International Journal of Knowledge
and Language Processing
Volume 12, Number 2, Dec **2021**

AKLP International ©2021 ISSN 2191-2734
pp. 32–50

A Review of Multimodal Data Fusion based on Unsupervised Learning

Yuxiao Zhao
Institute of Scientific and Technical Information of China
No. 15 FuXing Road, Haidian District, 100038
Beijing, China
zhao-yx@foxmail.com

ABSTRACT: *In the context of growing multimedia data volume, how to automatically extract and parse relevant knowledge from the Internet and build a knowledge graph through fully automated and human-free algorithms or platforms, and then form a systematic knowledge graph, are of great significance to improve the existing knowledge graphs and build new domain knowledge graphs. This paper presents an extensive survey and summary of the current state of research on multimodal data fusion based on unsupervised learning. Starting from the development of unsupervised learning, this paper focuses on feature learning and fusion learning of unimodal and multimodal data, which can help readers to quickly establish the current state of research in this field and understand the shortcomings of current research.*

Keywords: multimodal data fusion, unsupervised learning, knowledge graph

1 Introduction. As the Internet and mobile communications continue to grow, the amount of data housed on the global Internet is growing at an explosive rate. According to the data released by IDC, the global data volume was 10ZB in 2010, while the global data volume has exceeded 40ZB in 2020 [1]. As the Internet develops rapidly in recent years, the growth rate of global data volume will reach new heights. The new concept of "Big Data" has become a research hotspot in various academic circles and has received high attention from governments and enterprises. The growing volume of data on the Internet provides valuable data resources for data analysis in various industries and facilitates deeper mining of universal laws and hidden knowledge.

The development of the Internet has made the variety and scale of data grow more rapidly, however, people cannot directly obtain a large amount of hidden knowledge from these complicated data, and the disordered data bring difficulties for computers to automatically identify and process data. Therefore, since Tim Berners Lee et al. [2] proposed the concept of Semantic Web, the Semantic Web has been widely used on the Internet. In the knowledge system of the Semantic Web, Ontology is a key concept that represents a formal, explicit, and detailed description of a shared concept system [3]. Ontology describes all concepts and relationships between concepts in the domain, including the attribute relationships of concepts and the subordination relationships of concepts. The Semantic Web gives a feasible direction for data structuring.

The rise of Web 2.0 has brought about a huge amount of User Generated Content (UGC), making the workload of search engines increase dramatically. To solve this problem, Google introduced the concept of Knowledge Graph (KG) in May 2012, which organizes and presents the entities, events as well as attributes and relationships in data in a two-dimensional graph-like structure[4]. The idea of "Things, not strings" is to describe real things in the real world and to extend the relevant concepts of the Semantic Web into more understandable and cognitive entities (Entities). In the knowledge graph, an entity can represent an object or concept, which is a point inside the graph, while the lines between entities represent the relationship between entities. As an enrichment and supplement to the Semantic Web, Knowledge Graph reduces the cost of construction and use, and it is easy to build a Knowledge Graph ontology by just constructing a data schema and establishing various attribute information and relationship information of the ontology. Structured knowledge graphs can help people to better mine and acquire knowledge, and facilitate computers to automate processing.

Knowledge graphs have an important role in knowledge systems and automated intelligent services. Therefore, how to extract relevant knowledge from massive, complex, and heterogeneous raw data and construct a high-quality knowledge graph have become a research focus in the academic community for many years. Many universities and research institutions have developed proprietary automated knowledge graph construction systems based on different models and architectures. For example, T Mitchell et al. at Carnegie Mellon University [5] developed the fully automated information extraction system named "Never-Ending Language Learning (NELL)", which is a combination of four extraction subsystems: CPL, CSEAL, CMC, and RL. The CPL system can extract knowledge through contextual parsing; CSEAL can extract semi-structured data such as lists and tables from web pages; CMC can classify words according to a logistic regression model and extract knowledge through morphological features of different words; RL is responsible for logical reasoning, i.e., reasoning about new knowledge with the help of existing knowledge base. However, as S Russell [6] pointed out, the confidence level of the NELL system is very low and it relies heavily on manual error correction by domain experts and removal of meaningless knowledge systems.

The multi-source heterogeneity of big data poses a new challenge for in-depth analysis of the data, and multimodal data is a concept that focuses on solving the problem of diversity of big data. In early data processing work, people focus on solving some single-kind data types, such as text information, image information, video information, audio information, etc. However, the diversity of big data and the increasingly sophisticated information analysis theories have led the academic community to focus on the synergy from multiple kinds of data types to analyze the same thing or event from multiple perspectives. Yoshua Bengio et al. [7] argue that the learning effect of machine learning largely comes from extracting the features of data sources, which is also known as feature learning. The key to research now is how to extract good features from a dataset and how to evaluate these features. In Dana Lahat et al.'s view [8], the extraction of features from a dataset is the goal of multimodal data fusion. Multimodal data fusion is to study how to mine multimedia data and achieve comprehensive analysis across data types, and the core purpose of this comprehensive analysis is "complementarity", i.e., each modality brings an irreplaceable value to the overall analysis, thus enhancing the robustness and interpretability of the overall analysis. However, whether it is feature extraction or feature complementarity, the basis and key to multimodal data analysis is how to identify and align the same entities in multimodal data.

The major contributions of this paper are listed as follows.

(1) Introduce the related research on unsupervised learning

This paper introduces the related research on unsupervised learning and introduces the main research ideas and analysis directions of unsupervised learning, which helps readers to quickly become familiar with the main analysis ideas of unsupervised learning, and then apply them to multimodal data fusion analysis.

(2) Provide the current status of research on multimodal data fusion based on unsupervised learning

Based on multimodal data feature learning, this paper introduces research related to multimodal feature learning and multimodal knowledge fusion, and introduces several research ideas that elaborate on different approaches, which help readers better understand the research ideas and the current state of research on multimodal data fusion and analysis.

(3) Summarize the shortcomings and future developments of related research

Through the elaboration and analysis of existing studies, this paper summarizes the shortcomings of the current research on multimodal data fusion analysis based on unsupervised learning and proposes some possible future developments to provide readers with inspiration.

2 Overview. Knowledge graph brings the possibility of structured data analysis and semantic retrieval. Based on structured data analyzed by knowledge graphs, search engines can return more accurate answers instead of simple string matching. In addition, semantic data can also improve and optimize existing automated parsing and analysis tool services in the form of custom data sources, or assist in various complex analyses, research, and decision support, etc., to help people at various professional levels to carry out related work. As the study of semantic data deepens, existing data analysis processes are becoming increasingly difficult to meet the needs of professionals' work. The integration processing of multimodal data will become one of the research topics to improve or extend the existing semantic knowledge system. After reading the relevant literature, it is found that there are few studies on multimodal improvement of semantic knowledge systems in the academic field, and most of them use modal data such as pictures and videos to assist in the parsing and analysis of text data in a superficial way.

Since the concept of multimodal fusion was put forward, various methods have been proposed in the academic community to realize multimodal data fusion, which can be roughly divided into two ways: methods based on supervised learning, methods based on unsupervised learning, and methods based on few-shot learning. In the face of multimodal data, most of the traditional processing methods are to manually mark the same points among different modalities or use manual matching to divide the data of multiple modalities into the same group. Although manual processing has the feature of high flexibility, this approach also has its own disadvantages. For example, manual processing usually requires high human resources and small data size, and the standards relied on by manual processing are less transparent and difficult to be unified, etc. These disadvantages make it only applicable to small-scale sample data, and cannot cope with the increasingly large multimodal data. As a processing direction of multimodal data fusion, the core of unsupervised learning is to use the implicit knowledge contained in the data itself for learning and to mine and utilize deeper information through continuous optimization without human intervention in the whole process. Therefore, the research of multimodal data fusion based on unsupervised learning can help to process multimodal data in real-time, efficiently, and quickly, so as to improve the effectiveness of data usage and optimize decision support capability.

3 Status of unsupervised learning research. In traditional machine learning classification, machine learning models and methods can be classified into Supervised Learning and Unsupervised Learning, according to whether or not they require manually labeled information. Supervised learning uses a large amount of manually labeled data to train a model, and through the continuous input of training data and correct label pairs, the model learns through backpropagation, and eventually gains the ability to recognize features and relevant information in existing data, and to generalize to unknown new data. In contrast to supervised learning, unsupervised learning does not rely on any manual labeling information, but relies solely on the mining of features within the data to find relationships among data samples and complete tasks based on the relationships [9-10].

Traditional unsupervised learning is divided into two main categories: one is cluster analysis, such as Hierarchical Clustering (HCC) [11], K-Means [12-13], etc. The other category is dimensionality reduction analysis, such as Principal Components Analysis (PCA) [14], Singular Value Decomposition (SVD) [15], t-SNE [16], etc. As neural networks have entered the academy, more and more related researches have started to combine unsupervised learning and neural networks to build new models, expecting to achieve de-manualized and automated task processing with the help of more flexible, unsupervised deep learning methods.

3.1 AutoEncoder. The earliest introduction of neural networks into the field of unsupervised learning was the AutoEncoder constructed by DE Rumelhart et al. [17] and improved by DH Ballard [18]. The structure is shown in FIGURE 1. After continuous research and development in the academic community, AutoEncoder has evolved from a specific model approach to a research idea, that is, to construct a fully automatic encoding-decoding approach for neural networks, At the same time, the dimension of the hidden layer vector is much smaller than that of the input sample data, and enables the hidden layer vector to contain more basic features and semantic features of the data.

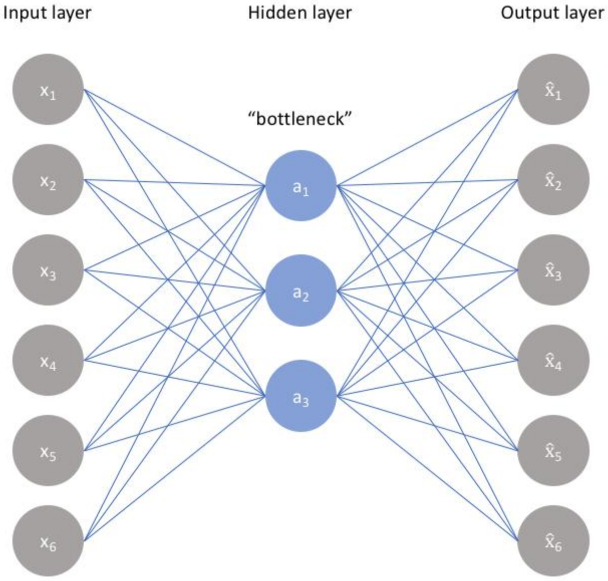


FIGURE 1 The structure of AutoEncoder

Unlike PCA, the AutoEncoder can characterize both linear and nonlinear transformations, greatly enhancing the flexibility of the model, whose goal is to learn a characterization function that allows the model to learn as many abstract features of the original data as possible. AutoEncoder can be classified into the following main categories.

(1) Denoising AutoEncoder

Denoising AutoEncoder is the most basic AutoEncoder, whose core purpose is to extract features on the original data with certain defects or noise, and the extracted features can repair the defects or remove the noise from the original data. Denoising AutoEncoder is widely used in image processing and other fields, such as Stacked Denoising Autoencoder (SDA) constructed by P Vincent et al.[19], which constructs Denoising AutoEncoder in the hidden layer using an unsupervised pre-training mechanism and fuses multiple Denoising AutoEncoder, improving the original Denoising AutoEncoder constructed by A Coates et al.[20].

(2) Sparse Autoencoder

Sparse AutoEncoder aims to discover the association between the original data features and the neural network nodes by reducing the activation of nodes in the hidden layer[21]. This training approach can force the neural network to use fewer nodes to accomplish the desired task, thus relying more on the features and hidden structures of the input sample data rather than the redundant information in the surface layer.

(3) Variational AutoEncoder

Variational AutoEncoder (VAE) is a generative neural network model based on Variational Bayes (VB) proposed by DP Kingma et al. [22]. Its character is to use probabilities for the observation of the potential space.

(4) Compressive AutoEncoder

Compressive AutoEncoder (CAE) features a regular term within its objective function that allows the model to obtain training data with subtle variations from the original data during training [23]. This input of training data allows the model to learn more about the hidden information contained in the data, while improving the robustness of the model mapping process.

3.2 Context-based unsupervised learning. The contextual information contained in the data itself is one of the features that can be learned by unsupervised learning models. For example, Word2vec [24-25], a classical model in the field of natural language processing, has the structure shown in FIGURE 2. Word2vec takes the contextual words of words into account when training the model parameters, and constructs the Skip-gram model that predicts the context of a word by its context and the CBOW model that predicts a word by its context. Combining these two models, Word2vec converts words into feature vectors that are used in various computational tasks downstream.

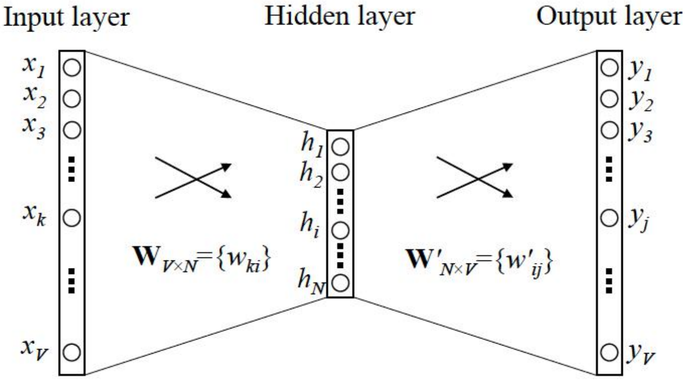


FIGURE 2 The structure of Word2vec

In the field of image processing, context encompasses multiple directions in the plane, including contextual similarity and spatial contextual structure. For example, R Zhang [26] constructed a "coloring Turing test" by building a model such as convolutional neural network and other models, and a new coloring method by a cross-channel encoder. M Caron et al.[27] proposed a new image feature extraction model, DeepCluster, which firstly divided the original image into several categories by image clustering, and applied the image-label mapping to the training and parameter updating of ConvNet, and achieved good results. C Doersch et al.[28] constructed a convolutional neural network to train the relative positions of two adjacent images with the help of the rich context information in the image space, trained the model to recognize the scene features, object features and some of the features, and applied them to the relative space reasoning, as shown in FIGURE 3.

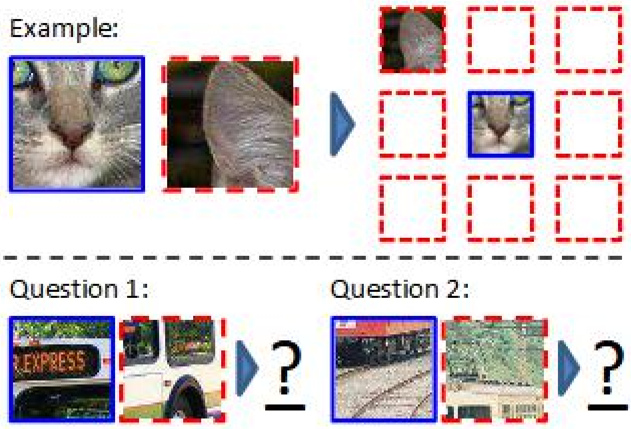


FIGURE 3 Relative space reasoning task [28]

Another type of task involving image contextual inference is missing image completion. Researchers remove a portion of an image and then train a model to achieve prediction and complementation. For example, D Pathak et al.[29] trained an encoder that converts image context information into feature vectors and a decoder that generates image data using feature vectors respectively, to construct an adversarial-like Encoder-Decoder structure, and combined them to accomplish the task of missing image completion. The results show that the model trained and learned by this task is able to understand the content of the images.

3.3 Time-series based unsupervised learning. A typical data modality for unsupervised learning based on temporal order, which uses information contained in data with temporal order as a constraint, is video data.

P Sermanet et al.[30] perform representation learning for video based on a parsimonious idea, that is, multiple frames adjacent to each other in time have more similarity, while frames far apart in time have less similarity. By constructing such similar frame pairs and dissimilar frame pairs, the Time-Contrastive Network (TCN) is trained to construct the semantic features of the video. W Zhu et al.[31] design the key object mining algorithm based on the information of image adjacent frames to identify the key video segments and give the classification results of the video segments at the same time. L Wang et al. [32] designed the Temporal Segment Network (TSN) based on image data to improve the learning effect of the convolutional neural network for action with the feature of image frame continuity. Ishan Misra et al.[33] extracted the video sequential information from the original temporal information in the video by constructing the correct temporal video data as positive samples and sampling the wrong of the temporal sequence video data as negative samples, and then inputting them into the convolutional neural network for training so that the model learned how to judge whether the video timing is correct or not.

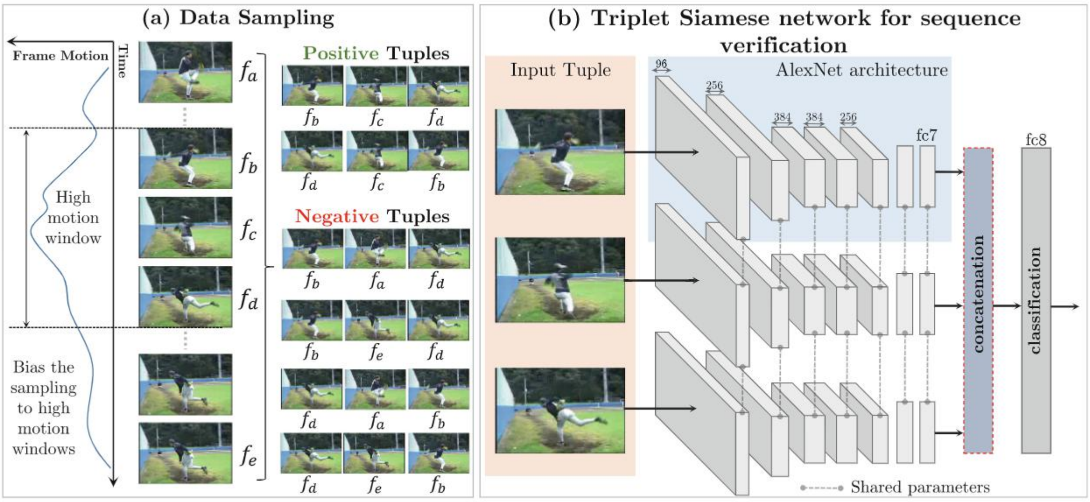


FIGURE 4 Video timing judgement model developed by Ishan Misra [33]

3.4 Comparison-based unsupervised learning. Contrast-based unsupervised learning is to discover and construct the characteristics of things by encoding and learning the similarity or dissimilarity of two things. Its core is to construct an eigenvector mapping function ***f*(*x*)** so that for any data ***x***, the following equation can be applied:

$$score(f(x), f(x^{+})) \gg score(f(x), f(x^{-})) \#(1)$$

Where x^{+} is the sample that is correlated with x (positive sample); x^{-} is the sample that is not correlated with x (negative sample). Thus, the neural network learns the features between the data by learning the differences among positive and negative samples.

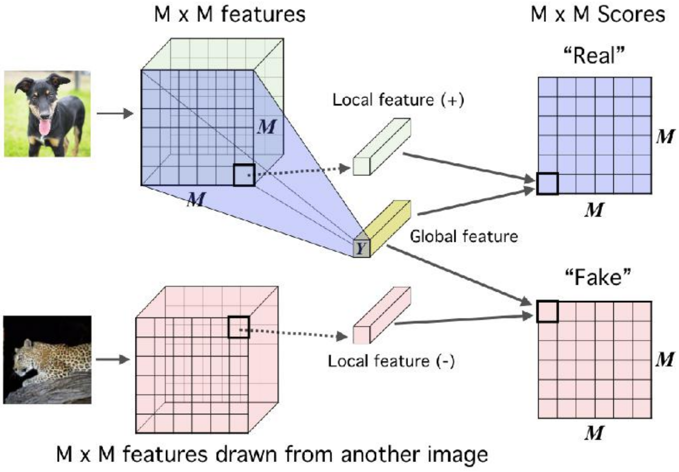


FIGURE 5 The structure of Deep InfoMax [34-35]

The Deep InfoMax model constructed by WL Hamilton et al. [34-35] is a typical contrast-based unsupervised learning, and its specific structure is shown in. Deep InfoMax uses both global features expressed as the final output of the encoder and local features expressed as the hidden layer of the encoder, and applies them to the learning of positive and negative image samples.

4 Current status of single-modal feature learning research based on unsupervised learning

4.1 Text Feature Learning. Among all the feature learning for modal data, the related research for text data is the most extensive and one of the focuses of the current tasks related to text computing. The core of text feature learning is to extract the implicit feature information in text data in the form of feature vectors for downstream model computation and analysis. At present, text feature learning based on the Latin alphabet is relatively mature, and there are a large number of models and corresponding open-source tools to support automatically generating feature vector representations from text data. However, the mainstream text feature learning models deal with Latin languages, and are less capable of handling non-Latin languages. As for Chinese, a typical non-Latin language, its minimum processing unit is a meaningful single Chinese character, rather than a word in Latin languages. Therefore, Chinese text feature learning is quite different from Latin text feature learning, and cannot simply use the same model for analysis and processing.

The prerequisite for text feature learning is the conversion of text into a representation that can be understood by a computer, i.e., a text representation. According to the model used, text representation can be divided into two types: statistical-based models and vector space-based models. Statistical-based models are represented by Boolean and bag-of-words models, which were commonly used for early text representation. With the development of text analysis, statistical-based models are gradually replaced by models based on vector space due to their discrete characteristics and difficulty in handling continuous text computation tasks.

Vector space-based models are the mainstream of text representation nowadays, such as Latent Dirichlet Allocation (LDA)[36-37] for unsupervised topic model generation methods, TF-IDF (Term Frequency-Inverse)[38] for paragraph and document representation, and so on. In 2003, Y Bengio et al.[39] proposed a method for building language models based on neural network models, creating a direction for creating text representations with the help of neural networks. The advantage is that the text representation obtained by neural networks is a low-dimensional dense vector, capable of representing semantic relevance. After introducing neural network models into text vectorization, it gradually became the mainstream direction for text feature learning because of its convenience, flexibility, and other advantages.

Word2vec, proposed by Mikolov et al. in 2013[26-27], is also a typical model of neural network models for text representation. After Word2vec, Q Le et al.[40] extended its idea to the whole document by directly treating document IDs as special words and putting in a large corpus for training, forming the Doc2vec model.

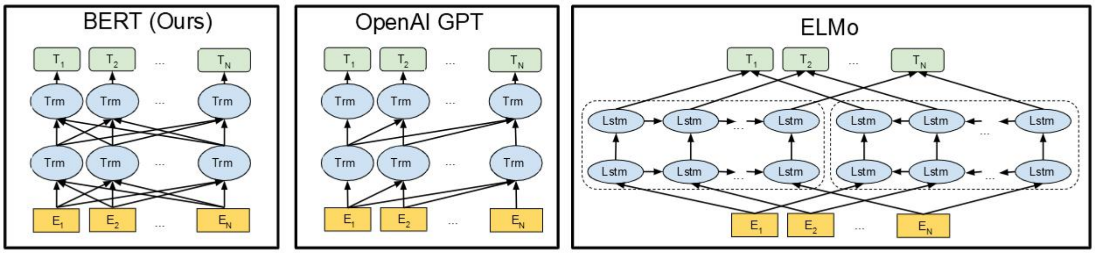


FIGURE 6 The structure of BERT, GPT and ELMo [42]

Word2vec and Doc2vec have made a large number of pre-training models a popular direction of research. ME Peters et al. [41] constructed the unsupervised text feature learning model ELMo (Embedding from Language Model) based on text features using bidirectional LSTM. J Devlin et al. [42] constructed BERT using a bidirectional Transformer model.

Based on existing models, there are many related kinds of research trying to fuse model approaches to improve the effectiveness of text feature learning. Piotr Bojanowski et al.[43] constructed FastText model for text classification based on the Skip-gram model and N-gram feature, which improves the speed of model training under large data and supports multiple language representation. Y. Kim et al.[44] fused the advantages of RNN in capturing dependencies and CNN in recognizing local features to build a TextCNN text classification model and achieved better results. P Zhou et al. [45] combined the bidirectional LSTM model and 2D convolution and proposed a BLSTM-2DCNN model to obtain a fixed-length vector representation of text data.

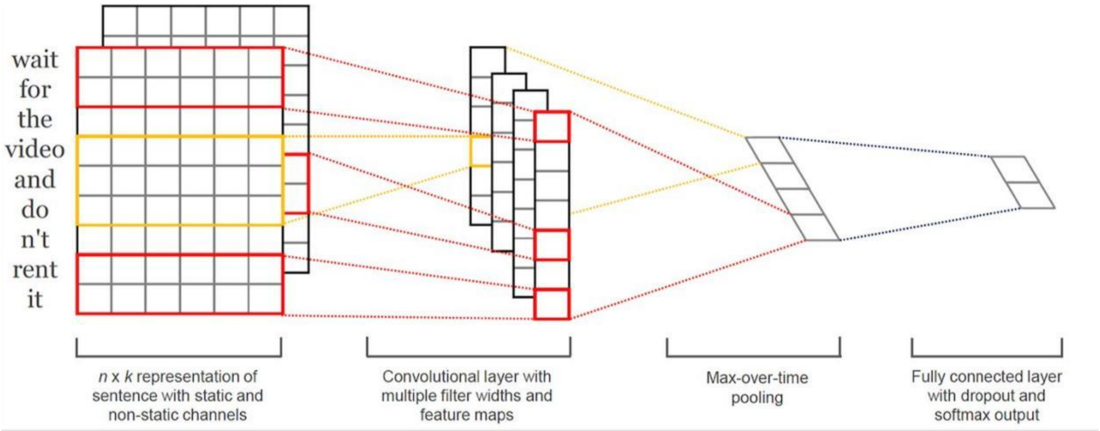

wait
for
the
video
and
do
n't
rent
it
n x k representation of sentence with static and non-static channels
Convolutional layer with multiple filter widths and feature maps
Max-over-time pooling
Fully connected layer with dropout and softmax output

FIGURE 7 The structure of TextCNN [44]

4.2 Image feature learning. Unlike text data, features of the smallest constituent pixel of image data are worthless. A feature has its value only if it has some structural or practical meaning. Traditional methods such as SIFT and HOG perform feature learning on images through the features of the histogram of map directions in the image. HOG feature units are smaller in size and less sensitive to local contrast changes in the image, and thus have better feature extraction for objects in complex environments; while SIFT features are usually calculated after a certain transformation of a square area in the image, and thus have good characteristics for rigid object feature extraction has good characteristics.

With the development of deep learning, image feature extraction through neural networks has become possible. Deep learning can better fit and extract the features of an image by self-learning, and its expression effect is generally higher than that of traditional image feature learning algorithms. For example, the concept of "image decoupling" proposed by R Zhang et al.[46] can be accomplished by splitting the original input image, such as dividing the original image into grayscale and color maps, and then predicting the image information from one part of the image information to another part of the image information, and finally synthesizing the predicted data. This is done by "decoupling" pairs of training data from the original data. This "self-learning" splitting of training data enables the model to understand and learn more about the semantic information contained in the image.

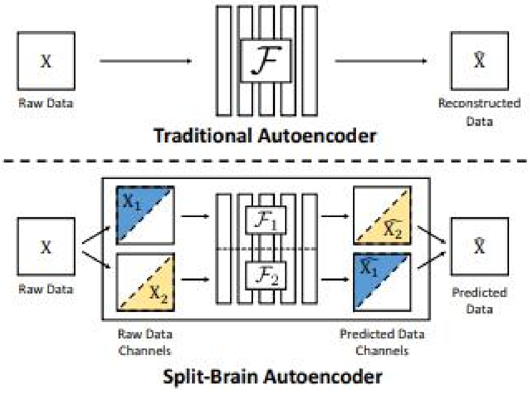
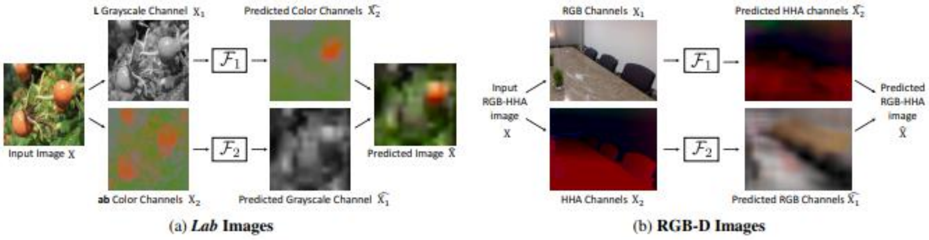


FIGURE 8 The structure of Split-Brain AutoEncoder [46]


L Grayscale Channel X_1
Predicted Color Channels $\hat{X}_2$
$\mathcal{F}_1$
$\mathcal{F}_2$
Input Image X
Predicted Image $\hat{X}$
ab Color Channels X_2
Predicted Grayscale Channel $\hat{X}_1$
RGB Channels X_1
Predicted HHA channels $\hat{X}_2$
Input RGB-HHA image X
Predicted RGB-HHA image $\hat{X}$
HHA Channels X_2
Predicted RGB Channels $\hat{X}_1$

(a) *Lab* Images (b) RGB-D Images

FIGURE 9 An example of the Split-Brain AutoEncoder [46]

A similar idea of "data augmentation" proposed by S Gidaris et al. [47] is also worth learning. In this paper, the input image is expanded into four images with different orientations by rotating the given input training image by 90 degrees, 180 degrees, 270 degrees, and itself, and passing them into ConvNet as a data set for training, which achieves an unexpected enhancement effect.

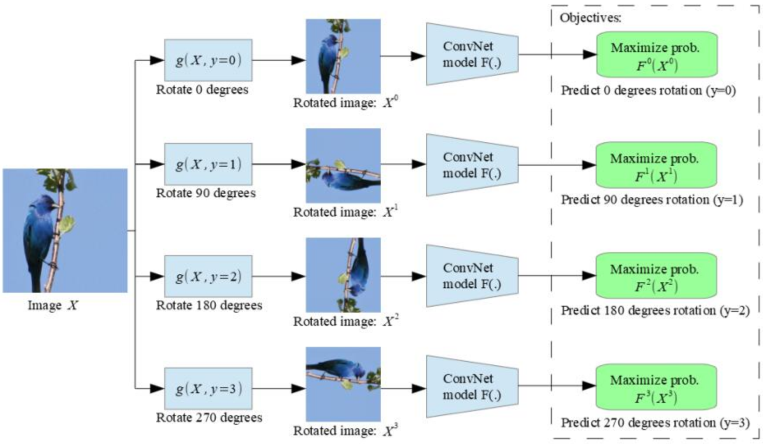


FIGURE 10 The self-supervised task designed by S Gidaris et al.[47]

5 Current status of research on multimodal feature learning based on unsupervised learning. While feature extraction for single-modal data is relatively mature, cross-modal analysis has become a focused area in the academic community today due to its importance in tasks including machine translation and object recognition. Current work on feature analysis of cross-modal data is mainly focused between text-based data and image-based data.

For unsupervised learning, multimodal feature learning mainly obtains information from co-occurrence, e.g., data of different modalities co-occurring in a document are related to each other, and thus may be semantically related.

Kaiye Wang et al.[48] broadly classified the unsupervised multimodal feature learning into three types: subspace mapping, topic modeling, and deep learning, and all three types can be combined into the same idea, i.e., by performing feature learning on text-based data, image-based data, and data from other modalities, respectively, and mapping them to common feature space for similarity computation.

C Sun et al.[49] performed the first joint entity extraction of text data and image data, paired known image data and its text labels in the process of image extraction, classified the image data according to the pre-processed text data, and compared them with the image similarity index. After a cross-sectional comparison with related models, the authors demonstrate that VCD (Visual Concept Discovery), a model for image similarity comparison after text data extraction and pairing, can effectively improve the efficiency of image similarity comparison.

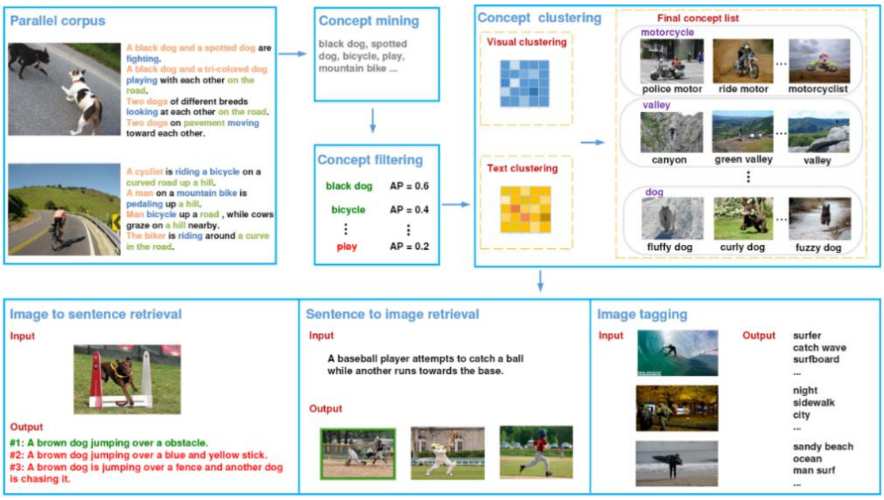

Parallel corpus
A black dog and a spotted dog are fighting.
A black dog and a tri-colored dog playing with each other on the road.
Two dogs of different breeds looking at each other on the road.
Two dogs on pavement moving toward each other.
A cyclist is riding a bicycle on a curved road up a hill.
A man on a mountain bike is pedaling up a hill.
Man bicycle up a road , while cows graze on a hill nearby.
The biker is riding around a curve in the road.
Concept mining
black dog, spotted dog, bicycle, play, mountain bike ...
Concept filtering
black dog AP = 0.6
bicycle AP = 0.4
play AP = 0.2
Concept clustering
Visual clustering
Text clustering
Final concept list
motorcycle
police motor
ride motor
motorcyclist
valley
canyon
green valley
valley
dog
fluffy dog
curly dog
fuzzy dog
Image to sentence retrieval
Input
Output
#1: A brown dog jumping over a obstacle.
#2: A brown dog jumping over a blue and yellow stick.
#3: A brown dog is jumping over a fence and another dog is chasing it.
Sentence to image retrieval
Input
A baseball player attempts to catch a ball while another runs towards the base.
Output
Image tagging
Input
Output
surfer
catch wave
surfboard
...
night
sidewalk
city
...
sandy beach
ocean
man surf
...

FIGURE 11 The structure of VCD

Q Fang et al. [50] combined and clustered classified image and text data, and used a clustering model for concept search, relationship extraction, and contextual relationship establishment, and the model significantly enhanced the effect of processing noisy and redundant labels. Y Zhu et al.[51] unified image data and its corresponding text by constructing a multimodal knowledge base structure, and creatively treated image data as the same level as text data, and then realized several popular functions such as a visual quiz.

6 Current state of Knowledge Fusion research based on unsupervised learning.

Knowledge Fusion is the downstream part of multimodal data fusion, and its core is to parse and map heterogeneous data from different sources, different modalities and different structures into a unified, single structured and interpretable semantic structure based on semantic features[52]. Its main operation is addition and deletion, i.e., adding new entities or relations to the semantic structure based on the parsed semantic knowledge, and removing duplicate or incorrect parts of the semantic structure to finally form a more complete and accurate knowledge graph. Multimodal data analysis is an indispensable part of a complete big data analysis process, and the difficulty lies precisely in the diversity of big data: the absence of modality, the incomplete and uneven data of different modalities, and the high dimensional properties of modal data, which all bring new challenges to multimodal data fusion.

Knowledge Extraction is a link between semantic alignment and knowledge fusion. The goal of multimodal knowledge extraction is to achieve the extraction of knowledge from multimodal data, which has been pre-processed, by machine automation. According to the current research status in the academic community, knowledge extraction can be classified into three categories: Entity Extraction, Attribute Extraction, and Relationship Extraction.

The core of entity extraction is to identify and extract entities from data, which is also the premise of attribute extraction and relationship extraction. Knowledge extraction for modal data such as images is an emerging research direction in recent years.

In addition to entity extraction, attribute extraction and relationship extraction are also important parts of knowledge extraction. Most of the existing studies consider the attributes of an entity as a formal concept, which in turn enables the association of attribute extraction with entity extraction to reduce the complexity of attribute extraction. The IMGpedia system proposed by S Ferrada et al.[53] in 2017 extracts a large amount of image visualization information from the Wikimedia dataset and constructs 15 million visualizations based on these visual content descriptors with 450 million visual similarities between images. In 2019, Y Liu et al.[54] connected three knowledge graphs containing digital text and images through the relational words of Same-As and successfully implemented relational inference among different knowledge graphs.

Earlier feature-based multimodal data fusion work was to merge and analyze the features of all modalities directly, and then complete the subsequent classification, clustering, and prediction based on the merged new features. This can greatly simplify the problem of articulation between features of different modalities, but it also ignores the association between features of different modalities and generates more dependencies[55]. The subsequent alignment task transforms alignment into the computation of similarity between entities or events. HB Newcombe [56] and IP Fellegi et al.[57] further discretize similarity with continuous results into a triple classification problem, i.e., three classes of exact matches, partial matches, and mismatches.

$$\begin{array}{c} similarity \leq t_1 \rightarrow mismatch \\ \{t_1 \leq similarity \leq t_2 \rightarrow partial\ match\ \#(2) \\ similarity \geq t_2 \rightarrow match \end{array}$$

Where t_1 and t_2 are the similarity thresholds for determining mismatch and match, respectively. The advantage of this alignment approach is that it can assign a unique weight to each entity pair, which is directly related to the final matching degree of the two, and thus affects whether the alignment operation is performed. However, the problem of this approach is that the relationship between two entities is only determined by a similarity value that includes the influence of all possible attributes between them, and the weight of different attributes on the similarity cannot be measured, which affects the granularity and interpretability of entity alignment.

Recent research, on the other hand, has relied more on deep learning and neural networks for a unified representation of features of different modalities. Different features are populated as input data to deep neural networks so that they automatically learn the features of different modalities and abstract representations of different data, which in turn are converted into more abstract internal feature parameters of deep neural networks. For example, N Srivastava et al.[58] constructed a multimodal Deep Boltzmann Machine (DBM) to fuse the data features of picture modality and text modality, and integrated the features of both modalities through a layer of top-level neural network to form a data fusion analysis model.

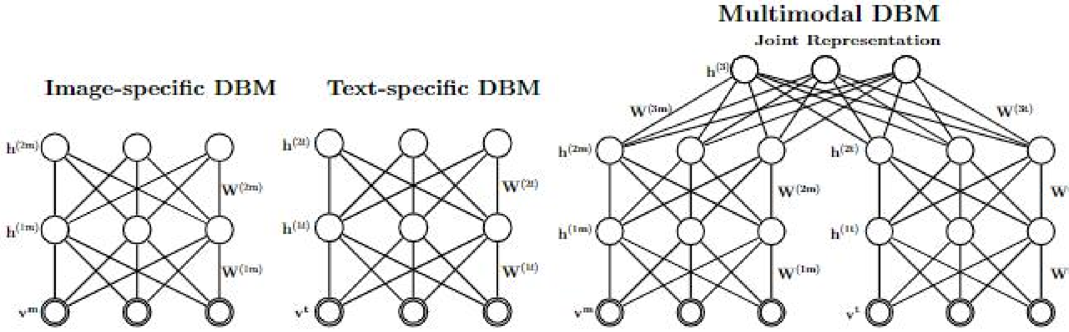


FIGURE 12 The structure of multimodal DBM [58]

In addition, Minjie Han [59] designed two corresponding deep neural network models for different modalities and fused the features extracted from the two models for the final action recognition and classification. By building a comprehensive neural network, Lei Zhao [60] processes the input data of different modalities separately and constructs a multilayer sub-neural network to correspond to them, and finally converts the heterogeneous modal features into the fused features of the same modality.

7 Future trends. Despite some progress in multimodal data fusion, existing multimodal data fusion still suffers from the following problems.

(1) Fewer multimodal datasets for large multidisciplinary domains

For most machine learning research, high-quality datasets are an indispensable part. There are already many high-quality datasets for single-modal data analysis, but the construction of multimodal public datasets, especially those involving specialized domains, is still in the development stage. The construction of more datasets will help researchers to analyze and evaluate the accuracy and performance of multimodal data fusion algorithms.

(2) Coarse granularity of multimodal data analysis

In the existing multimodal data analysis work, text data is mainly used as the main data, and the data of other modes is used as the auxiliary for enhancement learning, or the data of multiple modes are aligned and analyzed with entity granularity, and the hidden information contained in multimodal data is not fully explored. Mining semantic knowledge for multimodal data and fusion analysis with the granularity of knowledge structure will help to better explore the information implied in multimodal data and its association relationships.

(3) Large computational volume required by the model

The popularity of the Internet and mobile devices has led to a large amount of user-generated content, and more and more multimodal data are generated and rapidly disseminated on the Web, bringing challenges to multimodal data fusion analysis. Unsupervised learning-based multimodal data fusion is one way to address the growth of data volume. Another method is to reduce the complexity of the model and deploy it in mobile terminals to achieve "edge computing" for multimodal data fusion.

8 Conclusions. The study of multimodal data fusion based on unsupervised learning provides a fast and effective method to break the data isolation between different modalities and then unify the data of multiple modalities for analysis, which is more complex and promising for application than the traditional single-modal data analysis. In this paper, we start from unsupervised learning and its related ideas, based on single-modal feature learning, outline the multimodal data fusion techniques based on unsupervised learning in recent years, and discuss the problems and future research trends of multimodal data fusion based on unsupervised learning. We hope this paper will help readers understand the development of unsupervised learning and the current state of research on multimodal data fusion based on unsupervised learning, and inspire more work that is meaningful.

REFERENCES

[1] Xiao-Long, W. Y. Z. J., & CHENG, X. Q. (2013). Network Big Data: Present and Future. *Chinese Journal of Computers*, *6*.

[2] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific american*, *284*(5), 34-43.

[3] Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge acquisition*, *5*(2), 199-220.

[4] Blog, G. O. (2012). Introducing the knowledge graph: thing, not strings. *Introducing the Knowledge Graph: things, not strings*.

[5] Mitchell, T., et al. (2018). Never-ending learning. *Communications of the ACM*, *61*(5), 103-115.

[6] Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

[7] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798-1828.

[8] Lahat, D., Adali, T., & Jutten, C. (2015). Multimodal data fusion: an overview of methods, challenges, and prospects. *Proceedings of the IEEE*, *103*(9), 1449-1477.

[9] Hinton, G. E., & Sejnowski, T. J. (Eds.). (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

[10] Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural networks: Tricks of the trade* (pp. 599-619). Springer, Berlin, Heidelberg.

[11] Friedman, J. H. (2017). *The elements of statistical learning: Data mining, inference, and prediction*. springer open.

[12] Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100-108.

[13] Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461.

[14] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, *2*(11), 559-572.

[15] Golub, G. H., & Reinsch, C. (1971). Singular value decomposition and least squares solutions. In *Linear algebra* (pp. 134-151). Springer, Berlin, Heidelberg.

[16] Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

[17] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1985). *Learning internal representations by error propagation*. California Univ San Diego La Jolla Inst for Cognitive Science.

[18] Ballard, D. H. (1987, July). Modular learning in neural networks. In *AAAI* (Vol. 647, pp. 279-284).

[19] Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, *11*(12).

[20] Coates, A., Ng, A., & Lee, H. (2011, June). An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics* (pp. 215-223). JMLR Workshop and Conference Proceedings.

[21] Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, *72*(2011), 1-19.

[22] Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[23] Rifai, S., Mesnil, G., Vincent, P., Muller, X., Bengio, Y., Dauphin, Y., & Glorot, X. (2011, September). Higher order contractive auto-encoder. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 645-660). Springer, Berlin, Heidelberg.

[24] Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196). PMLR.

[25] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[26] Zhang, R., Isola, P., & Efros, A. A. (2016, October). Colorful image colorization. In *European conference on computer vision* (pp. 649-666). Springer, Cham.

[27] Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 132-149).

[28] Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision* (pp. 1422-1430).

[29] Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2536-2544).

[30] Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., ... & Brain, G. (2018, May). Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)* (pp. 1134-1141). IEEE.

[31] Zhu, W., Hu, J., Sun, G., Cao, X., & Qiao, Y. (2016). A key volume mining deep framework for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1991-1999).

[32] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L. (2016, October). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Springer, Cham.

[33] Misra, I., Zitnick, C. L., & Hebert, M. (2016, October). Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision* (pp. 527-544). Springer, Cham.

[34] Veličković, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., & Hjelm, R. D. (2018). Deep graph infomax. *arXiv preprint arXiv:1809.10341*.

[35] Velickovic, P., Fedus, W., Hamilton, W. L., Liò, P., Bengio, Y., & Hjelm, R. D. (2019). Deep Graph Infomax. *ICLR (Poster)*, *2*(3), 4.

[36] Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*(2), 945-959.

[37] Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using

multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164*(4), 1567-1587.

[38] Luhn, H. P. (1957). A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, *1*(4), 309-317.

[39] Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, *3*, 1137-1155.

[40] Le, Q., & Mikolov, T. (2014, June). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196). PMLR.

[41] Sarzynska-Wawer, J., Wawer, A., Pawlak, A., Szymanowska, J., Stefaniak, I., Jarkiewicz, M., & Okruszek, L. (2021). Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research*, *304*, 114135.

[42] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[43] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

[44] Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

[45] Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., & Xu, B. (2016). Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639*.

[46] Zhang, R., Isola, P., & Efros, A. A. (2017). Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1058-1067).

[47] Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*.

[48] Wang, K., Yin, Q., Wang, W., Wu, S., & Wang, L. (2016). A comprehensive survey on cross-modal retrieval. *arXiv preprint arXiv:1607.06215*.

[49] Sun, C., Gan, C., & Nevatia, R. (2015). Automatic concept discovery from parallel text and visual corpora. In *Proceedings of the IEEE international conference on computer vision* (pp. 2596-2604).

[50] Fang, Q., Xu, C., Sang, J., Hossain, M. S., & Ghoneim, A. (2016). Folksonomy-based visual ontology construction and its applications. *IEEE Transactions on Multimedia*, *18*(4), 702-713.

[51] Zhu, Y., Zhang, C., Ré, C., & Fei-Fei, L. (2015). Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*.

[52] Zeng, J. *Data Science: 7th International Conference of Pioneering Computer Scientists, Engineers and Educators, ICPCSEE 2021, Taiyuan, China, September 17–20, 2021, Proceedings, Part II*. Springer Nature.

[53] Ferrada, S., Bustos, B., & Hogan, A. (2017, October). IMGpedia: a linked dataset with content-based analysis of Wikimedia images. In *International Semantic Web Conference* (pp. 84-93). Springer, Cham.

[54] Liu, Y., Li, H., Garcia-Duran, A., Niepert, M., Onoro-Rubio, D., & Rosenblum, D. S. (2019, June). MMKG: multi-modal knowledge graphs. In *European Semantic Web Conference* (pp. 459-474). Springer, Cham.

[55] Xu, C., Tao, D., & Xu, C. (2013). A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*.

[56] Newcombe, H. B., Kennedy, J. M., Axford, S. J., & James, A. P. (1959). Automatic linkage of vital records. *Science*, *130*(3381), 954-959.

[57] Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*(328), 1183-1210.

[58] Srivastava, N., & Salakhutdinov, R. (2012, December). Multimodal Learning with Deep Boltzmann Machines. In *NIPS* (Vol. 1, p. 2).

[59] Min-jie, Han. (2017). Multi-modal Action Recognition Based on Deep Learning Framework. *Computer and Modernization*, (7), 48.

[60] Lei, Zhao. (2017). Feature Extraction and Selection from Multi-Modality Data Based on Deep Learning. *Tianjin University*.