Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Exploring Deep Learning Approach on Semantic Gap: A Comprehensive Review

JinHang Zhang, MD. Sah Bin Hj, Xianpeng Li, Jianbo Fan

Department of Emergent Computing, Faculty of Computing, Universiti Teknologi Malaysia 81310
Skudai, Johor Bahru, Johor, Malaysia

zhangyuoctopus@foxmail.com

Department of Emergent Computing, Faculty of Computing, Universiti Teknologi Malaysia 81310
Skudai, Johor Bahru, Johor, Malaysia.

sah@utm.my

Province Guizhou Qiannan College of Science and Technology(GCST) International technology
Mengjiang Avenue, Lianjiang Street, Huishui County, Guizhou 550600 China

jy02669100@126.com

Department of Emergent Computing, Faculty of Computing, Universiti Teknologi Malaysia 81310
Skudai, Johor Bahru, Johor, Malaysia.

jianbo@graduate.utm.my

ARTICLE INFO

Received: 14 Dec 2024

Revised: 18 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

With the prevalence of the Internet and smartphones, users upload a large number of images to the web. However, it is challenging for users to find what they really need from the vast sea of images. It is also difficult for Internet companies to effectively integrate their massive image data resources. In traditional content-based image retrieval, images are indexed by their low-level visual features, which leads to a key problem: the semantic gap between low-level features and high-level semantic concepts. To address this issue, semantic-based image retrieval has been proposed as a solution to bridge the semantic gap, making it a key technical challenge in the field of image retrieval. To tackle these challenges, this proposal presents a novel multi-annotation method for images and develops an image retrieval system based on deep learning and image semantic content. Preparatory work, including a literature review and methodology development, will be conducted to implement the semantic-based image retrieval system and efficiently utilize the vast number of available images.

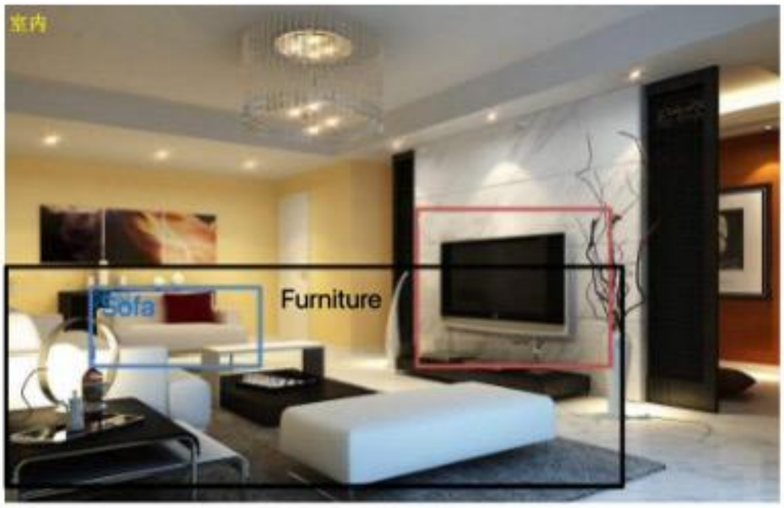
Keywords: deep learning, image semantic, image retrieval, image annotation

Journal of Information Systems Engineering and Management
2025, 10(38s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

1. INTRODUCTION

The demand for image retrieval technology is widespread across various aspects of human life, particularly in fields such as e-commerce, copyright protection, medical diagnosis, public safety, and street view maps, where its application holds significant commercial potential. For instance, in e-commerce, Google launched **Google Goggles**, a service that allows users to upload product images to a server. The image retrieval application on the server then helps users find links to stores offering the same or similar products [1]. In the realm of copyright protection, copyright service providers can utilize image retrieval technology to manage trademarks, such as verifying whether a trademark has already been registered. In medical diagnosis, image retrieval technology assists doctors by categorizing and retrieving images from medical image libraries, thereby enabling more accurate identification of patient lesions. Similarly, in applications like street view maps, image retrieval can help users identify objects in street scenes, enhancing safety by allowing users to detect and avoid potential dangers [2]. Overall, image retrieval technology has become deeply integrated into numerous fields, providing significant convenience to users in both their professional and daily lives. Common image retrieval techniques primarily include **Text-Based Image Retrieval (TBIR)**, **Content-Based Image Retrieval (CBIR)**, and **Sketch-Based Image Retrieval (SBIR)**. The advantage of TBIR lies in its simplicity of implementation, ease of understanding, alignment with human retrieval habits, and its ability to deliver precise results. However, TBIR requires substantial human effort for manually annotating images, which becomes impractical for large multimedia databases, especially when new data is continuously added [6]. Additionally, TBIR struggles to adapt quickly to new data and cannot resolve the subjectivity of annotators in perceiving and describing content. To address the limitations of TBIR, experts have proposed CBIR. The primary advantage of CBIR is its ability to extract features directly from image content and define image similarity by comparing these features, thereby reducing the need for human intervention [7]. Moreover, the approximate matching method used in CBIR offers faster retrieval and sorting speeds compared to TBIR. Nevertheless, CBIR also has its shortcomings. Images on the internet often originate from diverse environments and domains, and CBIR technology, which relies on low-level visual features, faces significant challenges in real-world applications due to the **semantic gap** problem. In image retrieval, the **semantic gap** generally refers to the inconsistency between the information extracted from visual data and the interpretation of this data by users in specific contexts. The semantic gap is a persistent challenge in image retrieval technology, particularly in **Content-Based Image Retrieval (CBIR)**. CBIR relies on low-level visual features of images, which cannot fully capture the semantic meaning of the images, making the semantic gap inevitable. In image retrieval tasks within the same category, images are often assigned multiple category labels based on their characteristics. The typical operation of image retrieval based on visual features involves extracting different features from an image and learning the mapping from these features to image categories through linear classification algorithms, thereby classifying images into the correct category labels. The implicit assumption in these methods is that, in the feature space, images with the same category labels are closer to each other, while images with different category labels are farther apart.However, in practical applications, category labels are often not mutually independent. For instance, as shown in Figure 1-1 (Category Labels), an image may be categorized as "sofa," "furniture," and "interior." These categories exhibit hierarchical relationships; for example, "sofa" is a subset of "furniture," and "furniture" is an important component of "interior".

Figure 1-1 Category labels



From the analysis above, it is clear that in the real world, images with similar visual features may sometimes be categorized under the same label, yet at other times be classified into different categories [5]. Conversely, images with different visual features, when categorized into different classes, may sometimes be grouped under the same category [8]. This indicates that some features in the visual feature space are not mutually independent, which leads to images not being strictly separable when using linear classification algorithms.The inability to strictly linearly classify images based on visual features can result in significant intra-class variation within the same category, i.e., images within the same category in feature space may be far apart. Meanwhile, inter-class differences between different categories might be small, i.e., the distance between images of different categories in feature space may be unexpectedly close. This suggests that images may not be strictly linearly separable in low-level visual features, and the difficulty of linear classification ultimately leads to a mismatch between the visual information obtained by algorithms and users' understanding of images, thus creating a semantic gap. This paper aims to explains the reasons for the semantic gap and proposes several methods to solve the problem through semantic embedding. The specific structure is as follows: Chapter 1 is Introduction, first briefly introduces the background and significance of the project; Then the present situation of image retrieval is analyzed and summarized. Secondly, it introduces the research content of this thesis -- image semantic feature extraction and retrieval technology based on deep learning [9]. Finally, the content structure of this thesis is explained. Chapter 2 shows the main contributions of the review. chapter 3 literature review, firstly, the relevant knowledge of semantic gap is briefly introduced, and then the basic principles and common models of deep learning are introduced. Then, the similarity between other method and deep learning network is simulated, and compared the different algorithms in deep learning to narrow the semantic gap.

2. CONTRIBUTIONS

2.1 The main contributions of the review

This review aims to present the development of image retrieval technology and analyze the reasons for the semantic gap in multi-objective semantic retrieval. It also compares different solutions to the semantic gap problem in content-based image retrieval (CBIR) and explores various algorithms based on deep learning technology to address this issue. By doing so, this review not only provides a comprehensive understanding of the current state of deep learning in bridging the semantic gap but also charts a course for future research and practical applications in multimedia retrieval systems.

3. COMPARISON WITH STRATEGIES TO SOLVE THE SEMANTIC GAP PROBLEM

3.1 Current Research in Image Retrieval Technology

In image retrieval tasks, extracting features from images is a necessary step. Based on the level of abstraction and complexity in describing image content, image features can be categorized as shown in Figure 2-1 (Image Feature Categories).

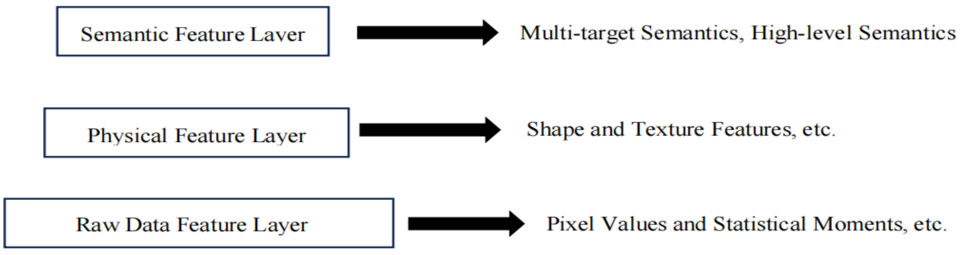

Semantic Feature Laver
Multi-target Semantics, High-level Semantics
Physical Feature Layer
Shape and Texture Features, etc.
Raw Data Feature Layer
Pixel Values and Statistical Moments, etc.

Figure 2-1. Image features categories

The features at higher levels of the pyramid are more complex and abstract, and higher-level features are often abstracted from lower-level features. Many mature algorithms already exist for extracting image features, and we will introduce commonly used image features by level.

Raw Data Layer Features.

The raw data layer features primarily consist of the original pixels of the image matrix or low-level statistical values of pixels. Among these, color features are the earliest and most widely used visual features in the raw data layer. Color features utilize all the pixels in an image region, and commonly used color features include color moments, color histograms, and color correlograms [4].

Physical Feature Layer Features.

Physical layer features of an image mainly include descriptions of the primary content within the image, with common physical layer features being shapes and textures. Generally, there are two types of methods to represent the shape features of an image: boundary feature methods, which focus on the object's boundary contours, and Fourier descriptors, which describe the entire shape area.

Semantic Layer Features.

Semantic layer features represent a higher level of abstraction for images. When retrieval results using lower-level features are unsatisfactory, it becomes necessary to introduce high-level semantic features to achieve retrieval functionality. Semantics generally refer to the meaning of the concepts represented by data in reality and the relationships between these conceptual meanings. In short, semantics are the interpretation and logical representation of data in a specific domain. Image semantics represent the meaning contained in the image content. Image semantics exist on multiple levels, and semantic features based on different levels of semantics have varying impacts on image retrieval.

3.2 Existing Methods to solve the Semantic Gap

3.2.1 Method based on processing scope

Based on whether image feature extraction is regional or global, it can be divided into region-based semantic extraction and global-based semantic extraction. The region-based semantic extraction method first uses object and scene classifiers to recognize the objects and scenes of the segmented image. It then obtains semantic information by mining the spatial and positional relationships between the objects and combining them with the scene. For example, Wang et al. developed the Simplicity system, which first segments images, extracts image features, and categorizes image objects using statistical methods to obtain simple semantic classes (e.g., texture-non-textured, indoor-outdoor, image-photo) [10]. Jeon et al. also proposed a cross-media correlation model (CMRM) based on regional feature extraction [11]. This method first segments the image and then obtains a few BLOBs by fusing adjacent areas based on color and shape features. The probability of co-occurrence of semantic keywords and BLOBs is then calculated using the CMRM model.

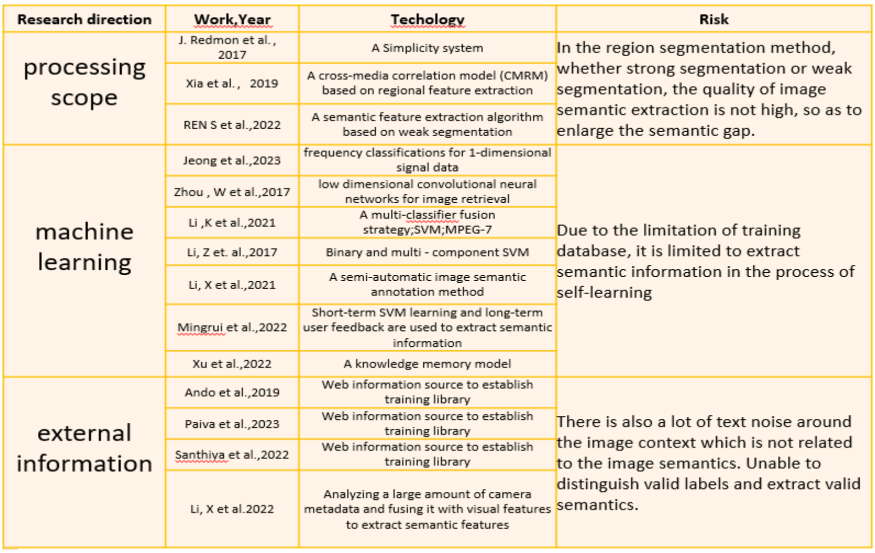
However, in real-life applications, objects are not precisely segmented (strong segmentation) and then understood. Instead, a relatively homogeneous region (weak segmentation) is roughly segmented according to the user's point of interest, and visual features are extracted [12]. Navon's psychological research shows that, in general, global features take precedence over local features in human visual perception—a "forest before trees" strategy [13]. Therefore, from the perspective of processing speed and efficiency, it is better to directly use global semantic features or weak segmentation methods. For example, Fan proposed a semantic feature extraction algorithm based on weak segmentation. This algorithm uses weak segmentation to obtain different semantic image features, mines the contextual relationships between salient objects in the image, and finally establishes a statistical model of salient objects and their maximum correlation semantics. In the research of global semantic feature extraction, Li et al. adopted a 2D hidden Markov model [14]. Although the region-based method can construct a more flexible model than the global-based method in extracting semantic information, it has higher requirements for image segmentation and target recognition. Currently, to balance the quality and performance of image semantic extraction, a multi-scale image semantic extraction method based on the fusion of local and global features can be adopted.

3.2.2 Methods based on machine learning

Machine learning algorithms are relatively complex algorithms that, driven by large amounts of data, can self-learn better complex mapping relationships and further reduce the semantic gap. Bayesian networks are probabilistic networks often used to solve problems of uncertainty and incompleteness. Aksoy et al. constructed a Bayesian framework to shorten the semantic gap. In region-based feature extraction, pixel spectrum, texture, and other information are used to repeatedly split and fuse the segmented image region, and the spatial relationship is used to model the image [15]. Finally, a naive Bayes classifier is used to train positive and negative samples and obtain semantic features. Support vector machines (SVM) are a common classifier based on the principle of minimizing structural risk. Han et al. proposed a multi-classifier fusion strategy based on multi-example learning. SVM based on multi-example learning extracts features from 3×3 image sub-blocks [16]. The global-based SVM combines the MPEG-7 color descriptor and edge histogram operator to address the deficiency of multi-example learning SVM in image deformation resistance. Goh et al. used binary and multipartite SVM to extract image semantic features [17]. They adopted a dynamic integration algorithm to propagate the representative confidence level to the multi-class SVM classifier layer by layer, continuously adjusting the classifier dynamically to improve classification accuracy.

3.2.3 Methods based on external information

The method of semantic extraction based on external information involves obtaining semantic information related to image content from external additional information. This generally refers to the image itself having a name or the context of the image reflecting keywords or sentences related to the image subject. Web image retrieval has the characteristics of large throughput, fast response, and high accuracy, making the method based on external information an important approach to extracting high-level semantics from web images. Shen et al. used the method of crawling image context and constructing lexical chains as the external information source of image semantics [18]. Yang et al. used data mining technology to extract semantics from the image context and obtained implicit fuzzy semantic information through information mining [19]. With the development of mobile internet, especially the rise of graphic-sharing applications such as Facebook, Moments, and Weibo, a large number of image resources are available online. These images are usually accompanied by captions and tags. Ames et al. discovered the phenomenon of users freely adding large labels to images, which represents a form of unstructured knowledge. Extracting image semantics from this unstructured knowledge is a current challenge. Boutell et al. found that metadata such as exposure time and flash usage have significant differences in indoor and outdoor scene recognition [20]. They leveraged this feature distinction and integrated it with visual features to extract semantic features, effectively addressing the problem of indoor and outdoor image classification. Using external information from images can effectively obtain high-level semantic information and is easily processed by computers, offering significant advantages. However, as there is often text noise around the image context unrelated to image semantics, it can be difficult to obtain effective information. Additionally, issues such as missing labels, inaccurate labels, semantic confusion, and garbage labels can arise, making this method less ideal in practice. One way to compensate is to dynamically adjust the ratio of visual features and tag semantics based on the abundance of external information.



Research direction	Work,Year	Techology	Risk
processing scope	J. Redmon et al., 2017	A Simplicity system	In the region segmentation method, whether strong segmentation or weak segmentation, the quality of image semantic extraction is not high, so as to enlarge the semantic gap.
	Xia et al., 2019	A cross-media correlation model (CMRM) based on regional feature extraction	
	REN S et al.,2022	A semantic feature extraction algorithm based on weak segmentation	
machine learning	Jeong et al.,2023	frequency classifications for 1-dimensional signal data	Due to the limitation of training database, it is limited to extract semantic information in the process of self-learning
	Zhou , W et al.,2017	low dimensional convolutional neural networks for image retrieval	
	Li ,K et al.,2021	A multi-classifier fusion strategy;SVM;MPEG-7	
	Li, Z et. al.,2017	Binary and multi - component SVM	
	Li, X et al.,2021	A semi-automatic image semantic annotation method	
	Mingrui et al.,2022	Short-term SVM learning and long-term user feedback are used to extract semantic information	
	Xu et al.,2022	A knowledge memory model	
external information	Ando et al.,2019	Web information source to establish training library	There is also a lot of text noise around the image context which is not related to the image semantics. Unable to distinguish valid labels and extract valid semantics.
	Paiva et al.,2023	Web information source to establish training library	
	Santhiya et al.,2022	Web information source to establish training library	
	Li, X et al.2022	Analyzing a large amount of camera metadata and fusing it with visual features to extract semantic features	

Figure 3-1. Existing Methods to solve the Semantic Gap

The processing scope encompasses the overall structure of an affective semantic retrieval system, along with proposed algorithms to address key challenges. However, in the region segmentation methods—whether strong or weak segmentation—the quality of image semantic extraction remains suboptimal, thereby exacerbating the semantic gap. Deep learning is implemented through multilayer neural networks for the first time, converting high-dimensional raw input signals into a lower-dimensional representation. This process yields low-dimensional information that can be effectively described through neural network feedback learning. Nevertheless, due to limitations within the training database, there are constraints on extracting semantic information during self-learning processes. External information presents a multi-layer emotion model based on emotional psychology that integrates emotion, mood, and personality mapping. Additionally, there exists considerable textual noise surrounding the image context that is unrelated to its semantics. This noise complicates the ability to distinguish valid labels and extract meaningful semantics effectively.

3.3 Application of Deep Learning in Narrowing the Semantic Gap

Compared to traditional methods for addressing the semantic gap, deep learning has demonstrated significant advantages. In recent years, with advancements in machine learning, deep learning—a branch of machine learning based on connectionism—has been increasingly applied in image retrieval to tackle the semantic gap issue. Deep learning constructs layers with local structures through interconnected computational units, enabling these layers to express complex functions. Improved optimization algorithms allow these layers to identify structures that effectively express these functions. Driven by larger datasets and more powerful hardware, deep learning has overcome many challenges that traditional shallow learning methods could not resolve.

In 2006, Hinton published a seminal study on Deep Belief Networks (DBN) [1], marking the beginning of the deep learning era. The paper proposed alternating supervised and unsupervised learning to address the vanishing gradient problem in deep neural networks. In 2011, the introduction of the ReLU activation function replaced traditional nonlinear activation functions, effectively mitigating the vanishing gradient issue during model training and enabling the training of deeper networks [21]. In 2012, Hinton's group designed the AlexNet model, which dominated the ImageNet competition by significantly reducing the error rate from 26% to 15%. AlexNet expanded the original LeNet5 model by increasing its depth, using ReLU as the activation function, and employing the Dropout method to optimize the network. This success established convolutional models as the preferred approach for processing image data. By 2015, Y. Lecun, Y. Bengio, and G. Hinton demonstrated that local minima in deep learning are effectively global minima, eliminating concerns about local minima hindering deep learning performance [21].

3.3.1 Comparison different deep learning Models in Narrowing the Semantic Gap

Deep learning has been widely and deeply applied in image recognition, showing significant potential in bridging the semantic divide—the gap between how computer vision systems understand image content and how humans perceive it. Below are some of the core deep learning algorithms and their applications in bridging the semantic gap in image recognition.

1.Convolutional Neural Networks (CNNs):

CNNs are the most commonly used deep learning models for solving image recognition tasks. By simulating the workings of the human visual system [22], CNNs can effectively recognize and classify objects in images. They extract hierarchical features, from simple edges to complex object parts, which helps narrow the semantic gap.

① AlexNet is the first deep learning model that significantly improved performance in the ImageNet challenge, and its success demonstrated the enormous potential of CNN in handling complex image data [23]. AlexNet can learn effective feature representations from a large amount of image data, which are not only applicable to classification tasks, but also provide useful information foundation for other visual tasks such as object detection and semantic segmentation [24]. The success of AlexNet has propelled the application of deep neural networks in computer vision, thereby inspiring research and development in more complex visual tasks including semantic analysis. Later, more advanced models such as VGG and ResNet, partially based on the design ideas of AlexNet, have demonstrated better performance in handling more complex visual understanding tasks such as image semantic segmentation and scene parsing [25]. Direct responses to the semantic divide usually require more advanced techniques and models, such as multimodal learning combined with natural language processing, or neural network architectures specifically designed to understand the context and deeper semantic information of image content. These approaches extend and deepen on the basic feature learning provided by AlexNet.

② The VGG network, which improves image recognition performance by using a very deep convolutional network structure, is an important resource for understanding deep CNN architecture [26]. It serves as an on-processing image classification model that makes the following contributions to solving the semantic divide problem:

- Depth of feature extraction: The multi-layer structure of VGGNet can extract rich hierarchical features from images, which can be used for more complex tasks such as semantic segmentation and object detection in images. In these tasks, deep-level features help to better understand the context and relationships between objects in the image.
- Transfer Learning: VGGNet is often used as a pre-trained model for other visual tasks due to its

excellent feature extraction capability. In transfer learning scenarios, models trained by VGGNet can be used to initialize or fine-tune networks for new tasks, which can help the new models better understand and process semantic information [27].

- Multimodal learning: While VGGNet itself does not contain mechanisms for processing multimodal data (e.g., combining images and text), its feature extraction capabilities can be combined with other types of neural networks (e.g., RNNs or Transformers for processing text) to work together on more complex semantic parsing tasks [28]. For example, in tasks such as image captioning or Visual Question Answering (VQA), VGGNet can provide visual features, while the text processing part parses and generates semantic descriptions related to the image content.

③ ResNet, proposed by Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun at Microsoft Research in 2015 [33], not only passes inputs through traditional layers but also adds them directly to the output of later layers through skip connections. This design allows the network to learn the residual (i.e., difference) between the input and output [30], which helps to address the problem of gradient vanishing in deeper networks. It effectively improves the training process and final performance, and is very helpful for understanding how to narrow the semantic gap through deep networks.

ResNet has made significant contributions in narrowing the semantic gap, that is, reducing the gap between machine vision systems and human visual understanding [29]. Although it was originally designed for image classification tasks, its deep structure and innovative residual learning mechanism make it perform exceptionally well in a variety of more complex visual understanding tasks. Here are several key contributions of ResNet in helping to narrow the semantic gap:

- Deeper feature learning: ResNet allows models to learn richer feature hierarchies through its depth and residual connections. These deep features can better capture the complex patterns and structures in the image, thereby helping machines approach human visual understanding.
- Improved Gradient Flow: Residual connectivity improves the flow of gradients through the deep network and solves the gradient vanishing problem. This allows the network to be trained deeper without loss of information, thus understanding and processing semantic information in images more efficiently.
- Multi-task learning and transfer learning: The structure of ResNet makes it an excellent feature extractor, commonly used for transfer learning. By pre-training on one task (e.g., image categorization) and migrating to other tasks (e.g., semantic segmentation, object detection), ResNet can apply learned visual features to new contexts that require more semantic understanding.
- Fine-grained and contextual understanding: In fine-grained image recognition and scene

understanding tasks, ResNet can help models distinguish subtle visual differences and understand the elements in complex scenes and their relationships, which is particularly important for narrowing the semantic gap.

- Multimodal applications: In multimodal learning tasks that combine images and text, such as visual question answering (VQA) and image captioning, ResNet can provide deep visual features that, when combined with textual information, can generate more accurate and richer semantic outputs.

Through these contributions, ResNet has not only demonstrated outstanding performance in traditional image processing tasks, but also played an important role in understanding and generating semantic information related to images, providing machine vision systems with a more human-like understanding capability. This is crucial for the development of intelligent systems that can understand complex visual scenes and interact with humans naturally.

2. Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs):

Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) are primarily used for processing sequential data, but they have also been applied in the field of image recognition, especially in tasks that require understanding temporal or spatial dependencies within images, such as video recognition or dynamic scene analysis. In these applications, RNNs and LSTMs can help the model understand the temporal dependencies within image sequences, thereby improving semantic understanding of dynamic scenes. Recurrent Neural Networks (RNNs) and Long Short-Term Memory Networks (LSTMs) are very important in the processing of sequential data, including language processing and time series analysis.

3. Generative Adversarial Networks (GANs):

GAN consists of a generator network and a discriminator network, which compete with each other to improve performance [31]. The application of GANs in image generation and style transfer can improve the semantic expression of images and make the generated images more similar to human visual perception.

GANs deal with the problem of semantic divide in image recognition mainly by improving the semantic comprehension capability of the generative models so that they can better model and generate images with complex semantic content. This capability can help improve the semantic consistency of images and thus be more effectively applied to image recognition and understanding tasks.

4. Variational Autoencoders (VAEs):

- VAEs are a type of generative model that generates new data samples by learning the latent representation of input data.

In image recognition, VAEs can be used to generate new images with similar semantic content, helping to understand and bridge the semantic gap in image content. Variational Autoencoders (VAEs) are a type of generative model widely used for learning the latent representation of input data. In image recognition, especially in dealing with the problem of semantic gap, VAEs can help models better understand and generate images with complex semantics.

① Xianxu Hou, Linlin Shen, Ke Sun, and Guoping Qiu(2018) proposed a DFC-VAE model that aims to maintain the consistency of deep features when reconstructing images [32], in order to improve the quality and semantic accuracy of generated images.

② Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, and Max Welling(2014) introduced a method that combines variational autoencoders and semi-supervised learning to improve semantic interpretation of images by utilizing unlabeled data.

③ VampPrior is a prior distribution used for Variational Autoencoder (VAE), which stands for "Variational Mixture of Posteriors prior". This concept was proposed by Jakub M. Tomczak and Max Welling in the paper "VampPrior: Improving the Variational Autoencoder with a Learned Prior". VampPrior aims to improve the standard prior used in VAE models, which is typically a simple Gaussian distribution, by introducing a more complex and flexible prior to enhance the learning capability and generative performance of the entire model. It is a new VAE prior used to improve the generative performance of the model, especially in terms of semantic coherence and image quality.

5. Attention Mechanisms:

- The attention mechanism allows the model to focus on the important parts of the information when processing the input data. In the image recognition task, the introduction of the attention mechanism allows the model to better focus on the key information in the image, such as the important parts of the object or the regions that are closely related to the task, so as to understand the semantics of the image more accurately.

Attention mechanisms have become a very important concept in deep learning, especially in the field of image recognition and processing. They improve the model's semantic understanding of image content by focusing on key parts of the image.

① The "Transformer" model and its attention mechanism proposed by Ashish Vaswani have been widely applied to image recognition tasks, demonstrating the ability to focus on key information when processing images. (Advances in Neural Information Processing Systems (NeurIPS).

② ViT (Vision Transformer), which applies the Transformer structure to image recognition, has demonstrated the effectiveness and superiority of attention mechanisms at the whole-image level. This paper proposes a stacked attention network for image question answering, which enhances semantic understanding of image content by iteratively attending to different regions of the image.

Table 1.1 Comparison different deep learning Modelsin Narrowing the Semantic Gap

Algorithm	Model	Year	Author	Title	Contribution
CNNS	AlexNet	2012	Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton	ImageNet Classification with Deep Convolutional Neural Networks	This paper introduced AlexNet,it is a significant convolutional neural network (CNN) that played a pivotal role in advancing the field of deep learning. and the success of this model demonstrates the great potential of CNNS in processing complex image data.
	VGGNet	2014	Karen Simonyan and Andrew Zisserman	Very Deep Convolutional Networks for Large-Scale Image Recognition	VGGNet significantly improved the state-of-the-art in image classification upon its release. Its architecture has influenced many subsequent CNN designs and has been widely adopted for various applications beyond image classification, such as in neural style transfer and feature extraction tasks.
	ResNet	2016	*Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun	Deep Residual Learning for Image Recognition	This paper introduced residuals network (ResNet), an architecture that trains very deep networks by using residuals connections, effectively improves the training process and final performance, and is very helpful for understanding how to narrow the semantic gap through deep networks.

RNNS	/	2016	Andrej Karpathy, Justin Johnson, Li Fei-Fei	Visualizing and Understanding Recurrent Networks	This paper explored the internal mechanisms and effects of RNNs and LSTMs in a variety of tasks, including image description, helping to understand how these models process and understand sequence data.
LSTM	CNNs+LSTM	2015	Jeff Donahue et al.	Long-term Recurrent Convolutional Networks for Visual Recognition and Description	The researchers showed how a combination of CNN and LSTM can process video data to identify actions in video and generate descriptions. This method is suitable for solving the semantic gap problem in video comprehension.
GANs	Pix2pix	2017	Phillip Isola et al.	Image-to-Image Translation with Conditional Adversarial Networks	The Pix2Pix model is proposed, which is a conditional GAN capable of image-to-image conversion. Although primarily used for style transformation and image editing, its ability to process semantic information is equally important for understanding image content.
	StackGAN	2017	Han Zhang et al.	StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks	StackGAN shows how to generate photo-grade realistic images from text descriptions. In this way, model learning transforms complex text information (with rich semantics) into corresponding visual content.
	SPADE	2019	Mingyu Park et al.	Semantic Image Synthesis with Spatially-Adaptive Normalization	This paper presented SPADE, a method for semantic image synthesis that improves the semantic accuracy of generated images by utilizing spatially

				(SPADE)	adaptive normalization. This approach produces high-quality images and performs well on multiple semantically segmented datasets.
VAEs	DFC-VAE	2017	Jakub M. Tomczak, Max Welling	Deep Feature Consistent Variational Autoencoder (DFC-VAE)	The study proposed a DFC-VAE model that aims to maintain consistency of depth features when reconstructing images to improve the quality and semantic accuracy of the generated images.
	VampPrior	2018	Xianxu Hou, Linlin Shen, Ke Sun, Guoping Qiu	Variational Inference with a Neural Network Prior: Approximation and Sampling Capabilities	This paper presented VampPrior model, a new VAE priori for improving model generation performance, especially in terms of semantic coherence and image quality.
	Semi-Supervised Learning with Deep Generative Models	2014	Diederik P. Kingma, Danilo J. Rezende, Shakir Mohamed, Max Welling	Semi-Supervised Learning with Deep Generative Models	In this paper, a method combining variational autoencoder and semi-supervised learning is introduced to improve the semantic interpretation of images by using unlabeled data, thus narrowing the semantic gap.
Attention Mechanisms	Transformer	2017	Vaswani et al.	Attention is All You Need	The "Transformer" model and its attention mechanism proposed in this article have been widely used in image recognition tasks, demonstrating the ability to focus on key information when processing images.

	Vision Transformer	2021	Alexey Dosovitskiy et al.	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale	The paper introduced ViT (Vision Transformer), which applies the Transformer structure to image recognition and proves the effectiveness and superiority of the attention mechanism at the full image level.

These papers demonstrate how attention mechanisms improve the accuracy of semantic understanding by focusing the model's "gaze" on key regions of the image. These algorithms each have their own merits. When solving the semantic gap problem in image recognition, they enhance the computer system's visual understanding ability in different ways, making it closer to human visual perception.

4. CONCLUSION

Due to the implicit relationships between image categories, images are not linearly separable based on low-level visual features. To address this limitation, I propose a deep learning-based approach that incorporates the relationships between image categories into image classification. This method leverages existing deep learning models from the literature and introduces prior information about category relationships through semantic embeddings. The goal is to transform the low-level visual features of images from non-independent relationships to mutually independent ones, enabling more effective classification.

The proposed method consists of two main steps:

1. Semantic Extraction
2. Semantic Interpretation

1. Semantic Extraction

In the first step, a bottom-up approach is used to extract object-level semantic information from images. This process involves the following sub-steps:

- Image Segmentation: Based on the low-level visual features of the images, the images are segmented into multiple candidate regions.

- Feature Extraction and Classification: Deep neural networks are employed to perform regression classification on the image features of these candidate regions. This step ultimately yields the spatial location and category information of objects within the regions.

By extracting object-level semantics, this step captures the essential visual elements of the image, providing a foundation for further semantic interpretation.

2. Semantic Interpretation

The second step focuses on interpreting the semantic content of the images. The human process of describing images typically follows a sequence of observing the image, extracting visual material, conducting inductive reasoning, and generating a description. This process treats semantic content as sequential data. Since recurrent neural networks (RNNs) excel at processing sequential data, they are well-suited for inferring semantic content.

In this step, the extracted object-level semantic information is fed into an RNN-based model. The RNN processes the sequential data to infer higher-level semantic relationships and generate a comprehensive interpretation of the image content. This approach mimics the human cognitive process, enabling the model to bridge the gap between low-level visual features and high-level semantic understanding.

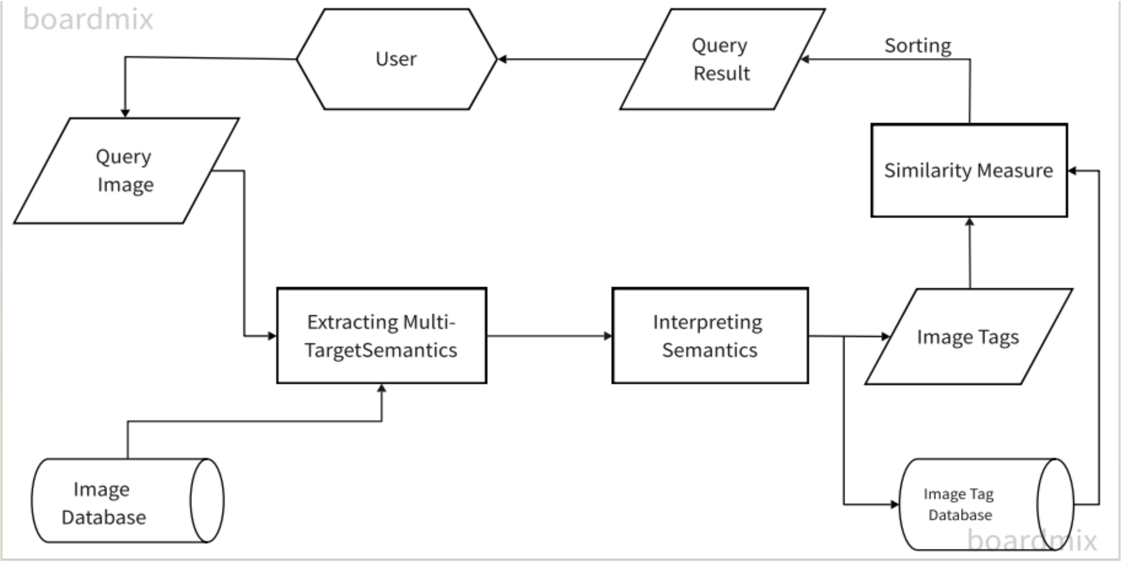


Figure4-1. Principle of Learning Model

The gap between abstract semantics and multimedia data semantics remains a significant challenge in image retrieval and semantic understanding. Abstract semantics, as defined and extracted, often represent only a portion of the total abstract semantics contained within multimedia data. This discrepancy arises because different individuals may interpret the same scene differently, and each person may only perceive or understand a subset of the abstract semantics involved. As a result, user feedback plays a crucial role in refining and enhancing semantic retrieval systems.

Bridging the gap between abstract semantics and multimedia data semantics requires a multi-faceted approach that integrates advanced technologies, multi-level correlation feedback, and user-centric mechanisms. By focusing on both horizontal and vertical integration, future research can more effectively narrow the semantic gap in image retrieval, ultimately achieving the goal of semantic retrieval. This will enable systems to better align with human perception and interpretation, providing more accurate and meaningful results.

REFERENCES

[1] G. E. Hinton, S. Osindero, Y. W. Teh. A fast learning algorithm for deep belief nets[J]. Neural Computation, 2006, 18(7):1527–1554

[2] Y. Bengio, A. C. Courville, P. Vincent. Representation Learning: A Review and New Perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8):1798– 1828

[3] I. Goodfellow, Y. Bengio, A. Courville. Deep Learning[M]. MIT Press, 2016

[4] C. Singh, K. P. Kaur. A Fast and Efficient Image Retrieval System Based on Color and Texture Features[J]. Journal of Visual Communication Image Representation, 2016, 41

[5] M. M. Alkhawlani, M. Elmogy, H. M. Elbakry. Content-Based Image Retrieval using Local Features Descriptors and Bag-of-Visual Words[J]. International Journal of Advanced Computer Science and Applications, 2015, 6(9)

[6] X. Glorot, A. Bordes, Y. Bengio. Deep Sparse Rectifier Neural Networks[J]. 2011, 15:315–323

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks[J]. 2012:1097–1105

[8] Y.Lecun,L.Bottou,Y.Bengio,etal. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278–2324

[9] Y. Lecun, Y. Bengio, G. Hinton. Deep learning[J]. Nature, 2015, 521(7553):436

[10] K. Lin, H. F. Yang, J. H. Hsiao, et al. Deep learning of binary hash codes for fast image retrieval[M]. 2015

[11] W. J. Li, S. Wang, W. C. Kang. Feature Learning based Deep Supervised Hashing with Pairwise Labels[M]. 2016

[12] F. Zhao, Y. Huang, L. Wang, et al. Deep semantic ranking based hashing for multi-label image retrieval[J]. computer vision and pattern recognition, 2015:1556–1564

[13] F.Schroff,D.Kalenichenko,J.Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering[J]. CoRR, 2015, abs/1503.03832

[14] A. Mojsilovic, J. Hu. A method for color content matching of images[M]. 2000, 649–652 vol.2

[15] T. Mikolov, M. Karafiat, L. Burget, et al. Recurrent neural network based language model[J]. 2010:1045–1048

[16] S. Hochreiter, J. Schmidhuber. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735–1780

[17] T. P. Vogels, K. Rajan, L. F. Abbott. Neural Network Dynamics[J]. Annual Review of Neuroscience, 2005, 28(0):357–376

[18] D. E. Rumelhart, G. E. Hinton, R. J. Williams. Learning internal representations by error propagation[M]. MIT Press, 1988, 318–362

[19] J.Redmon,S.K.Divvala,R.B.Girshick,etal. YouOnlyLookOnce: Unified, Real-TimeObject Detection[J]. computer vision and pattern recognition, 2016:779–788

[20] J. Redmon, A. Farhadi. YOLO9000: Better, Faster, Stronger[J]. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017

[21] G. H. Liu. Content-based image retrieval using computational visual attention model(C)[J]. Pattern Recognition, 2015, 48(8):2554–2566

[22] I. Sutskever, O. Vinyals, Q. V. Le. Sequence to Sequence Learning with Neural Networks[M]. 2014

[23]J.R.R.Uijlings,K.E.A.V.DeSande,T.Gevers,etal.SelectiveSearchforObjectRecognition[J]. International Journal of Computer Vision, 2013, 104(2):154–171

[24] S. Ioffe, C. Szegedy. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. international conference on machine learning, 2015:448–456

[25] Z. Wang, X. Wang, G. Wang. Learning Fine-grained Features via a CNN Tree for Large-scale Classification[J]. arXiv: Computer Vision and Pattern Recognition, 2015

[26] B. Xu, N. Wang, T. Chen, et al. Empirical Evaluation of Rectified Activations in Convolutional Network[J]. arXiv: Learning, 2015

[27] J. Chung, C. Gulcehre, K. H. Cho, et al. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling[J]. Eprint Arxiv, 2014

[28] S. C. Wong, A. Gatt, V. Stamatescu, et al. Understanding Data Augmentation for Classification: When to Warp?[J]. digital image computing techniques and applications, 2016:1–6

[29] R. Pascanu, T. Mikolov, Y. Bengio. On the difficulty of training recurrent neural networks[J]. international conference on machine learning, 2013:1310–1318

[30] O. Vinyals, A. Toshev, S. Bengio, et al. Show and tell: A neural image caption generator[J]. computer vision and pattern recognition, 2015:3156–3164

[32] D. Ganguly, D. Roy, M. Mitra, et al. Word Embedding based Generalized La..0.00nguage Model for Information Retrieval[J]. 2015:795–798

[33] K. He, X. Zhang, S. Ren, et al. Deep Residual Learning for Image Recognition[J]. CoRR, 2015, abs/1512.03385