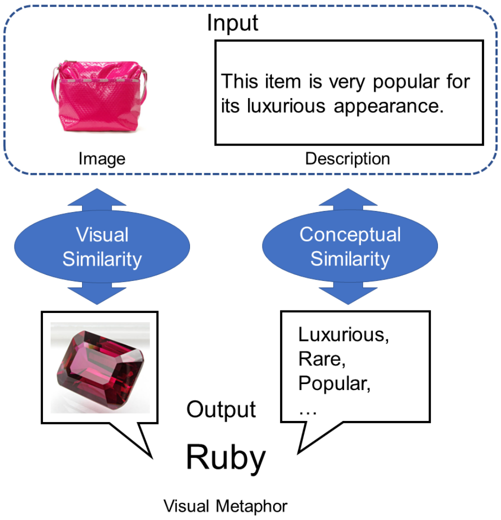


Figure 1 Example of input and output for this study

Using visual metaphors to express the characteristics of fashion goods can achieve a very good effect to consumers in product promotion. However, it is not an easy task for designers to find a appropriate visual metaphor for a fashion good. It takes a lot of imagination and knowledge to conjure up a visual metaphor that resonates with the public and reflects the visual features of a fashion good well. Therefore, we propose the method of visual metaphor generation to solve this difficult task and to support the understanding of visual metaphors and their application in various fields.

2. Related Work

We analyzed the related research for three aspects: (a) studies of visual metaphors in advertising, (b) studies about metaphor datasets. (c) studies of visual metaphor construction in various fields based on deep learning.

2. 1. Visual Metaphors in Advertising

Visual metaphors are often used in advertising because of the visual patterns and conceptual meanings they carry to enhance the persuasiveness of the message [10] [1]. Forceville's research on visual metaphors cites many forms of visual metaphors used in advertising [7]. Among them, visual metaphors are presented through various forms of media such as cartoons, comics, and posters. The promotional effect of using food as a visual metaphor for non-food brands was also highly evaluated [13]. Due to cultural and linguistic differences, using textual advertising in an international environment brings limited promotional effect, while pictorial advertising can be simply understood by people in different countries [3]. The positive effect of visual metaphors on advertising has also been mentioned in many recent studies [4] [20].

2. 2. Metaphor Dataset

The proposed metaphor datasets in previous studies were classified into three categories: textual, image, and text-image categories. A dataset that can be used for discriminating whether a syntactic construction is meant literally or metaphorically using lexical semantic features of the words that participate in the construction is proposed [17]. the VU Amsterdam Metaphor Corpus (VUAMC) [15], TroFi Example Base [2], and MOH-X [12]. MultiMET is a dataset including both text and images to facilitate understanding metaphorical information [18].

2. 3. Visual Metaphor Construction

Visual metaphor construction has been studied in a variety of different fields. Model that can distinguish metaphorical street art images from other rhetorical figures is suggested [14]. There is also method that can understand the meaningful patterns conveyed in artworks for artwork clustering [6]. There are also studies about uses of visual metaphor based on deep learning in advertising. There is a framework proposed for ads understanding. The framework considers both the visual and textual elements of the ad in order to properly understand the metaphorical message in the ad [19]. Ideation support tool is also proposed to inspire visual metaphor ideas and for streamlining development through multi-dimensional example exploration [11].

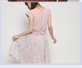
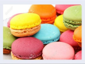
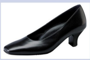
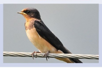
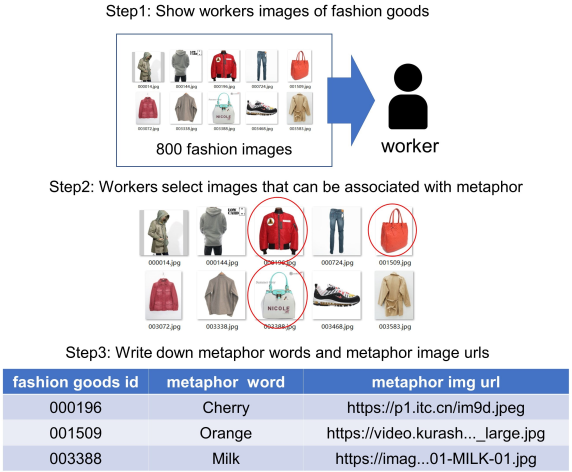
fashion img	metaphor word	metaphor img
	Moon	
	Chocolate	
	Ruby	
	Macaron	
	Swallow	

Figure 2 Fashion goods visual metaphor dataset


Step1: Show workers images of fashion goods
800 fashion images
worker
Step2: Workers select images that can be associated with metaphor
Step3: Write down metaphor words and metaphor image urls

fashion goods id	metaphor word	metaphor img url
000196	Cherry	https://p1.itc.cn/im9d.jpeg
001509	Orange	https://video.kurash..._large.jpg
003388	Milk	https://imag...01-MILK-01.jpg

Figure 3 Data collection process for visual metaphor dataset

3. Visual Metaphor Dataset

Fashion goods visual metaphor dataset defined in this study contains paired data of fashion goods and metaphors. As shown in Figure 2, dataset has three columns, which are fashion image, metaphor word, metaphor image.

3. 1. Data Collection

Data of visual metaphor dataset is collected manually. Worker is shown a large number of images of fashion goods and then provides metaphor words and metaphor images. The flow chart of data collection is shown in Figure 3. There were 15 workers that participated in data collection. Each worker was assigned a folder containing 800 fashion images and an answer sheet. Fashion images assigned to each worker were different. Workers selected the images that they could associate with metaphor words and recorded their ids, metaphor words and metaphor images on the answer sheet.

We suggest the following instructions for associative

Table 1　Visual metaphors collected based on color similarity

fashion image	metaphor image	metaphor word
		grape
		marshmallow
		sky
		cucumber salad

metaphor words: (a) Metaphor word can be anything other than something related to fashion (b) Metaphor word should have a general and specific impression. Workers were asked to associate metaphor words based on the above two instructions. Metaphor image is the image of metaphor that is consistent with the impression that comes to mind when workers look at a fashion goods. We asked workers to use image search engines to find metaphor images and to record image urls. We gave the following two instructions for workers to collect metaphor images: (a) Metaphor images can be in any form, including photographs, illustrations, paintings, and so on. (b) Metaphor images should be without any impurity such as logos, watermarks, and so on.

The data source for fashion images is from Rakuten Ichiba Dataset [注1]. We select items from the categories of women's fashion, men's fashion, shoes, bags, and hats as the images of the products that we show to the workers during data collection.

3. 2. Data Study

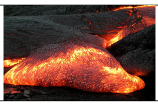
We collected 2,362 pairs of metaphor images and fashion images, and 1,049 words of visual metaphors. We divided it into training data and validation data in the ratio of 8:2. Then, we study the training data to explore the conditions under which the human viewpoint-based visual metaphor holds.

A typical visual similarity pattern that we first observed from the training set was color. As shown in Table 1, goods and metaphors have similarity in color or pattern. *Grapes* and *marshmallow* in the table are the simplest color-based visual metaphors. Both the goods and the metaphors are consistent in a single color. Color similarity is not only reflected in the same color, but also in the arrangement (pattern) of colors, as in the example *sky* and *cucumber salad* in the table. In this case, the goods corresponding to the *sky* are consistent in both blue and white colors. The goods corresponding to the *cucumber salad* have multiple colors, and even if the colors are not perfectly consistent, the consistency of the pattern can lead people to assume that the visual metaphor is valid. This similarity is so strongly persuasive that there are many examples in the dataset where other visual features can be ignored as long as color consistency holds. We can see the visual metaphor pairs shown in the table, which do not hold true in terms of similarity of shape or texture.

Table 2　Visual metaphors collected based on shape similarity

fashion goods	visual metaphor	metaphor words
		brownie
		rice cracker
		egg
		UFO

Table 3　Visual metaphors collected based on texture similarity

fashion goods	visual metaphor	metaphor words
		cotton
		magma
		gold bar
		macaroon

Beyond color, we found that people focused on the similarity in shape between the goods and the metaphor (Table 2). Shape-based visual metaphors sometimes hold without relying on color similarity (*UFO* and *egg* shown in the table), while sometimes they require both color similarity to hold (*brownie* and *rice cracker* shown in the table).

Another similar pattern is texture-based (Table 3). Fashion goods have a variety of textures due to different materials, ranging from warm-looking plush, to shiny, premium materials, to silky leather. Texture-based visual metaphors are considered to best characterize fashion goods and bring excitement to consumers. Texture-based visual metaphors often rely on similarities in color. Inconsistencies in color can make textural similarities difficult to understand.

We collected 1,049 metaphor words as visual metaphor

[注1]：`https://rit.rakuten.com/data_release/`

Table 4 Metaphor words

category	metaphor words
natural	sea, sky, fire, sun, universe, starry sky, forest, desert, snow, rainbow, red leaf, water
jewel	ruby, emerald, sapphire, pearl, gold, cinnabar, silver, lapis lazuli, opal, obsidian, alexandrite
sweets	chocolate, caramel, marshmallow, ice cream, whipped cream, chiffon cake, chocolate cake
animal	horse, dolphin, rabbits, penguins, butterflies, swans, wolf, shark, tiger, swallowtail butterfly
drink	milk coffee, milk chocolate, soda float, cream soda, chocolate shake, orange milk
flowers	cherry blossom, rose, carnation, poppies, tulips, marigolds, hydrangea, lilies, rape blossoms
fruits	lemon, strawberry, peach, banana, apple, orange, dragon fruit, raspberries, plums, pears, muscat

candidates. Table 4 shows some examples of metaphor candidates.

4. Visual Metaphor Generation

In this section we illustrate the method of extracting visual and conceptual features from fashion goods and metaphor candidates, then describe how to integrate the two similarities to finally output a visual metaphor words ranking for fashion goods.

4. 1. Visual Feature Extraction

In Section 3. 2, we mentioned that similarities on color, shape, texture dominate the similar patterns in the visual metaphor dataset, so we build three classifiers based on Inception V4 (pre-trained by ImageNet) that be able to understand basic visual feature of various objects [16] [8]. We define 16 labels for color, 9 labels for shape and 7 labels for texture, which are shown in Figure 4. We used each label as a query and collected more than 300 images for each label by Bing image search API. These collected images are used as for training data as shown in Figure 4. We remove the fully connected layer of classifiers to obtain feature networks that be able to extract color, shape, texture feature vector of images.

Further, we perform a second fine-tuning based on this feature model. The purpose of this fine-tuning is to enable the model to associate metaphors according to human thinking. Training data that used for second fine-tuning is visual metaphor dataset that mentioned in Section 3. We first split the 1,889 pieces of training data into a training set and a validation set by 8:2, and then construct separate negative sets based on them. We first draw out all the fashion images in the dataset and re-match each image with a different metaphor image than the original one. The metaphor image is randomly selected from all metaphor images in the visual metaphor dataset. Using this method we obtain a negative set of the same size as positive set. The network we choose is Siamese network based on Inception V4. Siamese network is constructed as shown in Figure 5, whose input is a pair of images that are fed into the same base network to obtain the feature vectors, respectively. We calculate the cosine similarity of these two vectors and pass Mean Squared Error for back propagation.

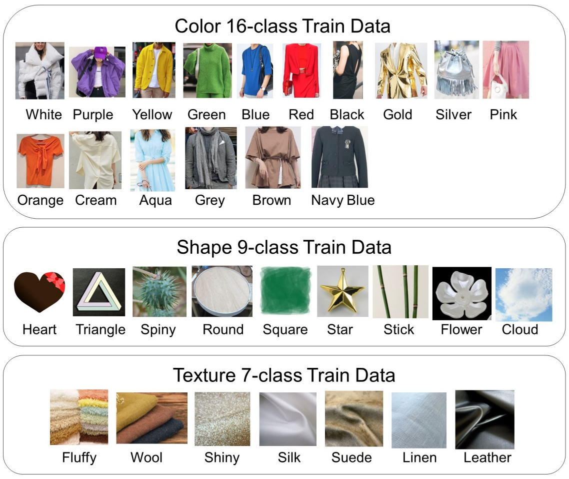


Figure 4 Training data for color, shape, texture classifiers

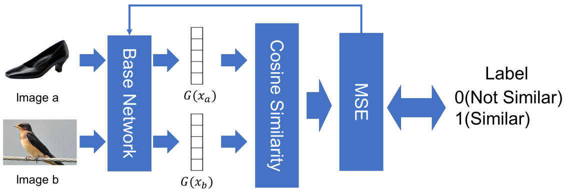


Figure 5 Siamese network

Following two fine-tuning, we obtained three visual feature extraction models. The visual feature extraction of fashion goods is to input the image of fashion good into the visual feature extraction model and output visual feature vector of fashion goods. As for visual feature extraction of metaphor word, we use a method that referring to DeepCluster [5]. First, we collected 40 images for each metaphor word using Bing Image Search. For each metaphor word, We use three visual feature models to get the feature vectors of the 40 images. We perform K-means clustering on shuffled 40 feature vectors. The number of clusters is fixed to 6. The images closest to the center points is selected as representative impressions of visual metaphors. The visual features of the metaphorical words are obtained by averaging 6 feature vectors of representative images. The number of clusters and the mean-based integration method of visual feature we chose were determined based on validation of the metaphor-generated performance. The process of valida-

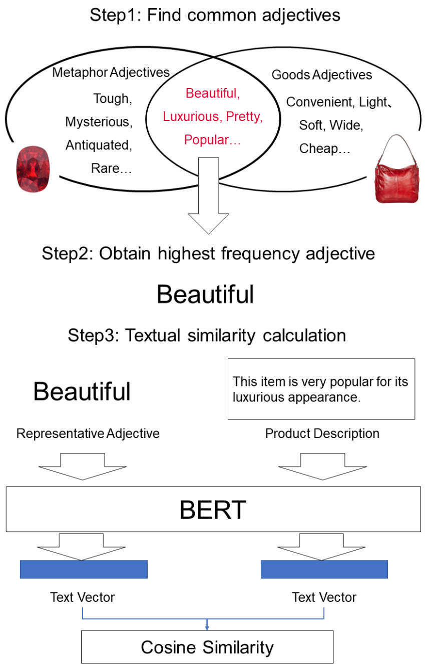


Figure 6 Process of conceptual similarity calculation

tion is to evaluate MMR (Mean Reciprocal Rank) of visual metaphor generation on test set (mentioned in Section 3. 2) for different number of clusters (3, 6, 9) and two feature integration methods (mean and maximum).

4. 2. Conceptual Feature Extraction

Calculation of the conceptual similarity between fashion goods and visual metaphor is shown in Figure 6. First, we collected adjectives of fashion goods and metaphor words from web articles. For fashion goods, we used their category names (e.g., sweater, sandals) as search terms and collected articles using Bing web search API. For metaphors, we use metaphor words as search terms. For each term, we obtained 1,000 articles and extract their adjectives. We select the adjective with highest frequency among adjectives shared by fashion goods and visual metaphor to calculate similarity with product description. The model we used to calculate the text similarity is Japanese pre-learning BERT model from Tohoku University [9]. Representative adjective of metaphor word and product description are transformed into text vectors, and their cosine similarity is calculated to obtain the conceptual similarity.

4. 3. Metaphorical Appropriateness Computation

An overview of the visual metaphor generation method is shown in Figure 7. Metaphorical appropriateness of visual metaphors and fashion goods consists of two parts: visual similarity and conceptual similarity. Based on different treatments of color, shape, and texture similarities, there are different approach to obtain metaphorical appropriateness. We will illustrate three methods of visual metaphor generation in this section. which are based on single visual similarity, double visual similarity, and three visual similarities.

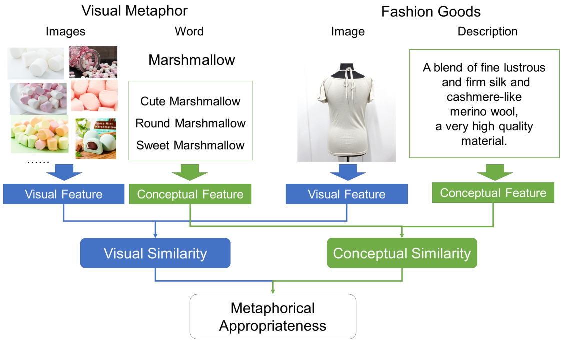


Figure 7 Overview of Metaphorical Appropriateness C

Visual metaphor generation based on single visual feature means that one of the visual features of color, shape or texture is emphasized in visual feature extraction and the visual similarity of the fashion goods and all visual metaphors candidates is calculated using the extracted visual feature vector.

Visual feature extraction based on two visual features means combining three independent visual features two by two to get three methods of color-shape, shape-texture and color-texture respectively. First, we obtain the visual similarity of fashion goods and all visual metaphors candidates based on single visual features. Then the two visual similarities are combined with the following formula. Sim_a and Sim_b are any two visual similarities in color, shape, and texture, respectively. The double visual similarity is determined by two values, one is the average of Sim_a and Sim_b. The second is the maximum value of Sim_a and Sim_b.

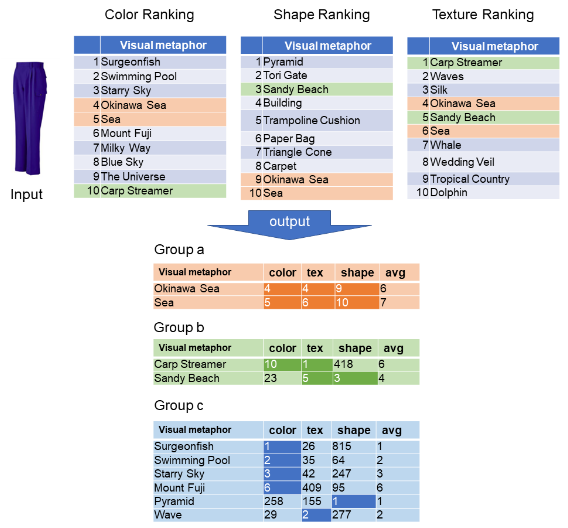
$$Sim = \sqrt{\frac{Sim_a + Sim_b}{2} \max(Sim_a, Sim_b)}$$

All visual metaphors were ranked by the similarity calculated from the above formula to obtain the visual metaphors of fashion goods based on the method of double visual similarity.

Method based on three visual similarity is to combine three visual similarities to present all visual metaphors in one ranking. Process of combination is shown in Figure 8. Firstly, We output single visual similarity rankings based on color, shape, texture. Then separate all metaphor words into three groups: (a) visual metaphors that appear on all three rankings, (b) visual metaphors that appear on two rankings, and (c) visual metaphors that appear on one ranking only. We sort each group by the mean of the ranks of each metaphor word.

5. Evaluation

The manual-based evaluation consists of two aspects: the



Color Ranking

	Visual metaphor
1	Surgeonfish
2	Swimming Pool
3	Starry Sky
4	Okinawa Sea
5	Sea
6	Mount Fuji
7	Milky Way
8	Blue Sky
9	The Universe
10	Carp Streamer

Shape Ranking

	Visual metaphor
1	Pyramid
2	Tori Gate
3	Sandy Beach
4	Building
5	Trampoline Cushion
6	Paper Bag
7	Triangle Cone
8	Carpet
9	Okinawa Sea
10	Sea

Texture Ranking

	Visual metaphor
1	Carp Streamer
2	Waves
3	Silk
4	Okinawa Sea
5	Sandy Beach
6	Sea
7	Whale
8	Wedding Veil
9	Tropical Country
10	Dolphin

Group a

Visual metaphor	color	tex	shape	avg
Okinawa Sea	4	4	9	6
Sea	5	6	10	7

Group b

Visual metaphor	color	tex	shape	avg
Carp Streamer	10	1	418	6
Sandy Beach	23	5	3	4

Group c

Visual metaphor	color	tex	shape	avg
Surgeonfish	1	26	815	1
Swimming Pool	2	35	64	2
Starry Sky	3	42	247	3
Mount Fuji	6	409	95	6
Pyramid	258	155	1	1
Wave	29	2	277	2

Figure 8 The process of visual metaphor generation based on three visual features

appearance aspect and the conceptual aspect. In appearance aspect, we evaluate the metaphor generation methods based on single similarity, double similarity, and three similarities in Section 4. 3. For conceptual aspect, we examined the effect of the different importance of conceptual similarity on visual metaphor generation. We introduced a parameter α to regulate the degree of visual and conceptual contribution to the establishment of metaphors. As shown in the following equation, M is the metaphorical appropriateness, Sim_v and Sim_c are the visual similarity and conceptual similarity. We generated metaphors for each item based on three α ($\alpha = 0.7$, $\alpha = 0.8$, $\alpha = 0.9$) methods.

$$M = Sim_v^{\alpha} Sim_c^{1-\alpha}$$

We selected 20 items with distinctive features in color, shape, and texture from the test set of the visual metaphor dataset for manual evaluation. We take top 10 ranked metaphors generated by each method as the result of each method and present shuffled results to workers and asked them to answer following two questions.

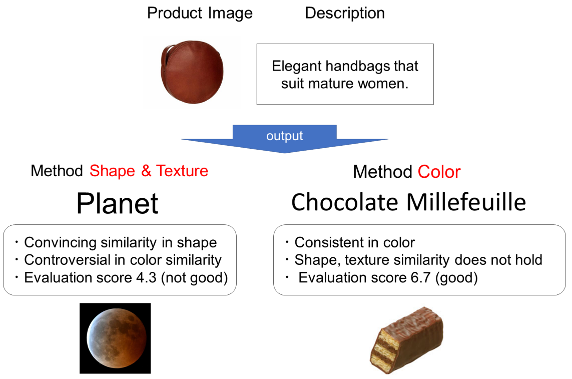
- Look at the product image and the visual metaphor. Would you say that this visual metaphor is valid for this **product image**?
- Look at the product image and the visual metaphor. Would you say that this visual metaphor is valid for this **product description**?

For each question, workers were asked to answer a number from 1 to 10, which indicates 10 levels of satisfaction (1 being the worst and 10 being the best). We averaged the scores answered by the three workers as the final rating value.

Table 5 shows results of appearance aspect. We can learn that people are most concerned with color similarity when judging the validity of visual metaphor. As shown in Figure 9, the visual metaphor based on shape and texture and the visual metaphor based on a single visual similarity of color are *planet* and *chocolate millefeuille*. We can observe that the product has a typical round shape as well as *planet*, which proves that *planet* and product have a very convincing similarity in shape. However, the similarity in color is doubtful, because the impressions of *planet* in terms of color are diverse. The less convincing color similarity between *planet* and commodities led workers to score the output low.

Table 5 Evaluation results of different methods for appearance aspect

method		best score	avg score
single-sim	color	6.4	4.4
	shape	5.7	3.0
	texture	5.9	3.2
double-sim	color&shape	6.4	4.2
	color&texture	6.4	4.2
	shape&texture	6.1	3.2
three-sim	color&shape&texture	6.4	3.8



Elegant handbags that suit mature women.

Method Shape & Texture

Planet

- Convincing similarity in shape
- Controversial in color similarity
- Evaluation score 4.3 (not good)

Method Color

Chocolate Millefeuille

- Consistent in color
- Shape, texture similarity does not hold
- Evaluation score 6.7 (good)

Figure 9 Color similarity is evaluated more than shape and texture

Table 6 shows performance of visual metaphors generated based on different importance degrees of conceptual similarity. It is concluded that the variation of α from 0.7 to 0.9 did not affect the performance of visual metaphors in reflecting descriptions. We use an example shown in Figure 10 to explore the reason behind this result. We assume that the workers interpreted the feature of "stylized design" as individuality and unexpectedness when they understood the description of this product. The design of product remind people that it is a performance costume, so *stage* is the closest to the product in terms of function, but it does not reflect individuality and unexpectedness. Although *wedding veil* is more unexpected and surprising than *stage*, its impression is more considered as a symbol of tenderness than highlighting individuality. While *pearl shell*, with its shiny appearance and rarity, gives it a flashy impression and fits the individ-

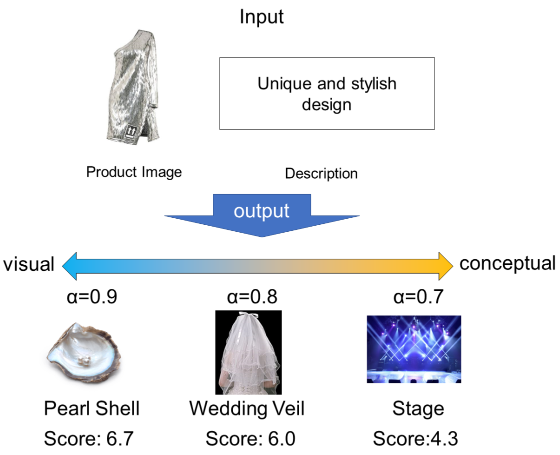


Figure 10 Visual metaphors obtained based on different conceptual similarity weights

Table 6 Evaluation results of different methods for conceptual aspect

	$\alpha = 0.9$	$\alpha = 0.8$	$\alpha = 0.7$
best score	6.4	6.4	6.3
avg score	3.2	3.2	3.2

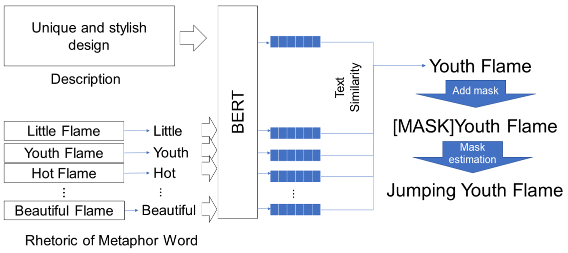


Figure 11 Process of catchphrase generation

uality part of the product description, all three metaphors have a positive impression in terms of appearance, but *pearl shell* is the one that best highlights the unexpectedness.

6. Catchphrase Generation based on Visual Metaphor

The process of generating catchphrases is shown in Figure 11. We complete the catchphrase by adding rhetoric in front of the metaphor word. The first step is to select the candidate most similar to the product description from the common rhetorics of the metaphor word and form a phrase. The second step is to add mask in front of this phrase to generate the second rhetoric by BERT mask prediction to complete catchphrase.

We collect rhetorics candidates of metaphor words by Bing web search API. We build queries for each metaphor word to get web articles and extract rhetorics from them. Query is built by prefixing metaphor word with a cue character. For one metaphor word, We use 4 cue characters in Japanese (“ta”, “no”, “na”, “i”) to build 4 queries. “Ta” is auxiliary verb in Japanese, which is used behind verb. “No” is auxiliary that connects noun to the metaphor word, which can be regarded as “of” in English. “I” is word used in Japanese to follow an adjective. “Na” is auxiliary verb that connect adjective and metaphor word, which can be regarded as “ful” at the end of an English adjective. We searched 1,000 results for each of query, collected 25 to 110 rhetoric candidates for each metaphor word. We obtained rhetoric with the highest similarity to product description based on BERT. Then use this rhetoric as first word in rhetoric part of catchphrase. We add a random cue character and `[MASK]` for mask estimation to complete catchphrase.

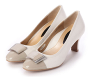
Product image	Product description	Catchphrase
	Cute colors create a soft atmosphere.	Thick colorful *cream*
		Baked whitish *roll cake*
		Calm comfortable *bedding*
	Warm and gentle comfort.	Energetic tender *teddy bear*
		Energetic tender *baby kangaroo*
		Energetic tender *poodle*
	Soft material with elasticity.	Fresh delicious *sesame pudding*
		Fresh tasty *cafe au lait*
		Smooth soft *sphynx cat*

Figure 12 Examples of catchphrase generation

Figure 12 shows examples of catchphrase generation for three items. We took three metaphor words from the top 10 of the visual metaphors generated for each product as the metaphorical part of the catchphrase, and then generated rhetoric based on them to complete the catchphrase.

The first example in Figure 12 is a pink and white knitted sweater with a gentle touch. Catchphrases “thick colorful cream” and “baked whitish roll cake” contain adjectives related to color because the color scheme of the product is highlighted in the product description. In addition, “calm comfortable bedding” focuses on the part of the product description that emphasizes the feel of the clothes, highlighting the softness of the product. In the second example in Figure 12, the metaphor words are all animals, and they share the common rhetoric “Energetic tender”. Since the first rhetoric is the most similar rhetoric to the product description among rhetoric candidates of the metaphor word, the repeated word is output when the word is also present in the rhetoric candidates of other metaphor words. The third example is a pair of light teal shoes with stretchy material and soft to wear. *Sesame pudding*-like shoes will make people associate the material of the shoes with the taste of pudding, and the smooth texture of the shoe material on the product picture also has some similarity with pudding. The other catchphrase, *cafe au lait*, not only expresses the smooth and delicate texture of the shoes, but also has a similar color to the shoes, which is visually more persuasive. The last catchphrase is the most surprising, because *Sphinx cat* does have a soft body, fully reflecting the description of the material in the shoe description. Moreover, *Sphinx cat* indeed has a similar color to the shoes. The rhetoric "smooth soft" further draws attention to the Sphinx's consistent characteristics with the product. For this example, without the modifier, it would have been difficult to capture the similarity between *Sphinx cat* and the product, as they are hardly identical in appearance alone. Catchphrases increase the volume of information that the visual metaphor can convey, so we can say that the visual metaphor and catchphrases have a complementary effect.

7. Conclusion

In this paper, we construct a visual metaphor dataset of fashion goods manually. We summarize three metaphor establishment patterns of color, shape, and texture, based on which we construct three feature extraction models to implement visual metaphor generation. Based on the conceptual similarity between visual metaphors and fashion goods, we discuss the use of visual metaphors in generating catchphrases. There are two improvement points in this paper that are worth exploring in the future. One is to increase the number of evaluation participants to analyze performance of visual metaphors more comprehensively. The other is to achieve a variety of results even for simple-looking products.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP21H03775, JP21H03774, JP21H03554, JP22H03905. In this paper, we used "Rakuten Dataset" (`https://rit.rakuten.com/data_release/`) provided by Rakuten Group, Inc. via IDR Dataset Service of National Institute of Informatics.

References

[1] Swee Hoon Ang and Elison Ai Ching Lim. The influence of metaphors and product type on brand personality perceptions and attitudes. *Journal of Advertising*, 35:39–53, 2006.

[2] Julia Birke and Anoop Sarkar. A clustering approach for nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 329–336, 2006.

[3] Sandy Bulmer and Margo Buchanan-Oliver. Visual rhetoric and global advertising imagery. *Journal of Marketing Communications*, 12:49–61, 2006.

[4] Shuo Cao, Huili Wang, and Xiaoxia Zou. The effect of visual structure of pictorial metaphors on advertisement attitudes. *International Journal of Marketing Studies*, 10:60–72, 2018.

[5] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision 2018*, pages 139–156, 2018.

[6] Giovanna Castellano and Gennaro Vessio. A deep learning approach to clustering visual arts. *International Journal of Computer Vision*, 130:2590–2605, 2022.

[7] Forceville Charles. Pictorial metaphor in advertisements. *Metaphor and Symbolic Activity*, 9:1–29, 1994.

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 8.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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*, pages 4171–4186, 2019.

[10] Se-Hoon Jeong. Visual metaphor in advertising: Is the persuasive effect attributable to visual argumentation or metaphorical rhetoric? *Journal of Marketing Communications*, 14:59–73, 2008.

[11] Youwen Kang, Zhida Sun, Sitong Wang, Zeyu Huang, Ziming Wu, and Xiaojuan Ma. MetaMap: Supporting visual metaphor ideation through multi-dimensional example-based exploration. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

[12] Saif Mohammad, Ekaterina Shutova, and Peter Turney. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 23–33, 2016.

[13] Boujena Othman, Ulrich Isabelle, Piris Yolande, and Chicheportiche Laëtitia. Using food pictorial metaphor in the advertising of non-food brands: An exploratory investigation of consumer interpretation and affective response. *Journal of Retailing and Consumer Services*, 62, 2021.

[14] Georgios Stampoulidis and Marianna Bolognesi. Bringing metaphors back to the streets: A corpus-based study for the identification and interpretation of rhetorical figures in street art. *Visual Communication*, 2019.

[15] Gerard Steen, Lettie Dorst, J. Herrmann, Anna Kaal, Tina Krennmayr, and Trijntje Pasma. *A Method for Linguistic Metaphor Identification: From MIP to MIPVU.* 2010.

[16] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, Inception-ResNet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.

[17] Tsvetkov Yulia, Boytsov Leonid, Nyberg Eric, Gershman Anatole, and Dyer Chris. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258, 2014.

[18] Dongyu Zhang, Minghao Zhang, Heting Zhang, Liang Yang, and Hongfei Lin. MultiMET: A multimodal dataset for metaphor understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 3214–3225, 2021.

[19] Huaizheng Zhang, Yong Luo, Qiming Ai, Yonggang Wen, and Han Hu. Look, read and feel: Benchmarking ads understanding with multimodal multitask learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 430–438, 2020.

[20] Haoxing Zhao and Xiaoyong Lin. A review of the effect of visual metaphor on advertising response. In *Proceedings of the 4th International Conference on Economy*, pages 29–34, 2019.