MultiMET: A Multimodal Dataset for Metaphor Understanding

Dongyu Zhang[1], **Minghao Zhang**[1], **Heting Zhang**[1], **Liang Yang**[2], **Hongfei LIN**[2]
[1]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province,
School of Software, Dalian University of Technology, China
[2]School of Computer Science and Technology, Dalian University of Technology, China
hflin@dlut.edu.cn

Abstract

Metaphor involves not only a linguistic phenomenon, but also a cognitive phenomenon structuring human thought, which makes understanding it challenging. As a means of cognition, metaphor is rendered by more than texts alone, and multimodal information in which vision/audio content is integrated with the text can play an important role in expressing and understanding metaphor. However, previous metaphor processing and understanding has focused on texts, partly due to the unavailability of large-scale datasets with ground truth labels of multimodal metaphor. In this paper, we introduce MultiMET, a novel multimodal metaphor dataset to facilitate understanding metaphorical information from multimodal text and image. It contains 10,437 text-image pairs from a range of sources with multimodal annotations of the occurrence of metaphors, domain relations, sentiments metaphors convey, and author intents. MultiMET opens the door to automatic metaphor understanding by investigating multimodal cues and their interplay. Moreover, we propose a range of strong baselines and show the importance of combining multimodal cues for metaphor understanding. MultiMET will be released publicly for research.

1 Introduction

Metaphor is frequently employed in human language and its ubiquity in everyday communication has been established in empirical studies (Cameron, 2003; Steen, 2010; Shutova et al., 2010). Since Lakoff and Johnson (1980) introduced conceptual metaphor theory (CMT), metaphor has been regarded as not only a linguistic, but also a cognitive phenomenon for structuring human thought. Individuals use one usually concrete concept in metaphors to render another usually abstract one for reasoning and communication. For example, in the metaphorical utterance "knowledge is treasure," knowledge is viewed in terms of treasure to express that knowledge can be valuable. According to CMT, metaphor involves the mapping process by which a target domain is conceptualized or understood in terms of a source domain.

(a) A fire in the sky tonight. (b) Smoking causes lung cancer.

Figure 1: Examples of multimodal metaphor

As a means of cognition and communication, metaphor can occur in more modes than text alone. Multimodal information in which vision/audio content is integrated with the text can also contribute to metaphoric conceptualization (Forceville and Urios-Aparisi, 2009; Ventola et al., 2004). A multimodal metaphor is defined as a mapping of domains from different modes such as text and image, text and sound, or image and sound (Forceville and Urios-Aparisi, 2009). For example, in Figure 1 (a), the metaphorical message of fire in the sky is conveyed by a mapping between the target domain "sky" (sunset) and the source domain "fire" from two modalities. Figure 1 (b) offers another example with the metaphor of lungs made from cigarettes so a relation is triggered between two different entities, lung and cigarette, with the perceptual idea that smoking causes lung cancer. The source domain "cigarette" comes from the image, while the target domain "lung" appears in both text and image. Understanding multimodal metaphor

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
August 1–6, 2021.

requires decoding metaphorical messages and involves many cognitive efforts such as identifying the semantic relationship between two domains (Coulson and Van Petten, 2002; Yang et al., 2013), interpreting authorial intent from multimodal messages (Evan Nelson, 2008), analyzing the sentiment metaphors convey (Ervas, 2019), which might be difficult for computers to do.

Qualitative studies have investigated the interplay between different modes underlying the understanding of multimodal metaphors in communicative environments such as advertisements (Forceville et al., 2017; Urios-Aparisi, 2009), movies (Forceville, 2016; Kappelhoff and Müller, 2011), songs (Forceville and Urios-Aparisi, 2009; Way and McKerrell, 2017), and cartoons (Refaie, 2003; Xiufeng, 2013). In particular, with the development of mass communication, texts nowadays are often combined with other modalities such as images and videos to achieve a vivid, appealing, persuasive, or aesthetic effect for the audience. This rapidly growing trend toward multimodality requires a shift to extend metaphor studies from monomodality to multimodality, as well as from theory-driven analysis to data-driven empirical testing for in-depth metaphor understanding.

Despite the potential and importance of multimodal information for metaphor research, there has been little work on the automatic understanding of multimodal metaphors. While a number of approaches to metaphor processing have been proposed with a focus on text in the NLP community (Shutova et al., 2010; Mohler et al., 2013; Jang et al., 2015, 2017; Shutova et al., 2017; Pramanick et al., 2018; Liu et al., 2020), multimodal metaphors have not received the full attention they deserve, partly due to the severe lack of multimodal metaphor datasets with their challenging and time- and labor-consuming creation.

To overcome the above limitations, we propose a novel multimodal metaphor dataset (MultiMET) consisting of text-image pairs (text and its corresponding image counterparts) manually annotated for metaphor understanding. MultiMET will expand metaphor understanding from monomodality to multimodality and help to improve the performance of automatic metaphor comprehension systems by investigating multimodal cues. Our main contributions are as follows:

- We create a novel multimodal dataset consisting of 10,437 text-image pair samples from a range of resources including social media (Twitter and Facebook), and advertisements. MultiMET will be released publicly for research.
- We present fine-grain manual multimodal annotations of the occurrence of metaphors, metaphor category, what sentiment metaphors evoke, and author intent. The quality control and agreement analyses for multiple annotators are described.
- We quantitatively show the role of textual and visual modalities for metaphor detection; whether and to what extent metaphor affects the distribution of sentiment and intention, which quantitatively explores the mechanism of multimodal metaphor.
- We propose three tasks to evaluate fine-grained multimodal metaphor understanding abilities, including metaphor detection, sentiment analysis, and intent detection in multimodal metaphor. A range of baselines with benchmark results are reported to show the potential and usefulness of the MultiMET for future research.

2 Related Work

2.1 Metaphor Datasets

Although datasets of multimodal metaphors are scarce, a variety of monomodal datasets for metaphor studies have been created in recent years. Table 1 lists these datasets with their properties.

Numerous text metaphor datasets have been published for metaphor processing in the NLP community including several popular ones, e.g., the VU Amsterdam Metaphor Corpus (VUAMC) (Steen, 2010), TroFi Example Base (Birke and Sarkar, 2006), and MOH-X (Mohammad et al., 2016). The largest one, VUAMC, consists of over 10,000 samples spread across 16,000 sentences, while others contain less than 5,000 samples. However, most existing metaphor datasets contain only textual data. Image metaphor datasets are few and they are pretty limited in the size and the scope of the data, such as VisMet (Steen, 2018), which is a visual metaphor online resource consisting of only 353 image samples. Although Shutova et al. (2016) constructed both text and image samples, their images were obtained by using a given phrase and queried Google

Metaphor Dataset	Sample Size (%Metaphor)	Modality	Data Source	Annotation
TroFi (Birke and Sarkar, 2006)	3,737 (44%)	Text	WSJ	metaphor (metaphoricity)
VUAMC (Steen, 2010)	16,000 (12.5%)	Text	BNC Baby	metaphor
TSV (Tsvetkov et al., 2014)	3,334 (50%)	Text	Web	metaphor, affect
LCC (Mohler et al., 2016)	16,265 (19%)	Text	ClueWeb09	metaphor
MOH (Mohammad et al., 2016)	1,639 (25%)	Text	WordNet	metaphor
Zayed's Tweets (Zayed et al., 2019)	2,500 (54%)	Text	Twitter	metaphor
Visual Met (Steen, 2018)	353 (100%)	Image	Adv, Arts, Cartoons	metaphor
Shutova et al. (2016)	2,415 (50%)	Text,Image	WordNet	metaphor
MultiMET (Ours)	**10,437 (58%)**	**Text,Image**	**Social Media, Adv**	**metaphor, sentiment, intent**

Table 1: Comparison of various metaphor datasets

images. In that way, words and images in their work may be not suitably presented by each other.

The cognitive nature of metaphor implies that not only one modal isolation, but rather integrated multimodal information may contribute to metaphor expression and understanding, which makes our dataset MultiMET, which is large scale and contains both natural text and image messages and their annotations, different from existing datasets and more important for metaphor studies.

2.2 Metaphor Understanding

Automatic metaphor understanding requires accomplishing certain tasks to decode metaphorical messages. In this paper, we focus on three important tasks for NLP in understanding metaphor: metaphor detection, sentiment analysis, and author intent detection. There has been increasing interest in NLP in various approaches to metaphor detection based on monomodal text. Early metaphor studies have focused on hand-constructed knowledge and machine learning techniques (Mason, 2004; Turney et al., 2011; Tsvetkov et al., 2014; Hovy et al., 2013). Others have also used distributional clustering (Shutova et al., 2013) and unsupervised approaches (Shutova et al., 2017; Mao et al., 2018). More recently, deep learning models have been explored to understand metaphor. However, little has been explored in multimodal metaphor detection except by Shutova et al. (2016), who are among the very few to explore the fusion of textual and image modalities to detect multimodal metaphor. Their results demonstrate the positive effect of combining textual and image features for metaphor detection.

However, in their work, image features are extracted from a small size of constructed examples rather than natural samples of texts integrated with images, like MultiMET in our work. In addition, apart from multimodal metaphor detection, the tasks related to metaphor understanding like sentiment detection and author intent detection in multimodal metaphor also have rarely been studied, although there exist similar multimodal studies in different tasks (Wang et al., 2017; Zadeh et al., 2017; Kruk et al., 2019).

3 The MultiMET Dataset

3.1 Data Collection

With the goal of creating a large-scale multimodal metaphor dataset to support research on understanding metaphors, we collect data that contains both text and image from a range of sources including social media (Twitter and Facebook), and advertisements. Table 2 shows an overview for the statistics of the dataset.

Social Media. To collect potential metaphorical samples from Twitter and Facebook, we retrieved posts by querying hashtags metaphor or metaphorical. We collected publicly available Twitter and Facebook posts using Twitter and Facebook APIs complying with Twitter and Facebook's terms of service. What the author labels as metaphorical is not always aligned with the actual definition of metaphor in our study. To collect metaphors whose nature accorded with what we define as multimodal metaphors, we re-annotated "metaphorical or literal" in the below section to potential Twitter and Facebook posts that other authors annotated as metaphor with hashtags.

Advertisements. Based on our review of linguistic literature on multimodal metaphor, we focused on an important source that is the main context of study: advertisements. Metaphorical messages abound in advertisements , which offer a natural and rich resource of data on metaphor and how textual and visual factors combine and interact (Sobrino, 2017; Forceville et al., 2017). We collected

(a) Life is a highway.

(b) Sometimes, without knowing why, your heart beats faster. New Beetle.

(c) A kitten is kissing a flower. Butterflies are not insects.

Figure 2: Examples of metaphor categories

Item	Social Media	Adv	Total
Total Samples	6,109	4,328	10,437
Metaphorical Samples	3,489	2,537	6,026
Literal Samples	2,620	1,791	4,411
Total Words	79,417	51,936	131,353
Avg Words of Samples	13	12	13
Train Set Size	2,791	2,029	4,820
Validation Set Size	349	254	603
Test Set Size	349	254	603

Table 2: MultiMET dataset statistics

potential metaphorical samples of advertising from a large, publicly released dataset of 64,832 image advertisements that contain both images and inside text (Ye et al., 2019). To obtain the textual information, we extracted inside text from images using the API provided by Baidu AI. After that, human annotators rectified the extracted inaccurate text, removed any blurred text, and obtained text + image pairs from advertisements.

3.2 Data Filter

For text data, we removed external links and mentions (@username); we removed non-English text using the LANGID (Lui and Baldwin, 2012) library to label each piece of data with a language tag; we removed strange symbols such as emojis; we removed "metaphor" or "metaphoric" when they were regular words rather than hashtags, because explicit metaphorical expressions are not our interest (e.g., "This metaphor is very appropriate"); we removed text with fewer than 3 words or more than 40 words. For image data, we removed text-based images (all the words are in the image), as well as images with low resolution. Because this task is about multimodal metaphor, it is necessary to maintain consistency of data between models. In other words, either both the image data and the text data should be removed, or neither. In addition, in the de-duplication step, we considered removal only when both text and images were repeated.

3.3 Annotation Model

We annotated the text-image pairs with the occurrence of metaphors (literal or metaphorical); (if metaphorical) relations of target and source domain (target/source: target/source vocabulary in text or verbalized target/source vocabulary in image); target/source modality (text, image, or text + image), metaphor category (text-dominant, image-dominant, or complementary); sentiment category (the sentiment metaphors evoke, namely very negative, negative, neutral, positive, or very positive), and author intents (descriptive, expressive, persuasive, or other). The annotation model was AnnotationModel = (Occurrence, Target, Source, TargetModality, SourceModality, MetaphorCategory, SentimentCategory, Intent, DataSource). Figure 3 is an annotation example.

Figure 3: An example of a text+image annotation

3.4 Metaphor Annotation

Metaphor category. There are a variety of ways in which texts and images are combined in multimodal content (Hendricks et al., 2016; Chen et al., 2017). Based on our review of the literature and observation of the samples in our dataset, we follow Tasić and Stamenković (2015) and divide multimodal metaphor into three categories: text dominant, image dominant, and complementary. Sometimes metaphors are expressed through texts with a mapping between source and target domains while the accompanying images serve as a visual illustration of the metaphors in the text, which is text dominant. As in Figure 2 (a), the text itself is sufficient to convey metaphorical information and can be identified as metaphorical expressions. "Highway" is a visual illustration of the source domain in a textual modality. By contrast, in the image dominant category, images play the dominant role in conveying metaphorical information and they provide sufficient information for readers to under-

stand the metaphors. In Figure 2 (b), where we see the metaphorical message "Beetle (cars) are blood cells," the text enriches the understanding of metaphorical meaning by adding an explanation "your heart beats faster" to the visual manifestation. The complementary category involves a roughly equal role of texts and images in rendering metaphorical information. The understanding of metaphor depends on the interaction of and balance between different modalities. If texts and images are interpreted separately, metaphors cannot be understood. In Figure 2 (c), when people read the text, "A kitten is kissing a flower," and the inside text "Butterflies are not insects," they do not realize the metaphorical use until they observe the butterfly in the corresponding image and infer that the target "butterfly" is expressed in term of the source "flower".

Metaphorical or literal. Our annotations focus on the dimension of expression, which involves identification of metaphorical and literal expressions by verbal means and visual means (Forceville, 1996; Phillips and McQuarrie, 2004). The metaphor annotation takes place at the relational level, which involves the identification of metaphorical relations between source and target domain expressions. For text modality, source and target domain expressions mean source and target domain words used in metaphorical texts. For image modality, source and target domain expressions mean words' verbalized source and target domain in the visual modality. That is, the annotation of metaphorical relations represented in the modality of image involve the verbalization of the metaphor's domains. Annotations involve naming and labeling what is linguistically familiar. Unlike text modality, which relies on explicit linguistic cues, for image modality, metaphorical relations are annotated based on perceptions of visual unities, and they determine the linguistic familiarity of images as well as existing words in the metaphor's domains. Following Šorm and Steen (2018), annotators identified the metaphorical text+image pairs by looking at the incongruous units and explaining one non-reversible "A is B" identity relation, where two domains were expressed by different modalites.

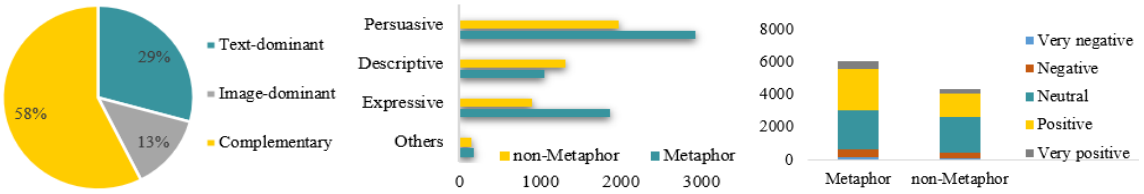
3.5 Intent and Sentiment Annotation

Interpreting authorial intent from multimodal messages in metaphor seems to be important for understanding metaphors. As mentioned above, within CMT, the essence of metaphor is using one thing from a source domain to express and describe another from a target domain. This implies that one important intent of creating metaphor could be to enable readers to understand the entities being described better. "Perceptual resemblance" is a major means of triggering a metaphorical relation between two different entities (Forceville and Urios-Aparisi, 2009). We name it descriptive intent, which involves visual and textual representations regarding the object, event, concept, information, action or character, etc. Moreover, in modern times, the increasing ubiquity of multimodal metaphors means that people cannot ignore its power of persuasion (Urios-Aparisi, 2009). People often leverage metaphor in communication environments such as advertisements and social media to persuade readers to buy or do things. We name this intent as persuasive. In addition, inspired by a variety of arousing, humorous, or aesthetic effects of metaphors (Christmann et al., 2011), the expressive is included in our intent annotation within the enlarged definition: expressing attitude, thought, emotion, feeling, attachment, etc. Based on these factors as well as investigation of the samples in our datasets, we generalized their taxonomy and listed the categories of the author intent in metaphor as descriptive,persuasive, expressive, and others.

Numerous studies show that metaphorical language frequently expresses sentiments or emotions implicitly (Goatly, 2007; Kövecses, 1995, 2003). Compared to literal expressions, metaphors elicit more emotional activation of the human brain in the same context (Citron and Goldberg, 2014). Thus we also added the sentiment in our annotation, to test whether the sentiment impact of metaphors is stronger than literary messages from a multimodal perspective. The sentiment was placed in one of the five categories of very negative, negative, neutral, positive, or very positive.

3.6 Annotation Process

We took two independent annotation approaches for two different types of tasks: selecting types of sentiment and intent and the annotation of metaphor. To select the options for sentiment and intent, we used a majority vote through CrowdFlower, the crowdsourcing platform. The participants were randomly presented with both the text and vision components with the instruction on the



(a) Metaphor Category. (b) Metaphor and non-Metaphor Intent. (c) Metaphor and non-Metaphor Sentiment.

Figure 4: Dataset Distribution

top of each text + image pair for options.

The annotation of metaphors includes metaphor occurrence, metaphor category and domain relation annotation. For metaphor annotation, we used expert annotators to complete the challenging annotation task, which required relatively deep understanding of metaphorical units and the complete task of verbalization of domains in image. The annotator team comprised five annotators who are postgraduate student researchers majoring in computational linguistics with metaphor study backgrounds. The annotators formed groups of two, plus one extra person. Using cross-validation, the two-member groups annotated, and the fifth person intervened if they disagreed.

3.7 Quality Control and Inner Agreement

Annotations of multimodal metaphors rely on annotators' opinions and introspection, which might be subjective. Thus we took corresponding, different measures for different types of annotations to achieve high-quality annotation. To select options, we established strict criteria for the choice of category. Each text-image pair was annotated by at least 10 annotators and we used a majority vote through CrowdFlower, the crowdsourcing platform. Following Shutova (2017), we chose the category of annotated options on which 70% or more annotators agreed as the answer to each question (final decision) to provide high confidence of annotation. For metaphor annotation, we added a guideline course, detailed instruction, and many samples, and we held regular meetings to discuss annotation problems and matters that needed attention. The guidelines changed three times when new problems emerged or good improvement methods were found. The kappa score, κ, was used to measure inter-annotator agreements (Fleiss, 1971). The agreement on the identification of literal or metaphorical was $\kappa = 0.67$; identification of text dominant, image dominant or complementary was $\kappa = 0.79$; the identification of source and target domain relation was $\kappa = 0.58$, which means they are substantially reliable.

4 Dataset Analysis

Metaphor Category. We analyzed the role of textual and visual modalities to detect metaphors. From Figure 4 (a), we can see a complementary category among the three kinds of multimodal metaphors, which requires the interplay of textual and visual modality to understand the metaphorical meaning. It accounts for the largest proportion of metaphors, followed by the text-dominant and image-dominant categories. It shows the contribution of visual factors, which are similarly important in detecting metaphors. We therefore present a quantitative study of the role of textual and visual modalities in metaphor detection through human annotations and confirm the role and contribution of visuals in metaphor occurrence in natural language.

Author Intent. Figure 4 (b) shows that expressive and persuasive intentions occur most frequently in the metaphorical data. However, descriptive intention occurs most frequently in the non-metaphorical data. This suggests that on the one hand, we are more likely to use metaphorical expressions when expressing our feelings, expressing emotions, or trying to persuade others. On the other hand, we tend to use literal expressions to make relatively objective statements.

Sentiment. Figure 4 (c) shows that there are some differences in the distribution of sentiment between the metaphorical data and non-metaphorical data. In the non-metaphorical data, neutral sentiment accounted for the largest proportion of 51%, followed by positive sentiment (33%), strong positive sentiment (7%), negative sentiment (7%), and strong negative sentiment (2%). In the metaphorical data, positive sentiment accounted for the largest proportion of 42%, followed by neutral sen-

Hyper-Parameter	Value
Word embedding size	300
Hidden size of LSTM	256
Dropout	0.4
Text padding	30
Batch size	48
Learning rate	5e-4
Gradient clipping	10
Early stop patience	10

Table 3: Hyperparameters.

timents (39%), strong positive sentiment (8%), negative sentiment (8%), and strong negative sentiment (3%). It turns out that there are more non-neutral sentiments in metaphor expression than in non-metaphorical expression, and that metaphors are more frequently used to convey sentiments. Our findings accord with the results of previous studies on monomodal textual metaphors that metaphors convey more sentiments or emotions than literary text (Mohammad et al., 2016). We confirm the stronger emotional impact of metaphors than literary messages from a multimodal perspective.

In positive sentiment, the most common words in the source domain are person, face, and flower; the most common words in the target domain are love, life, and success. In negative sentiment, heart, food, and smoke are the most common words in the source domain, and the world, disaster, and life are the most common words in the target domain. This shows that sentiment tendency can influence the category in the source and target domains to some extent.

5 Experiment

For the dataset constructed for this paper, we propose three tasks and provide their baselines, namely multimodal metaphor detection, multimodal metaphor sentiment analysis, and multimodal metaphor author intent detection.

We used the model shown in Figure 5 to detect metaphors, metaphorical sentiments, and metaphorical intentions. For text input, we used a text encoder to encode the text and to get the feature vector of the text. This paper used two different methods to encode the text, namely the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al., 2019) and Bidirectional Long-Short Term Memory (Bi-LSTM) networks (Medsker and Jain, 2001). Similarly, for image input, we used an image encoder to extract image features. We used three different image pre-training models: VGG16 (Simonyan and Zisserman, 2014), ResNet50 (He et al., 2016), and EfficientNet (Tan and Le, 2019). These methods have been widely used by researchers in feature extraction for various tasks.

After obtaining the text feature vector and the image feature vector, we used four different feature fusion methods to combine the vectors, namely concatenation (Suryawanshi et al., 2020), element-wise multiply (Mai et al., 2020), element-wise add (Cai et al., 2019), and maximum (Das, 2019). Finally, we inputted the fusion vector into a fully connected layer and obtained the probabilities of different categories through the softmax activation function.

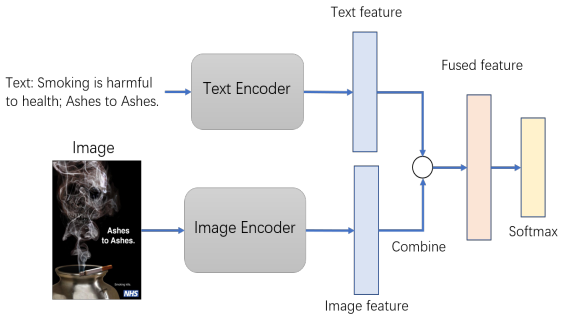


Figure 5: Multimodal model for integrating text and image data.

5.1 Experiment Settings

We used Pytorch (Paszke et al., 2019) to build the model. The pre-trained models are available in Pytorch. The word embeddings have been trained on a Wikipedia dataset by Glove (Pennington et al., 2014). In the training process, we did not update the parameters in the pre-training models. When the model gradually tended to converge, we updated the parameters of the pre-training models with training data to avoid overfitting. We used the Adam optimizer (Kingma and Ba, 2014) to optimize the loss function, and the training method of gradient clipping (Zhang et al., 2019) to avoid gradient explosion. Other hyper-parameter settings are shown in Table 3.

5.2 Results

The classification results are shown in Table 4. "Random" means that random predictions were made using the data as a baseline. In general, the model performed best on metaphor detection, followed by metaphor intention detection, and finally metaphor sentiment detection. For image and

			Metaphor		Sentiment		Intention	
Type	Text	Image	Validation	Test	Validation	Test	Validation	Test
Random	-	-	0.5063	0.4923	0.2222	0.2023	0.3416	0.3609
Text	Bi-LSTM	-	0.7458	0.7434	0.5705	0.5714	0.6597	0.6593
	BERT	-	**0.7742**	**0.7736**	**0.5958**	**0.5927**	**0.6794**	**0.6720**
Image	-	VGG16	0.7315	0.7345	0.5953	0.5914	0.6672	0.6658
	-	EfficientNet	0.7467	0.7405	0.5563	0.5548	0.6441	0.6324
	-	ResNet50	**0.7677**	**0.7646**	**0.5715**	**0.5714**	**0.6658**	**0.6653**
Text + Image	Bi-LSTM	VGG16	0.7735	0.7658	0.6195	0.6157	0.6843	0.6812
	Bi-LSTM	EfficientNet	0.7832	0.7795	0.5723	0.5714	0.6672	0.6732
	Bi-LSTM	ResNet50	0.7988	0.7912	0.6263	0.6220	0.7036	0.6843
	BERT	VGG16	0.8033	0.8072	0.6289	0.6188	0.7012	0.7000
	BERT	EfficientNet	0.7975	0.8033	0.6152	0.6125	0.6833	0.6757
	BERT	ResNet	**0.8276**	**0.8286**	**0.6462**	**0.6422**	**0.7278**	**0.7245**

Table 4: Results on three tasks with a combination method of concatenate.

	Metaphor		Sentiment		Intention	
Combination Methods	Validation	Test	Validation	Test	validation	Test
Add	0.7868	0.7834	0.6205	0.6186	0.6827	0.6779
Multiply	0.7596	0.7583	0.5685	0.5636	0.6442	0.6457
Maximum	0.7827	0.7759	0.6113	0.6074	0.7035	0.6993
Concatenate	0.8276	0.8286	0.6462	0.6422	0.7278	0.7245

Table 5: Results on different multimodal combinations for BERT + ResNet.

multimodal classification, the ResNet50 performed best, followed by VGG16, and finally EfficientNet. Because ResNet solved the problem of gradient disappearance through the method of residual connection, the classification performance was better than VGG16 and EfficientNet. For text and multimodal classification, BERT performed better than Bi-LSTM. BERT has been fully trained in a large-scale corpus, using transfer learning technology to fine-tune our three tasks and data, so it can achieve better performance. From the perspective of different features, multimodal features perform best, followed by text-only features, and finally image-only features. Multimodal fusion helps to improve the classification performance by 6%. This shows that the combination of image and text features is indeed helpful for the detection and understanding of metaphors, especially the detection of sentiments and intentions in metaphors. In addition, the importance of text modal data is explained. Without text description, it is difficult to detect metaphors correctly using only visual modal data.

To verify the influence of feature fusion on classification, we compared four different feature fusion methods. The results are shown in Table 5. The concatenate method to merge image and text features produces the highest accuracy. It shows that concatenate can make full use of the complementarity between different modal data, eliminate the noise generated by the fusion of different modal data, and improve the detection effect. In contrast, the other three fusion methods cannot effectively eliminate the influence of noise introduced by different modal data, and it therefore interferes with the training of the model. Overall, the multimode model that combines the BERT text function and the ResNet50 image function through the concatenation method performs best on our three tasks.

6 Conclusion

This paper presents the creation of a novel resource, a large-scale multimodal metaphor dataset, MultiMET, with manual fine-gained annotation for metaphor understanding and research. Our dataset enables the quantitative study of the interplay of multimodalities for metaphor detection and confirms the contribution of visuals in metaphor occurrence in natural language. It also offers a set of baseline results of various tasks and shows the importance of combining multimodal cues for metaphor understanding. We hope MultiMET provides future researchers with valuable multimodal training data for the challenging tasks of multimodal metaphor processing and understanding ranging from metaphor detection to sentiment analysis of metaphor. We also hope that MultiMET will help to expand metaphor research from monomodality to multimodality and improve the performance of automatic metaphor understanding systems and contribute to the in-depth understanding and research development of metaphors. The dataset will be publicly available for research.

Ethical Considerations

This research was granted ethical approval by our Institutional Review Board (Approval code: DUTIEE190725_01). We collected publicly available Twitter and Facebook data using Twitter and Facebook APIs complying with Twitter and Facebook's terms of service. We did not store any personal data (e.g., user IDs, usernames) and we annotated the data without knowledge of individual identities.

We annotated all our data using two independent approaches (expert based and crowdsourcing based) for two different types of tasks: the annotation of metaphor and the selection of types of sentiment and intent. For metaphor annotation, a deep understanding of metaphorical units was necessary. This challenging task was completed by five researchers who involved in this project. To annotate sentiment and intent, we used CrowdFlower, the crowdsourcing platform. To ensure that crowd workers were fairly compensated, we paid them at an hourly rate of 15 USD per hour, which is a fair and reasonable rate of pay for crowdsourcing (Whiting et al., 2019). We launched small pilots through CrowdFlower. The pilot for sentiment options took on average 43 seconds, and crowd workers were thus paid 0.18 USD per judgment, in accordance with an hourly wage of 15 USD. At the same time, the annotation of author intent took on average 23 seconds, and we thus paid 0.10 USD per judgment, corresponding to an hourly wage of 15 USD.

Acknowledgments

We would like to thank the anonymous reviewers for their insightful and valuable comments. This work is supported by NSFC Programs (No.62076051).

References

Julia Birke and Anoop Sarkar. 2006. 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, Trento, Italy.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy.

Lynne Cameron. 2003. *Metaphor in educational discourse*. A&C Black, London, UK.

Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the 2017 IEEE International Conference on Computer Vision*, pages 521–530, Venice, Italy.

Ursula Christmann, Lena Wimmer, and Norbert Groeben. 2011. The aesthetic paradox in processing conventional and non-conventional metaphors: A reaction time study. *Scientific Study of Literature*, 1(2):199–240.

Francesca MM Citron and Adele E Goldberg. 2014. Metaphorical sentences are more emotionally engaging than their literal counterparts. *Journal of Cognitive Neuroscience*, 26(11):2585–2595.

Seana Coulson and Cyma Van Petten. 2002. Conceptual integration and metaphor: An event-related potential study. *Memory & Cognition*, 30(6):958–968.

Dipto Das. 2019. *A multimodal approach to sarcasm detection on social media*. Ph.D. thesis, Missouri State University.

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)*, pages 4171–4186, Minneapolis, USA.

Francesca Ervas. 2019. Metaphor, ignorance and the sentiment of (ir) rationality. *Synthese*, pages 1–25.

Mark Evan Nelson. 2008. Multimodal synthesis and the voice of the multimedia author in a japanese efl context. *Innovation in Language Learning and Teaching*, 2(1):65–82.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Charles Forceville. 1996. *Pictorial metaphor in advertising*. Psychology Press, East Sussex, UK.

Charles Forceville. 2016. Visual and multimodal metaphor in film. In *Embodied metaphors in film, television, and video games: Cognitive approaches*, pages 17–32. Routledge, Abingdon, USA.

Charles Forceville and Eduardo Urios-Aparisi. 2009. *Multimodal metaphor*, volume 11. Walter de Gruyter, Berlin, Germany.

Charles Forceville et al. 2017. Visual and multimodal metaphor in advertising: Cultural perspectives. *Styles of Communication*, 9(2):26–41.

Andrew Goatly. 2007. *Washing the brain: Metaphor and hidden ideology*, volume 23. John Benjamins Publishing, Amsterdam, Netherlands.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016*, pages 770–778, Las Vegas, USA.

Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016*, pages 1–10, Las Vegas, USA.

Dirk Hovy, Shashank Srivastava, Sujay Kumar Jauhar, Mrinmaya Sachan, Kartik Goyal, Huying Li, Whitney Sanders, and Eduard Hovy. 2013. Identifying metaphorical word use with tree kernels. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 52–57, Atlanta, Georgia.

Hyeju Jang, Keith Maki, Eduard Hovy, and Carolyn Rose. 2017. Finding structure in figurative language: Metaphor detection with topic-based frames. In *Proceedings of the 18th Annual SIGDIAL Meeting on Discourse and Dialogue*, pages 320–330, Saarbrücken, Germany.

Hyeju Jang, Seungwhan Moon, Yohan Jo, and Carolyn Rose. 2015. Metaphor detection in discourse. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 384–392, Prague, Czech Republic.

Hermann Kappelhoff and Cornelia Müller. 2011. Embodied meaning construction: Multimodal metaphor and expressive movement in speech, gesture, and feature film. *Metaphor and the Social World*, 1(2):121–153.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv e-prints*, page arXiv:1412.6980.

Zoltán Kövecses. 1995. Anger: Its language, conceptualization, and physiology in the light of cross-cultural evidence. *Language and the Cognitive Construal of the World*, pages 181–196.

Zoltán Kövecses. 2003. *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press, Cambridge, UK.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4614–4624, Hong Kong, China.

George Lakoff and Mark Johnson. 1980. *Metaphors we live by*. University of Chicago press, Chicago, USA.

Jerry Liu, Nathan O'Hara, Alexander Rubin, Rachel Draelos, and Cynthia Rudin. 2020. Metaphor detection using contextual word embeddings from transformers. In *Proceedings of the 2nd Workshop on Figurative Language Processing*, pages 250–255, Online.

Marco Lui and Timothy Baldwin. 2012. langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea.

Sijie Mai, Haifeng Hu, Jia Xu, and Songlong Xing. 2020. Multi-fusion residual memory network for multimodal human sentiment comprehension. *IEEE Transactions on Affective Computing*, pages 1–15.

Rui Mao, Chenghua Lin, and Frank Guerin. 2018. Word embedding and wordnet based metaphor identification and interpretation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1222–1231, Melbourne, Australia.

Zachary J Mason. 2004. Cormet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics*, 30(1):23–44.

Larry R Medsker and LC Jain. 2001. Recurrent neural networks. *Design and Applications*, 5:1–391.

Saif Mohammad, Ekaterina Shutova, and Peter Turney. 2016. Metaphor as a medium for emotion: An empirical study. In *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics*, pages 23–33, Berlin, Germany.

Michael Mohler, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the 1st Workshop on Metaphor in NLP*, pages 27–35, Atlanta, USA.

Michael Mohler, Mary Brunson, Bryan Rink, and Marc Tomlinson. 2016. Introducing the lcc metaphor datasets. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 4221–4227, Portorož, Slovenia.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *arXiv e-prints*, page arXiv:1912.01703.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, Doha, Qatar.

Barbara J Phillips and Edward F McQuarrie. 2004. Beyond visual metaphor: A new typology of visual rhetoric in advertising. *Marketing Theory*, 4(1-2):113–136.

Malay Pramanick, Ashim Gupta, and Pabitra Mitra. 2018. An lstm-crf based approach to token-level metaphor detection. In *Proceedings of the Workshop on Figurative Language Processing 2018*, pages 67–75, New Orleans, USA.

Elisabeth El Refaie. 2003. Understanding visual metaphor: The example of newspaper cartoons. *Visual Communication*, 2(1):75–95.

Ekaterina Shutova. 2017. Annotation of linguistic and conceptual metaphor. In *Handbook of linguistic annotation*, pages 1073–1100. Springer, New York, USA.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, USA.

Ekaterina Shutova, Lin Sun, Elkin Darío Gutiérrez, Patricia Lichtenstein, and Srini Narayanan. 2017. Multilingual metaphor processing: Experiments with semi-supervised and unsupervised learning. *Computational Linguistics*, 43(1):71–123.

Ekaterina Shutova, Lin Sun, and Anna Korhonen. 2010. Metaphor identification using verb and noun clustering. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1002–1010, Beijing, China.

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv e-prints*, page arXiv:1409.1556.

Paula Pérez Sobrino. 2017. *Multimodal metaphor and metonymy in advertising*, volume 2. John Benjamins Publishing Company, Amsterdam, Netherlands.

Esther Šorm and Gerard Steen. 2018. Towards a method for visual metaphor identification. In *Visual metaphor: Structure and process*, volume 18, pages 47–88. John Benjamins Publishing Company, Amsterdam, Netherlands.

Gerard Steen. 2010. *A method for linguistic metaphor identification: From MIP to MIPVU*, volume 14. John Benjamins Publishing, Amsterdam, Netherlands.

Gerard J Steen. 2018. *Visual metaphor: Structure and process*, volume 18. John Benjamins Publishing Company, Amsterdam, Netherlands.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (multioff) for identifying offensive content in image and text. In *Proceedings of the 2nd Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of International Conference on Machine Learning 2019*, pages 6105–6114, California, USA.

Miloš Tasić and Dušan Stamenković. 2015. The interplay of words and images in expressing multimodal metaphors in comics. *Procedia-Social and Behavioral Sciences*, 212:117–122.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor detection with cross-lingual model transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 248–258, Maryland, USA.

Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Scotland, UK.

Eduardo Urios-Aparisi. 2009. Interaction of multimodal metaphor and metonymy in tv commercials: Four case studies. *Multimodal Metaphor*, 11:95–116.

Eija Ventola, Cassily Charles, and Martin Kaltenbacher. 2004. *Perspectives on multimodality*, volume 6. John Benjamins Publishing, Amsterdam, Netherlands.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2017. Select-additive learning: Improving generalization in multimodal sentiment analysis. In *Proceeding of 2017 IEEE International Conference on Multimedia and Expo*, pages 949–954, Hong Kong, China.

Lyndon CS Way and Simon McKerrell. 2017. *Music as multimodal discourse: Semiotics, power and protest*. Bloomsbury Publishing, London, UK.

Mark E Whiting, Grant Hugh, and Michael S Bernstein. 2019. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, pages 197–206, Hilversum, The Netherlands.

Zhao Xiufeng. 2013. The conceptual integration model of multimodal metaphor construction: A case study of a political cartoon. *Foreign Languages Research*, 5:1–8.

Fan-Pei Gloria Yang, Kailyn Bradley, Madiha Huq, Dai-Lin Wu, and Daniel C Krawczyk. 2013. Contextual effects on conceptual blending in metaphors: an event-related potential study. *Journal of Neurolinguistics*, 26(2):312–326.

Keren Ye, Narges Honarvar Nazari, James Hahn, Zaeem Hussain, Mingda Zhang, and Adriana Kovashka. 2019. Interpreting the rhetoric of visual advertisements. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark.

Omnia Zayed, John P McCrae, and Paul Buitelaar. 2019. Crowd-sourcing a high-quality dataset for metaphor identification in tweets. In *Proceedings of the 2nd Conference on Language, Data and Knowledge*, pages 1–17, Leipzig, Germany.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. 2019. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv e-prints*, page arXiv:1905.11881.