Retraction Notice

The Editor-in-Chief and the publisher have retracted this article, which was submitted as part of a guest-edited special section. An investigation uncovered evidence of systematic manipulation of the publication process, including compromised peer review. The Editor and publisher no longer have confidence in the results and conclusions of the article.

ZC and LZ agreed with retraction.

Mimicking human vision systems: deep-learning-based feature fusion for semantic image retrieval

Zhongzhe Chen[a,*] **and Luming Zhang**[a]
[a]Jinhua Polytechnic, Key Laboratory of Crop Harvesting Equipment Technology of Zhejiang Province, Jinhua, China

Abstract. Cross-form feature combination is an important multifeature fusion technique where the purpose is to implicitly discover the relationship between samples from different modalities, i.e., to retrieve another image encoded by similar semantics through one example image. In the past decade, cross-modal image retrieval has becoming a hotspot investigated by many academicians. Moreover, it is now a significant tool for the future performance enhancement of image retrieval. A long-short term memory (LSTM)-based feature fusion model is proposed. First, aiming at the competitiveness of nonmixed deep architecture for image retrieval, the mechanism of LSTM is introduced in detail. Among them, ground-truth-based methods are used to improve cross-modality. We notice that LSTM can mimic human visual understanding of image semantics well. To improve the accuracy of oblique-form image retrieval, systems based on binary representation are proposed to improve cross-modal similarity and effectiveness of message recovery. Second, we use a quality model to measure the commonly used image low-/high-level visual features, where the disqualified features are abandoned accordingly. This in turn achieves an optimal set of highly descriptive features for image retrieval. Furthermore, we use LSTM and the refined visual features to build a biological model for image retrieval, wherein the multimodel features can be optimally incorporated at the temporal level. Extensive experimental validations on multiple well-known image sets have shown the superiority of our method. © *2023 SPIE and IS&T* [DOI: 10.1117/1.JEI.32.6.062509]

Keywords: multimodal; feature fusion; deep learning; long-short term memory; image retrieval.

Paper 221335SS received Nov. 20, 2022; accepted for publication Jan. 11, 2023; published online Feb. 6, 2023.

1 Introduction

Many forms of image retrieval models attempt to understand the multiple high-level intelligences from multiple visual channels. In the past decade, owing to the rapid development of deep/shallow learning architectures, multimodal feature learning has becoming a hot topic since humans naturally support multichannel visual cues. Yong et al.[1] proposed to apply the state-of-the-art computer vision techniques on the tongue language, which was subsequently leveraged in the Acoustic Optical Tongue Notification project and has became a universal multimodal prototype. Afterward, multimodal information processing has undergone a long evolution. Hinton et al.[2] classified the existing techniques of multimodal fusion with the collaborative learning and fusion level into multiple types. Srivastava and Salahutdinov[3] discussed a multimodal hidden restricted random extension to improve multimodal feature classification. Hermann et al.[4] proposed an AI-guided multiform hashing with feature discriminative ability regularization system based on rectangular constraints to weaken the information pretemporal continuum of multimodal representations. Mao et al.[5] divided the mainstream topics of multimodal learning into multimodal vision, multimodal interpretation, multimodal alignment, multimodal joint, and multimodal collaborative perception are proposed accordingly. More recently, multimodal learning has successfully become an indispensable technique to enhance applications, such as face

*Address all correspondence to Zhongzhe Chen, 20151030@jhc.edu.cn

1017-9909/2023/$28.00 © 2023 SPIE and IS&T

recognition, human peacock estimation, multimodal retrieval, misfortune-form recovery, semantic display understanding, and tremor notification.

Cross-modal learning means to optimally combine multiple types of features to a unified framework. It can be deemed as a subtopic of multimodal feature learning that seamlessly integrate deep learning to calculate the interform representation, transformation, and alignment strategies in multimodal learning. Cross-modality learning is similar to multiform fusion in that both handle data from all modalities. However, the difference is that the data for the former is only preserved in a certain modality, whereas the data for the latter is tailored for a particular modality. Here, cross-modal image retrieval is one of the standard image-based applications toward cross-modal learning, also known as cross-media retrieval. Available modalities[6] are constructed from that the purpose of cross-modal retrieval is to describe the instruction interaction between pairwise different modalities. In the stage of recalling failed retrieved images, the corresponding techniques have gradually become the frontier as well as the hotspot of the researchers. Nowadays, it is becoming a significant information management tool for the future development of intelligent cities. Meanwhile, cross-modal similarity and textbook retrieval is an important research direction in cross-modal retrieval nowadays.

For adversarial modal similarity and textbook recovery tasks, keyword-to-image retrieval can be deemed a pseudo-"peevish-formal" problem since its key objective is to carefully calculate the similarity between keywords and the revealed semantic labels. It supports cross-modal similarity identification of visual data and local highly similarly samples well. It can investigate and model the interaction of both modalities, actors, and subjects. Its objective is to recover image semantics by subject (display) doubts without any useful information. Cross-modal image semantics and topic recovery can be divided into two stages: picture search and textual labels calculation. Lecun et al.[7] believe that the semantic relationship between images and text can be divided into the extended eight categories, that is, nonessential relationship, complementary relationship, interdependence relationship, bifurcation relationship, diagram relationship, contrast relationship, bad diagram relationship, and bad bifurcation relationship.

Aiming at characterizing the complex semantic interaction between actors and texts, unwritten image semantics recovery primarily adopts some statistical analytic models, such as canonical relation analysis and dysphoric modal substitution analysis (peevish-modal component analysis). These methods can reduce the dimension of the original features well by optimally fusing multiple features. Modal component analysis[8] is very effective in modeling complex relationships of different forms of data in many real-world scenarios. Rasiwasia et al.[9] designed the collaborative modeling of messages and concepts in multimedia information and subsequently leveraged the orthodox relational analysis to learn the analogies between the two modalities. However, the multimodal learning in Ref. 9 is substantially a linear map, which cannot accurately formulate other channel data. This can be reflected in the higher modality relevance.

More recently, the popularity of deeply learned vision models has become the state-of-the-art alternative for the previous shallow cross-modal presentation and information retrieval. This trend has gradually become hot and mainstream in computer vision nowadays.[10] On the one hand, compared with the previous shallow methods, deep networks are more suitable and effective for mining intramodality features and intermodality semantic relations due to their highly nonlinear construction. The processing of massive-scale data is typically leveraged in many deep models.[11] Cross-modality techniques improve textbook restoration studies based on complex feature relationships mining, which have received lots of attention for the remarkable achievements.

Using multimodal content, such as visual and contextual information, can prevent DSRs from the misunderstanding of image semantics. The distribution of recent deep models with multimodal similarity-supported metrics outperform the standard multimodal style understanding. Ji et al.[8] proposed the application of long-short term memory (LSTM)-guided image semantics understanding. Wang et al.[9] proposed a robotic system that can summarize human-based native language instructions to translate everyday oppositions. Feng et al.[10] established a semantic analysis system that supports human behavior and clothing learning and recommendation. Feng et al.[11] proposed to learn a convolutional neural network (CNN)-supported object-prescribed capability recognition course for multimodal humanized computer interaction.

Based on existing methods for facilitating natural speech understanding, combined with deep knowledge and multimodal advertising, we in this paper propose a hybrid and nonhybrid knowledge-based multimodal image semantics understanding and retrieval framework. It focuses on predicting the entire closed target-origin supports and obtains a choice of section order for each butted design pair. Finally, Generative adversarial network (GAN) is leveraged to retrain and update the training data to transform the accuracy of target object prediction in the image domain. The experimental results have shown that the proposed method can change the rationality of the Jiayi robot's command awareness. In addition, the image retrieval performance is outstanding by comparing with a set of state-of-the-art deep models.

2 Related Work

Among the methods designed using deeply learned feature representations, some researchers proposed a learning framework based on the maximum probability rule to optimize the deep neural network parameters through backpropagation and stochastic gradient descent. For modal exact shapes, deep learning[12] leverages the modally determined dense structure (MSDS) scheme. The model combines the CNN and WCNN architectures to accurately characterize image and textual representations. They further update the deep model parameters of CNN and WCNN through the enhanced standard back-propagation paradigms. WCNN can model sequences of different lengths in order to acquire the feature vectors with the same dimension.[4] This can effectively encode the shape feature of various objects. Experimental results have shown that such modality-specific deep feature engineering can intrinsically better learn modality representations in large-scale datasets. In this way, the so-called WCNN exhibits the capacity for textual feature extraction, which is more effective than that of deep CNN. Furthermore, on the basis of Meiwen,[8–11,13–17] Goodfellow et al.[6] proposed a method that supports deep bidirectional visual feature learning model. In detail, it leverages the appearance and textual channels in message transmission to improve the descriptiveness of deep features.

Bidirectional structure understands the so-called counting lines, wherein the unmatched textual feature sets are incorporated into the similarity of counting errors. Experimental comparisons have shown that bidirectional deep features are more expressive than the single-line learned mismatch pairs. It can guide the learning of highly discriminative information in the training data. For the cross-channel concept-topic recovery problem of single-drop toward multilabel samples, in order to bridge the gap between features learned from inconsistent semantic channels, feature engineering[18] using deep convolutional activation forms (DeCAF) is proposed to learn the indicator words. Herein, the 1000-dimensional predicted deeply encoded feature produced by the CNN implementation is considered as the input optical features trained from the well-known ImageNet. Experimental evaluations have validated that DeCAF can facilitate the learned deep architectures to learn highly descriptive features, wherein the feature extractor is highly effective. Since the previously learned CNN architecture can be directly utilized to learn image deep features, for the same settings, Lecun et al.[18] and the previously trained CNN model were proposed to produce a deep semantic feature (unmixed semantic matching, cunning-SM). The shield dataset is treated as a separate loss function, associated with an elegantly designed CNN and a fully connected neural network, projects images and textual information into the descriptive semantic space. Empirical results have demonstrated that the fine-tuning technique can correct its adaptability onto different datasets. This can effectively reduce the gap between appearance and consistent semantics. Lecun et al.[7] leveraged the same observation to calculate feature embeddings by applying the adjustments to enhance deep CNN engineering, which effectively avoid the possible noises from the semantic feature channel.

Among the feature representation-based methods, some researchers have established a Lore framework based on the maximum fidelity criterion to make optimal mesh parameters through backpropagation and stochastic gradient descent. For mode-specific shapes, Goodfellow et al.[12] intended to use a mode-specific depth configuration (MSDS) model. The bifurcation uses CNN and WCNN to partially extract image and SMS representations, and update the parameters of CNN and WCNN through standard back-diffusion techniques worn. WCNN can process sequences of different lengths and obtain ascending feature vectors with the same metric,[4] which

can refer to the textual form thoroughly. Experiments show that modality-specific feature learning can revise and extract input modality representations in large-scale datasets, and that the topic feature extraction ability of WCNN is better than that of deep CNN. Furthermore, on the basis of Goodfellow et al.,[12] belles-lettres[6] transformed an approach to support a literature model of deep bidirectional representation, which augments features with Rosalia and textural directives in textual descriptions, and uses a bidirectional structure Learn about matched and unmatched similar text. Yoke relationship, increasing the similarity of twin (former name) pairs. Experimental comparisons show that the bidirectional representation fork has the correct effect than the unidirectional form for imperfectly matched impairments, and the pattern can learn rich discriminative teaching in the semantic space.

When multimodal technique is leveraged to bridge the gap between concepts and their corresponding semantic concepts, for vision tasks like topic retrieval of split-label or multicategory samples for latent modality understanding, Ayyavaraiah and Venkateswarlu[18] designed an intelligent distorted incentive form index algorithm. It can obtain 1000 variants of CNN. Dimensional reduction algorithms are leveraged as the input features trained by leveraging ImageNet. Experiments have shown that the proposed DeCAF can make the learned optical features with qualified portrait descriptiveness. Noticeably, the feature extraction implementation is commendable for this algorithm. Since the pretrained CNN models can be practically transferable to image semantic understanding. For the same issue,[18] we fine-tuned the pretrained CNN models and designed a novel fuzzy semantic matching model (deep-SM). Different multimodal image datasets typically employ different visual semantic understanding models, by leveraging the upgraded CNNs and training neural networks to project images and text into a highly descriptive and homogeneous semantic space. Comprehensive experimental results have shown that the fine-tuning technique can substantially enhance its application to the target dataset. Moreover, it can effectively overcome the gap between similarity and perceptual semantics. Lecun et al.[7] leveraged the same idea to produce the visible embeddings of portraits by fine-tuning a deep CNN architecture. This technique can strongly avoid the noises inside some semantic intelligence.

3 Our Proposed Method

This paper leverages a hybrid deep learning multimodal natural language understanding approach to optimally understand and refine the predicted semantic labels. Our proposed method can predict and display all the image semantic labels and images pairs based on the given information well, by mimicking human temporal visual information perception. It can manage the predicted data inferior to GAN to correct semantic labels for classification fidelity. Our algorithm uses the indicative feature matrix as the input, where the various scenery pictures are leveraged as the input. It further uses the possible area as the target-ascent pair, where the misclassified result refers to the appearance of the image (e.g., bolt head or progeny) acquired using LSTM. Here, the source sends to the target's ascent data using LSTM (such as images or their noisy labels).

LSTM has its own forward and backward propagation learning mechanism, which can better capture the context information of the unidirectional LSTM. LSTM is an actively learned discontinuous neural fret (RNN) that is comprised by the input gates, memory gates, forget gates, and product gates. These can constitute the opt-in information that occurs through logging or deletion operations. The LSTM structure for unit function t is elaborated in the following:

$$f_t = \sigma(W_f \cdot [h_{t-1}, f_t] + \tau_f), \tag{1}$$

where W_f represents the memory one, it represents the input gate, f_t represents the forgetting gait, π describes the production path, h displays the input sequence, and σ represents the median transition content. In this equation, the product of the input gate are denoted by C_t. This formulation ignores the generation of the mode f_t and the generation of the memory content C_{t-1} of $t-1$. Here, we incorporate the above two terms to quantify the importance of the memory unit C_t.

A GAN can be considered as a deep learning framework that estimates the generative model through an adversarial process. It is strongly biologically inspired and has nowadays been pervasively utilized in image understanding. At present, more and more researchers are devoted to natural image semantics interpretation. The basic framework of GAN consists of two key components: the generative component G and the discriminative component D. The generative deep network G is trained using the real data block x to breed a refurbished data pattern $G(z)$. Meanwhile, the discriminative the mesh D is a binary classifier. According to the data x of the present feature engineering, $G(z)$ is formulated using the generative mode. Here, the objective function of GAN is formulated as

$$(D, G) = E_x * P(x)[\log\ D(x)] + Ez, P(z)[r(1 - D(G(z)))], \tag{2}$$

where z denotes the image representation, x denotes the true match, P denotes function as the probability distribution, and E represents the mean value. The first term accumulates all the image features here, whereas the last term calculates the weights of different image representations. This equation is the key to the GAN architecture.

CNN, as one of fundamental model in deep learning, concludes an active neural network with convolutional process and deep construction. This model is naturally biologically inspired. More specifically, CNN consists of convolutional layers, pooling bases, and fully joint boosting. The convolutional components are leveraged in the original target data set. The pooling layers are seamlessly combined with the convolution layer. Appearances are local regions of a shape where the shape involves some spatial invariance. In our image understanding framework, the effective VGG19 network structure in ImageNet is the practically based network structure. The main contribution of the VGG19 deep architecture is the effectiveness of the very small 3×3 convolution kernels.

Features extracted from LSTMs can be described as follows. The two multilayer perceptrons (MLPs) are useful in feature learning. After that, we customize the learned MLP to predict the position for the target object supported by the language and vision MLP product. Noticeably, word embedding is a competent modeling technique in natural text understanding. Commonly used information embedding methods include Word2vec. After the word embedding, the communication embedding shape is leveraged as the input of the bidirectional LSTM. In this work, the subword processing model BERT[14] is leveraged to initialize the embedding vector instead of the account-based embedding scheme. Actually, the BERT model is a language encoding method supported by a bidirectional transformer with higher flexibility and robustness. BERT is ante-exercise on 3.5 billion words. In this way, the data are uncommon for those rare words. Additionally, BERT does not rely on word-backed tokenization but subword tokenization. This attribute is more effective against word misspellings. The BERT standard message feature transmission is the input to those multiagent Bi-LSTM. The method in this work leverages the upgraded Bi-LSTM to encode language semantics. This can revolutionize the contextual information of opinions. Meanwhile, a 19-foot VGG19 is incorporated to encode the visual features of the backgrounds. These backgrounds closely related to a multiband perceptron network in two deep neural networks, namely MLP_I and MLP_V, whose output is utilized to predict the probability of an object. The outputs OI of the two MLPs represent visual features and OV represent glottal features. Here, the prediction of the deep model is maintained from OI and OV via MLP_U operations.

In order for our adopted LSTM to predict the accuracy of target objects in image semantic understanding tasks, GANs are interested in augmenting the training data. The GAN framework contains two adversarial clusters, a generator G and a discriminator D. The generator G generates artificial data by simulating the given data distribution, wherein the discriminator D predicts whether the input data are real or unreal. With its unbiased adversarial deep neural network, G is trained to produce more positive data while enhancing D's discriminative ability. The GAN has three inputs: the language shape O_I, the nonuniform input O_V, and a multidimensional input z randomly sampled from an exact assignment. To classify the testing data x_{real} and the testing fake data x_{fake}, from source data, alternately input discriminator D, the output of D can be obtained by $D(x) = PD(S = x_{\text{real}}|x)$. Here, the failure functions of G and D are denoted by

J_G and J_D, respectively. During the visual semantic understanding, the training of D and G was calculated iteratively.

4 Experimental Results

In the experimental evaluation, the parameters of the experimental settings are illustrated in the first place, the 24-layer fully connected BERT model is leveraged for gate-message transmission, and the dimensionality of the deep feature vector is 1024. The VGG19 pretraining model is deployed to learn the CNN. In MLP-1 and MLP-V, batch normalization and ReLU activation function are leveraged to model each layer. In MLP-S, in addition to applying ReLU to accelerate calculation, the last layer also leverages the Softmax function. Both the generator G and the discriminator D in the GAN consist of deep learning using four layers of ReLU activation functions, where batch normalization is adopted to these operations. The product layer of G is distributed in the semantic space, where the Softmax function is superimposed with the output layer of discriminator D. Our semantic understanding model's parameters are set as follows: $\lambda_1 = 1$ and $\lambda_2 = 0.7$. Meanwhile, the rest parameters are tuned based on our cross validation.

To evaluate the performance of our method in the real-world human semantic understanding, our method is applied to the PFN-PIC dataset, which has 89,861 axioms and 25,517 bounding boxes in the image set. It includes 898 sceneries and 352 bounding boxes in the validation image set. It demonstrated that under conditions of positive and negative sample cost γ, the calculation of proposed bioinspired LSTM using the BERT architecture. However, the deep feature is not derived in the standard BERT way, which hurts the accuracy of feature classification. This can be achieved by using a bidirectional transformer-based language encoding BERT framework. This makes the parsing of image semantics more accurate and the deep model will be much more complicated. Such a long BERT has 3.5 billion pretrained parameters in total. Thus the data for the semantically learned words will be huge. Furthermore, the learned LSTM model is supported by the noise-tolerant features. This makes the learned feature highly descriptive. This is more robust to handle the spelling errors. In order to evaluate the effectiveness of our deep features, the method in this experiment is obtained with other existing methods, understanding: CNN + LSTM complex erudition mode,[9] rules and knowledge-supported semantic analysis mode,[10] CNN-based deep feature engineering of human–computer interaction are leveraged for multimodal fusion.[11] The results are presented in Table 1.

We give the brief introduction for all the compared methods under the same confident and matching criteria γ. It is observable that the method proposed by us has the highest accuracy in predicting the semantics of images in the natural language acquisition task under different γ. By the way, according to Ref. 9, we can explain why the multimodal algorithm[10] leverages a highly competitive approach; this experiment[11] employs a CNN framework to categorize visual semantics from a unimodal method. We subsequently leverage[9] a CNN + LSTM regularity to fuse features from multiple channels and scenarios. Here, our method adopts a CNN + Bi-LSTM + GAN hybrid technique.

Thorough scientific approach, especially to encode and predict glottal shape and language-free features, and influence GAN data expansion have shown the usefulness of our feature fusion. Here, the target object semantics predictions is conducted. In addition, our proposed

Table 1 Comparison of different feature fusion techniques in image retrieval.

Model	Recall at 1	Recall at 6	Recall at 30
CNN	0.021	0.075	0.122
LSTM	0.032	0.103	0.212
CNN + LSTM	0.112	0.154	0.324
CAAN	0.213	0.183	0.332
Ours	**0.435**	**0.432**	**0.497**

Note: bold values represent the best performers.

feature fusion can receive a 99.8% semantics prediction accuracy. Except for the system described in this paper and the method in Ref. 9, which can give rise predictions, the other two methods can only predict the target semantics in a unimodal way. In order to receive a number of outcome of this multimodal feature fusion scenario, we noticed that collecting a set of real pictures can provide the requirement performance for the scheme. During the performance enhance of multiple model-based semantic understanding, we in this paper extract the natural language semantic interpretation and classification of some experimental data from the PFN-PIC data set. It can be concluded that the model in this paper achieves the highest performance. In this paper, we leverage the CNN + Bi-LSTM + GAN hybrid deep feature learning model. The training and learning time are significantly longer than the other three methods. However, the unimodal approach of the deep learning has the highest accuracy. The prediction accuracy of the different deep feature parts will converge to reduce the testing time of our multimodal feature fusion algorithm.

It can be concluded that the IMRAM algorithm achieves better performance in cross-format image and SMS recovery. The experimental context for the algorithm is the Pytorch v1.0, where 29,000, 1000, and 1000 samples are collected from the Flickr30k dataset. Images are leveraged for multimodal feature integration. Notably, for feature validation and testing, the 1000 similar samples are also leveraged for validation and testing on MSCOCO datasets, especially on image semantic datasets (Flickr30k) and large fine-grained visual classification dataset (MSCOCO). The high performance is achieved by leveraging the robustness of our algorithm and also demonstrates the necessity to explore deeply learned visual descriptors. Additionally, the CAAN algorithm program can confirm its usefulness by utilizing 29,000, 1000, and 1000 samples from the Flickr30k dataset. The algorithm receives an accuracy enhancement of 0.000 2 in the first 15 epochs and 2 in the last 15 epochs. Here, the PVSE algorithm leverages 113,287 images to create visual patterns on the MSCOCO dataset and is experienced on the full 5,000 experimental images. We generally employ >5 sets of 1000 images. The literature rate is 0.0366 and decreases when more visual pattern are used, and the training size is adjusted to 128 for 50 epochs. The CAAN algorithm performs satisfactorily and the PVSE algorithm also achieved good evaluation results on the small image dataset (Flickr30k) and massive dataset (MSCOCO), respectively.

This can show the advantages of performance enhancement in cross-modal image and textual channel combination. And these methods all emphasize the combination of topic features and deep semantic features to improve the discriminative ability of features. The feature channel weights are calculated to reduce the feature noises within this range. The results are presented in Table 2.

It can be seen that ACMR produces lots of correcting MAP noises on the dataset MSCOCO. The ACMR algorithm program leverages 66,226 and 16,557 image feature pairs for deep model training and testing, respectively. Here, the deep feature extraction used for experimentation is 4096 dimensions. VGGNet, the message shape lineage network is a 3000-dimensional BoW (pocket-of-words) with a batch size of 64. The algorithm exploits the adversarial attribute of GAN, which makes the underlying cross-modality semantic structure data better understood. The final results have shown that on the population-scale dataset (MSCOCO), in the MAP values, the actual value representation of knowledge generally outperforms the binary description scholarship approach. These results can clearly show the maturity of our designed method.

Among them, the CYC-DGH algorithm program achieves better termination on MSCOCO than other binary representation-based feature learning methods. The initial learning rate of the algorithm rule is 0.0002, the first 100 epochs remain unchanged, and the last 100 epochs are linearly reduced to 0. The ReLU component is combined with a dropout rate of 0.5. CYC-DGH

Table 2 Quantitative results of feature combination accuracy of the compared methods.

Methods	IOU at 0.1	IOU at 0.7	Recall at 1	Recall at 6	Recall at 30	Latency (ms)
Fully supervised	0.983	0.961	0.435	0.584	0.672	112
Weakly supervised	0.943	0.788	0.432	0.553	0.604	93

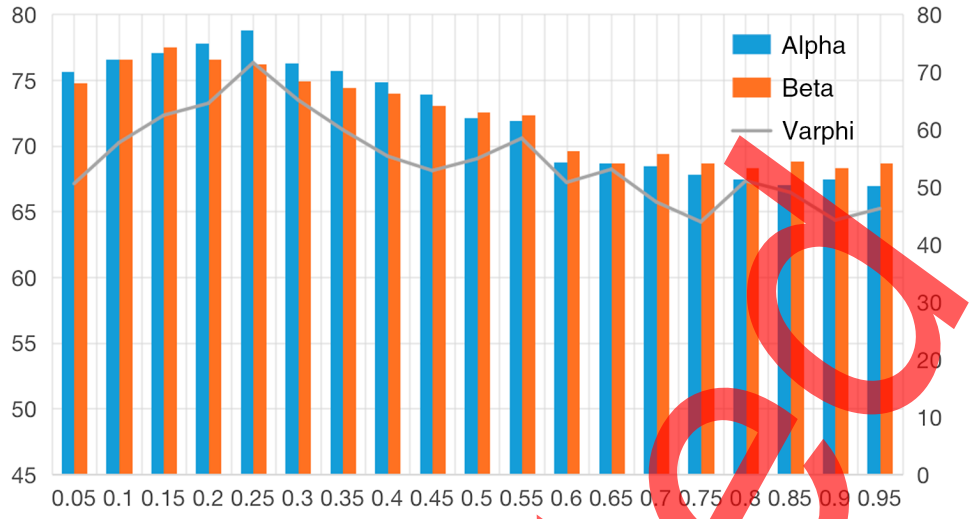

Fig. 1 Results of the CYC-DGH algorithm.

also utilizes the idea of GAN mesh network. This can explicitly exploit the relationship between its own information and samples from other modalities to the top feature. This can effectively enhance the input data. Therefore, it can be concluded that the concept of GAN is highly helpful for improving the performance of adversarial formal image and textual feature learning, and further produce a research model with appropriate value for feature enhancement. Statistics of them are presented in Fig. 1.

5 Discussion and Future Work

Humans can naturally perceive multichannel visual features. With the pervasive application of deep neural networks in image processing, multimodal feature engineering has received lots of attention in the literature. In order to improve the classification accuracy of multimodal feature learning, a bioinspired natural language instruction classification method based on hybrid deep learning is proposed. CNN encodes visual features and their correlation features. After the MLP processing, the prediction of target–source pairs is calculated. In order to improve the accuracy of NLP instruction classification by DSR, GAN is leveraged to expand and classification accuracy of the data. Extensive experimental results have shown that the proposed method improves the accuracy of target object prediction in semantic visual understanding task and its performance is better than other existing methods.

As the positive and negative sample rate increases, the accuracy of the proposed method for instruction classification increases. This verifies the feasibility of the proposed method and its high effectiveness. In practice, our method can be applied onto various tasks in computer vision. We plan to adopt an attention mechanism to enhance feature fusion in the future. In addition, we notice that the feature fusion is conducted only in a shallow mode. In the future, we plan to incorporate more deep learning components into our method. Further, more experiments will be designed to validate the pros/cons of our method.

References

1. L. Yong, L. Shaohe, and W. Xiaodong, "A review of human behavior perception technology based on WiFi signal," *Chin. J. Comput.* **42**(2), 231–251 (2019).
2. G. E. Hinton, S. Osindero, and Y. W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.* **18**(7), 1527–1554 (2014).
3. N. Srivastava and R. Salahutdinov, "Multimodal learning with deep Boltzmann machines," in *Proc. 26th Annu. Conf. Neural Inf. Process. Syst.*, Lake Tahoe, 3–6 December 2012 Curran Associates, Red Hook, pp. 2231–2239 (2012).

4. K. M. Hermann, T. Kociský, and E. Grefenstette, "Teaching machines to read and comprehend," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, 7–12 December 2015, Curran Associates, Red Hook, pp. 1693–1701 (2015).
5. J. H. Mao, W. Xu, and Y. Yang, "Explain images with multi-modal recurrent neural networks," arXiv:1410.1090 (2014).
6. I. J. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, Montreal, 8–13 December 2014, Curran Associates, Red Hook, pp. 2672–2680 (2014).
7. J. X. Gu, Z. H. Wang, and J. Kuen, "Recent advances in convolutional neural networks," *Pattern Recognit.* **77**, 354–377 (2018).
8. Z. Y. Ji, W. N. Yao, and W. Wei, "Deep multi-level semantic Hashing for cross-modal retrieval," *IEEE Access* **7**, 23667–23674 (2019).
9. C. Wang, H. J. Yang, and C. Meinel, "Deep semantic mapping for cross-modal retrieval," in *Proc. 27th Int. Conf. Tools with Artif. Intell.*, Vietrisul Mare, 9–11 November 2015, IEEE Computer Society, Washington, pp. 234–241 (2015).
10. F. X. Feng, X. J. Wang, and R. F. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 2014 ACM Int. Conf. Multimedia*, Orlando, 3–7 November 2014, ACM, New York, pp. 7–16 (2014).
11. F. X. Feng, *Deep Learning for Cross-Modal Retrieval*, Beijing University of Posts and Telecommunications, Beijing (2015).
12. I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, Cambridge, Massachusetts (2016).
13. N. Rasiwasia, J. C. Pereira, and E. Coviello, "A new approach to cross-modal multimedia retrieval," in *Proc.18th Int. Conf. Firenze*, 25–29 October 2010, ACM, New York, pp. 251–260 (2010).
14. K. Wang, Q. Yin, and W. Wang, "A comprehensive survey on cross-modal retrieval," arXiv:1607.06215 (2016).
15. Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circ. Syst. Video Technol.* **28**(9), 2372–2385 (2018).
16. Z. Y. Li, Z. F. Huang, and X. M. Xu, "A review of the cross-modal retrieval model and feature extraction based on representation learning," *J. China Soc. Sci. Tech. Inf.* **37**(4), 422–435 (2018).
17. M. Ayyavaraiah and B. Venkateswarlu, "Joint graph regularization based semantic analysis for cross-media retrieval: a systematic review," *Int. J. Eng. Technol.* **7**, 257–261 (2018).
18. Y. Lecun, L. Bottou, and Y. Bebgio, "Gradient-based learning applied to document recognition," *Proc. IEEE* **86**(11), 2278–2324 (1998).

Biographies of the authors are not available.