Neural Semantic-Based Net for Image-Text CBIR

Mohamed Aboali [1], Hossam E. Abd El Munim[2] and Islam Elmaddah [3]

[1 2 3] Ain Shams University Cairo, Egypt

Abstract. In this paper, Content Based Image Retrieval, CBIR, methodology proposed. The methodology uses query inputs image and text modifiers. The proposed methodology maps the problem into known space, the textual space, in which feature composition performed. The composed features mapped to intended image feature that is used in the recall. Image feature extracted using the deep convolution neural network the ResNet, and textual features using recurrent neural network the LSTM networks. The proposed architecture, named semantic net, replaces jointly the flatten layers of the two networks Resnet, and LSTM. The semantic network contains three single-hidden-layer feedforward networks in cascade structure NetA, NetC, and NetB. NetA translate image features into textual feature. NetC, the compositional network composes the textual features produced by NetA with text-modifier feature to produce target textual captions features. The target textual caption features mapped to intended image features using NetB. These image features used to recall image from the image base based on cosine similarity metric. The proposed architecture tested using ResNet18, 50, and 152. The results of the study indicate promising results and comparative with most recent studies to our knowledge.

1 Introduction

Computer vision is one of the highly active leading research fields as it has numerous modern applications. Progress in computer vision energized by recent technological advances in storage, processing, networking, and data encoding. Computer vision includes acquisition, enhancement, restoration, analysis, recalling, encoding, and digital images descriptions. Computer vision mimics human transformations that occur when visual images (as an input to the human retina) that govern human behaviors during daily activities as well as in support of performing job duties. Computer vision builds intelligence based on images which includes not only the use of images as inputs but also as potential output. Consequently, nonconventional storing and recalling of images is one of the main foundations of computer vision progress [2].

Cheap handy digital image acquisition devices as well as advances in networks, processors production and storage capabilities are trends for decades. Those trends make it possible for digital images to be part of our daily life. Digital images use increases exponentially by time. The fact that images stores human's moments in a unique way that encapsulates unusual information. The historic quote "image is better than one thousand words" is a landmark. The Internet social media as well as bandwidth capabilities for transmitting and sharing images adds a new dimension to the use of images. Clouds storage and processing capabilities also adds another new dimension.

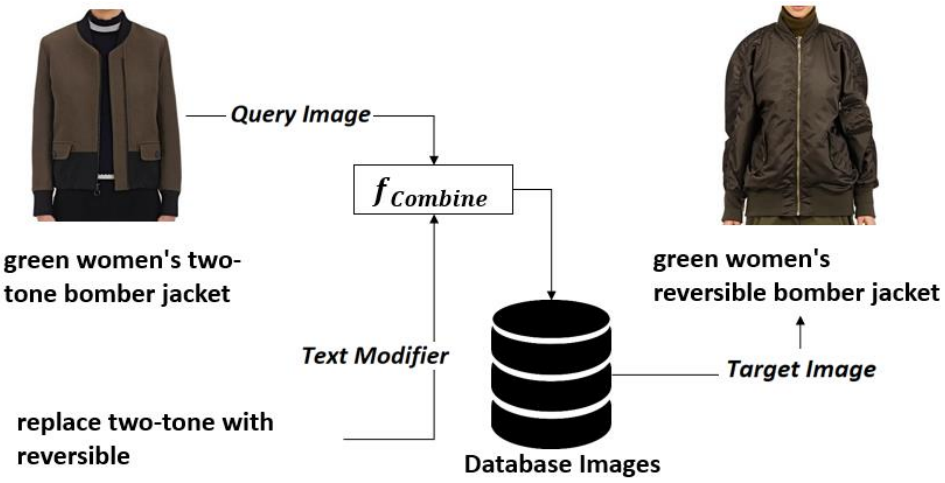

Fig. 1. Represents the image retrieval process in which the input is image and text combined

With this massive use of digital images dealing with images store-recall as datafile is inappropriate. One core problem in this is recalling intended images from large image-datasets. Efficient image-recall not only reduces effort in a painful search process but also adds more computer vision applications, makes image storage more efficient. Moreover, it adds a main foundation in intelligent-machines. Fig. 1 represents a typical image query formation using fashion 200K dataset as example.

Content Based Image Retrieval, CBIR, is a non-conventional digital image retrieval. Real-digital images data files are matrices of two-, or three-dimensional represents a scene. A scene is a set of objects in interaction. Computer vision applications interests could be in, an object, more than one object, subset of the interaction, or the entire interaction of a scene. This makes the numbers in the matrices as an absolute value useless for computer vision out of their local and/or global context. Effective computer vision applications mandate non-conventional image-recall relates to objects or image scene interactions it contains.

Comparing Images is not pixels by pixel-colors based process as these affected dramatically by many factors insignificant to image-objects or scene-interaction. These factors such as scale, rotation, translation, illumination, pose, orientation, displacements, and many other factors [4]. Consequently, images-similarities applied on features extracted from images and/or images regions. Features are high-level abstraction of image contents. Effective image query formulation is more complicated, in general, than data predicates as images or images abstraction is one of the potential inputs. Moreover, images text-modifiers are likely used. Clearly the recall query base, in effective CBIR, include use of image, visual descriptors, textual descriptors, and any combinations of these as potential inputs.

CBIR applications include fashion, graphic design, games, simulations, publishing, advertising, historical surveys, architectural engineering, crime prevention, medical diagnosis, geographical information, and remote sensing systems, etc. [5]. A typical image retrieval application example is in the clinical decision-making process; it is critical and important to find other images of the same modality, the same anatomic region of the same disease.

In this paper, a CBIR approach is proposed that uses images with textual modifiers. The proposed approach uses neural networks as feature extraction mean from both images and the text descriptors. The networks used are deep convolutional neural network for images and recurrent neural networks for text specifically, ResNet, and LSTM consequently. These networks proved to have significant ability to extract features with high discrimination abilities [6-8]. In the proposed model, the network architecture is semantically viable, uses known intermediate-feature spaces, and combine features of the same domain. The provided architecture tested against the well-known Fashion 200K standard dataset and compared with recent methodologies. The results indicated that the approach is comparable. Moreover, it is promising as its bases coincide with the problem solution logics.

2 Related work

Our work pursues image recall using image/picture modified by textual modifiers. The image scene builds a semantic concept which combined with modifier text semantic concept to generates conceptual semantics of the target image. The conceptual target-image represented as target image features suitable for fetching target image from the image dataset that store images with their image features. The fetching process requires metric function to carry out features comparison. The problem, in the former context, includes concepts/features extractions, concepts/features composition, distance metrics as well as mapping functions. Features extractions needed for both text and images then composition. The study focused on machine learning using neural networks. The study was built upon the conclusions drawn from our former research [1].

Concepts/features composition is a historically rooted research domain. The use of convolutional Neural Networks as a mean of feature representations of multiple semantic is used in [9]. Objects composition for recognition systems uses Deformable Part Models in [12], grammars in [13], and AND-OR graphs in [14]. Composition as pivotal element is used for visual question answering in [15], handwriting recognition in [16], and zero-shot detection in [17]. A lot of research has been done to improve features composition retrieval performance by user's feedback on relevance. A Multimodal Compact Bilinear Pooling as a feature fusion mechanism to combine image and text was proposed in [19]. In [20], the text feature is incorporated by mapping into parameters of a fully connected layer within the image CNN. Another domain that incorporates text images composition is visual question answering [21]. The residual connection

used to enforce composition in [22]. In [23] recurrent model to colorize images given text descriptions. In [24] adding text descriptors to localize objects within input images. Compositional Learning of an image and a text proposed in [25]. The attribute embedding operator was used in [18].

Another area related domain to our work is using neural networks in feature extraction. The use of neural feature extracted is incorporated in wide range of applications such as product search, face recognition or image geo-localization. Cross-modal image retrieval allows using other types of queries, examples include text to image retrieval, sketch to image retrieval or cross view image retrieval, and event detection. Feature extraction was not limited to image features but also text and sketches [25]. Neural networks used in feature extraction includes feed forward and recurrent networks both shallow and deep [26-28].

3 Content Based Image Retrieval

The term "Content-Based" implicitly requires content-comprehend that mandates analysis and a prior knowledge. That is contrary to using metadata such as keywords, and descriptions associated with the object/concept. "Content" of images implicitly refers to colors, shapes, textures, or even objects interactions inferred from the image. Field researchers refers to information extracted to represent these contents by features. Query formation, features extractions, combination of features and distance metrics are the main research topics of the CBIR [28-29] and are discussed briefly in the following subsections.

3.1 Query Formation

Query formation is driven by applications need. Applications specifies query inputs as well as the form of data or information available for the recall process. The CBIR queries are significantly difficult as it relates to abstract hypothetical concepts, images or thoughts formed in users' mind. Such query expression requires widening the means in which the formulation takes place. Therefore, researcher used varieties of types for query formation that include images, keywords, sketches, color maps, canvases, context maps, as well as icons [11]. The more common used are Query by keywords, Query by example, Query by canvas, Query by spatial icons. Combinations of the former mentioned formations was used in query formation [30-31]. In this research, the text features extracted from the text, rather than keywords associated with images combined with extracted image features to form combined images features to use in dataset image recall-queries.

3.2 Feature extraction

Is pivotal axil in the success of many computer vision applications. In case of engineered features, designers study the possible features. Moreover, likely, an

experimentation study over sample set is done to specify a set of features suitable for the application which called feature selection. Then, an algorithm designed for the extraction process of the selected features. This engineered approach assures that the selected features carry the desired properties which ensures the detection of insignificant distances for similar image concepts and vice versa [10]. Moreover, small variance for scale, rotation, noise, pose, and translation.

Feature learning or automatic representation is embedding with basic set of techniques. These techniques allow systems to automatically discover features from raw data. The engineering process herein is for feature learning rather than extracted itself. Feature learning motivated by the fact that machine learning, in general, mandates feature extraction step ahead. Feature learning can be either supervised or unsupervised. Supervised learning requires training and validation data to be labeled. From the supervised networks perceptron, radial bases, and Convolution neural networks. In unsupervised learning, features are learned without the need for labels. From the unsupervised architectures networks, Kohonen network, self-organizing map, and Hopfield network. Features could be extracted either in a local or global fashion. In global features extraction the feature operator applies to the whole image. Local features extraction is regional image search for any structured data, specific techniques like convolutional methods using hand-crafted kernels or syntactic and structural methods are used. These techniques encode problem specific knowledge into the features [10].

3.3 Features vectors combination

Is needed when different sources of features exist such as case of text and image and a new feature to be generated that carry both features. In [25] a Text Image Residual Gating, TIRG, function is used to combine image and text features. The TIRG function is weighted sum of Residual function and gating function. The weights supposed to be adapted through, potentially neural, learning process. The author's intent to "modify" the query image feature instead "feature fusion" to create a new feature from existing ones. Multilayer perceptron used to concatenate text and image features in [32-34]. LSTM, as recurrent model, fed by image features followed by text words used in [25]. Text used to form transfer matrices for image features in [18]. Visual question answering methodologies, that focus on finding answer to natural language question on a given image, were used in [20] [35-36].

3.4 Similarity Metrics

Content-based image retrieval (CBIR) not only needs efficient extraction of features and effective combination of features but also an effective similarity metric to measure the distance between features vectors. Various distance metrics are used to measure distance between vectors which include Euclidean distance, city block distance, Canberra distance, maximum value metric, Minkowski distance, Mahalanobis Distance, Histogram Intersection Distance, and Quadratic Form Distance [11]. Distance metrics are characterized by their accuracy and the time complexity.

4 Convolution and Recurrent Networks

4.1 Convolution Neural Networks

Convolution kernel filters applied to search for features throughout the image. Stride option provides allowance for convolution filters to skip pixels as features detected likely be redundant and at same time reduce computational complexity. Padding facility controls the size as the process goes on. Nonlinear functions enable removal of redundant concepts and a mean data domain control. Pooling functions remove duplicate or insignificant features as well as features translation control. The convolution and pooling layers finally make features map. Convolution layers are followed by traditional feed forward neural network layers for classification, recognition, or many other functions based on extracted feature map.

ResNet architecture is a convolution network characterized with skip or shortcut connection. This feature avoids the loss of information at later layers of the network because of the operators of the earlier layers. ResNet most famous architectures are RestNet18, ResNet34, ResNet50, ResNet101, and ResNet152 [6] [25].

4.2 Recurrent Neural Networks and LSTM

Judging an event highly depends on its context. Understanding word in the middle of a sentence requires comprehending the previous words and, likely, requires semantic memories from even previous sentences or paragraphs. Non-recurrent neural networks, traditional neural networks, cannot do that. The recurrent architecture as it goes through states according to the current state and the current input could be equipped with such feature. The architectures of recurrent neural networks include Hopfield, GRU, and LSTM. The Long-Short Term Memory, LSTM, network, as its structure, can carry concepts from long back in iterations as well as significant learning parameters [7][27].

5 Proposed Methodology

The problem under the study on the same dataset was explored recently in [25][3] and also in [1]. The proposed model, Fig. 2, uses ResNet and LSTM networks for image and text features extraction consequently. The semantic network replaces the fully connected layer of the two networks jointly with three stages of single hidden-layer feed-forward fully connected networks. These network maps the features to known intermediate spaces, textual, and image feature spaces. The composition of feature done in uniform space. The network architecture as well as the performed function coincide with the problem semantics [1].

To point out the problem that the design deals with, we should go through the standard dataset used in the study. The dataset is the fashion 200k which contains 200K+ dataset elements. The dataset splits into training subset of 172K training and 33k testing

elements chosen randomly. The dataset contains 200K images of different classes of fashions items. Each image of the dataset comes with brief text about the item. Our quarries created the same way of [25][3], modifier text contains a target word replaces a specific word in the query image captions that makes the target image caption. The training and testing datasets formed in this way that makes supervised training dataset. An element in the training and testing datasets contains element ID, query image 'Q_I ', Query caption 'Q_C ', modifier string 'Q_M', target image ID, target image'T_I' and target image caption 'T_C '. Fig. 1 depicts a typical example.

The query semantics, implicitly, assumes that the network will extract image caption or its features and replaces words to form target caption or associated features. Then, target features used to recall the intended target-image. One can also easily infer when humans see fashion model or even any-named-object our brain recalls its name or caption. Moreover, if we are subject to such a query, we will perform the text replacement form target captions or captions-features. Then, the target conceptual image or image-features formed in our brain based on former knowledge of object-images. Therefore, it is logically intuitive and semantically problem-fit to search for mapping that combines the two texts or their features in the text-feature space.

Therefore, the proposed architecture semantically-fit to the query-problems, Fig. 2. The semantic net architecture composed of three single-hidden nonlinear feedforward fully connected subnetworks NetA, NetB, and NetC. Two mapping networks, NetA, NetB and a compositional network NetC. The first two networks, we call them the translators' networks, mapping functions from image feature space to textual feature space and vise versal consequently.

NetA function maps the φ_{Qx} is the query image extracted by *ResNet to* φ_{Qt} which is an approximate image caption feature. NetC, the compositional network, inputs are the two text features, φ_{QMt} query modifier text and φ_{Qt} both represent textual features likely make composition effectively and smoothly operational. NetC outcome, φ_{Tt}, is an approximate target caption feature. The final stage, NetB, input is φ_{Tt} and it maps it possible target image features.

For more clarity and overall functional specifications of the proposed solution. Let us assume: - η_L, η_R are the LSTM network, and RESNET transfere functions to the inputs flatten vector of input of the FC layer of that networks, and τ_A , τ_B , τ_C are the networks NetA, NetB, and NetC transfer functions consequently then the semantic net overall function could be summarized as: -

$\varphi_t = \eta_L(t)$, for caption query, target and modifier text

$\varphi_x = \eta_R(I)$, for query and target image

$\overline{\varphi_{Tx}} = \tau_B\ (\tau_C(\tau_A\ (\varphi_{Qx}), \varphi_{QMt}))$

where: -

$$\tau_A\ (\varphi_{Qx}) = \overline{\varphi_{Qt}} = Tanh(\varphi_{Qx} * W_A^1) * W_A^2$$
$$\tau_B\ (\overline{\varphi_{Tt}}) = \overline{\varphi_{Tx}} = Tanh([\overline{\varphi_{Tt}}] * W_B^1) * W_B^2$$
$$\tau_C\ (\overline{\varphi_{Qt}}, \varphi_{QMt}) = \overline{\varphi_{Tt}} = \sigma([\overline{\varphi_{Qt}}: \varphi_{QMt}] * W_C^1) * W_C^2$$

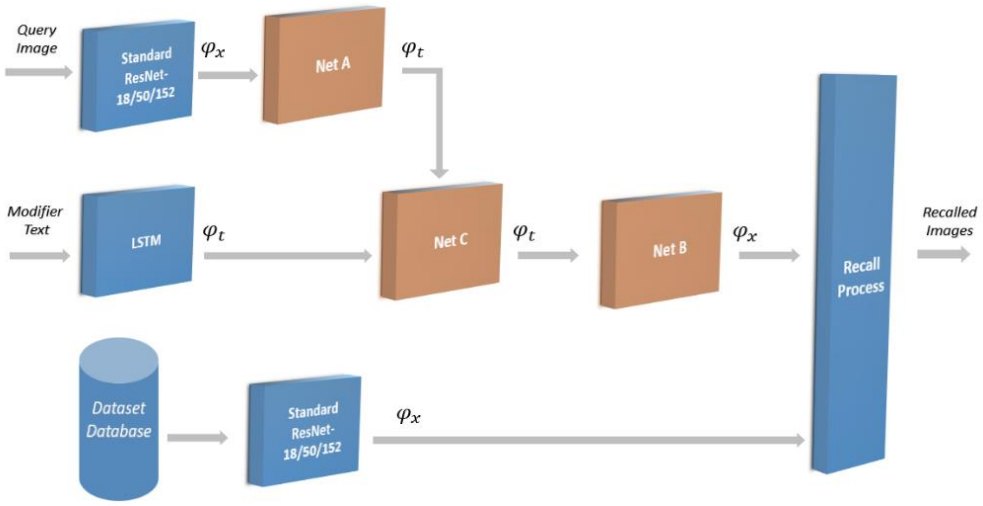

Fig. 2. Proposed semantic network

6 Experimental Results

The proposed three models trained using the 200K fashion dataset training dataset. The training dataset is the 172K dataset drawn from the 200K fashion dataset. The networks trained in two phases parallel phase and tuning phase. In the first phase, the parallel phase, the three networks are trained in parallel and on perfect inputs on the training dataset.

Table 1. Semantic net performance on the dataset

Semantic Net	Training Dataset			Test Dataset		
	Net A	Net B	Net C	Net A	Net B	Net C
ResNet 18	0.0201	0.1454	0.0157	0.052	0.1319	0.0443
ResNet 50	0.0159	0.1313	0.0128	0.022	0.1168	0.0214
ResNet 152	0.0150	0.1353	0.0124	0.027	0.1203	0.0248

In our experimentations, NetB ResNet 18, 50, and 152 models were used. The network feature vectors produced by the networks are of sizes 512,2048, and 2048 consequently. The later numbers are NetA input vector and NetB output vector per semantic network.

The hidden layers of the three network A, B and C fixed for the three semantic models as 1000, 2500, and 1800 in sequence. NetC, the compositional network, trained until MSE 0.003 and used for the three models. NetA trained per semantic model on perfect training dataset. NetB inputs affected by the errors generated by the former two networks in cascade effect. To make the point clear, let us assume, $P_{AE}, P_{BE}, and\ P_{CE}$ are the network error probability due to its imperfect mapping. Then, the expected error probability at the output of the network in the semantic formation $\overline{P_{AE}}$, $\overline{P_{BE}}, and\ \overline{P_{CE}}$ will be as follows: -

$$\overline{P_{AE}} = P_{AE}$$
$$\overline{P_{CE}} = P_{AE} + P_{CE} - P_{AE} * P_{CE}$$
$$\overline{P_{BE}} = \overline{P_{CE}} + P_{BE} - \overline{P_{CE}} * P_{BE}$$

The error propagation to NetB, mandates two stages of training. The networks trained until mean error below 0.12. NetB performance in semantic performance after this stage of training in semantic formation found to be 0.27, 0.22 0.21. In the second phase, which we call it tunning phase, the network inputs replaced by NETC-output. The training continues when it turns below 0.16 validation performance evaluation every 10 iterations using random 1000 testing dataset elements. The result of the training performance on the training dataset summarized in Table 1.

Table 2. semantic net results in comparison with [25] [3].

Method	R@1	R@10	R@50
TIRG	14.1±0.6	42.5±0.7	63.8±0.8
TRIG with BERT and complete text query	19.9±0.6	51.7±1.5	71.8±1.3
Compose AE	22.8±0.8	55.3±0.6	73.4±1.5
Semantic Net 18	19.86	46.12	67.54
Semantic Net 50	23.44	48.68	72.275
Semantic Net 152	25.4	52.85	74.51

Comparing the performance of these networks with recent studies on the same dataset conducted on both of [25] and [3]. Table 2 shows the results in comparison with the results presented in [25] [3].

During the final tunning of the NetB of the three semantic nets training behavior were different. ResNet18, the network learning was oscillating around the reported results without any significant progress in error reduction or improvement in the recall capabilities. ResNet152 and ResNet50 around the reported results, the network was slowly progressive learning. However, the ResNet50 was lower in error, the recall ability of ResNet152 was above ResNet50. The improved recall performance of the ResNet152, even at higher mean error, apparently come from the discriminative ability and the features extracted quality of the network. ResNet18 performance was between the

results reported in [25] and [3]. This architecture was unable to cross over [3]. That likely due to the insufficient number of features extracted which are 512 features.

Fig. 3. Typical recall example

Samples of the intermediate performance, ResNet50-semantic-network, recalls are in Fig. 3. The Figure contains three typical recall queries. The query image is on left with caption above. The arrow labeled with the modifier string. The arrow points to first four recalled images the boxed ones are targets. Query A first recalled image was hit and the rest carries significant intended features. Query B was a miss case however one can easily see that the multicolor property is there. Query C 2nd and 3rd were hit the rest was not far from query-targets.

7 Conclusion

In this paper, a proposed text and image composition model for CBIR was introduced. The model based on semantic network. The semantic network replaces the fully connected layers of the two well-known deep networks, LSTM and ResNet. The LSTM basic roll is text features extraction and ResNet extracts the image features. The semantic network uses semantically viable architectures to compose features. The semantic net composed of three non-linear feedforward single-hidden layer networks called NetA, NetB, and NetC. The NetA base role extract from query image the image text-

features. NetC compose the Modifier text features with query image text feature and map it to target captions features. NetB maps target-image captions features to target-image features. The formed target image features used to recall images using cosine similarity metric. The proposed semantic network components could be trained in parallel and later tuned to compensate for cascade errors. The proposed model tested on the well-known Fashion 200K dataset. During the study of the proposed architectures, three different types of ResNet were used 18, 50, and 152. The network performance found to be comparable to recent studies [25][3].

References

1. M. Aboali, I. Elmaddah and H. E. -D. Hassan, "Augmented TIRG for CBIR Using Combined Text and Image Features," 2021 International Conference on Electrical, Computer and Energy Technologies (ICECET),2021, pp. 1-6, doi: 10.1109/ICECET52533.2021.9698617.
2. Davies, E. Roy., "Computer vision: principles, algorithms, applications, learning fifth edition", Academic Press, 2018.
3. Anwaar, Muhammad & Labintcev, Egor & Kleinsteuber, Martin, Compositional Learning of Image-Text Query for Image Retrieval, IEEE Winter Conference on Applications of Computer Vision (WACV), 2021, 1139-1148.
4. Kavita. R. Singh , Mukesh. A. Zaveri, Mukesh. M. Raghuwanshi ," Illumination and Pose Invariant Face Recognition: A Technical Review", International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM), ISSN: 2150-7988 Vol.2 (2010), pp.029-038.
5. Jing Li, "The application of CBIR-based system for the product in electronic retailing," 2010 IEEE 11th International Conference on Computer-Aided Industrial Design & Conceptual Design 1, Yiwu, China, 2010, pp. 1327-1330.
6. Evaluating the Performance of ResNet Model Based on Image Recognition, Riaz Khan-Xiaosong Zhang,2018.
7. A review on the long short-term memory model Greg Van Houdt1 · Carlos Mosquera2 · Gonzalo Nápoles1, 2020.
8. Paul Ferré, Franck Mamalet, and Simon J. Thorpe. 2018. Unsupervised Feature Learning With Winner-Takes-All Based STDP. Frontiers in Computational Neuroscience 12 (2018).
9. Ishan Misra, Abhinav Gupta, and Martial Hebert," From Red Wine to Red Tomato: Composition with Context ", IEEE Conference on Computer vision and pattern recognition (CVPR), Honolulu, Hi, 2017, pp. 116—1169.
10. A review on image feature extraction and representation techniques, TianDongping Tian,2013.
11. Tatiana Jaworska, Query techniques for CBIR, Systems Research Institute, Polish Academy of Sciences, 6 Newelska Street, Warsaw, Poland, 2016.
12. P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.
13. Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. International Journal of computer vision, 63(2):113–140, 2005.
14. Z. Si and S.-C. Zhu. Learning and-or templates for object recognition and detection. IEEE transactions on pattern analysis and machine intelligence, 35(9):2189–2205, 2013.

15. J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Learning to compose neural networks for question answering. In NAACL, 2016.
16. B. M. Lake, R. Salakhutdinov, J. Gross, and J. B. Tenenbaum. One shot learning of simple visual concepts. In Proceedings of the 33rd Annual Conference of the Cognitive Science Society, volume 172, page 2, 2011.
17. V. Krishnan and D. Ramanan. Tinkering under the hood: Interactive zero-shot learning with net surgery. arXiv preprint arXiv:1612.04901, 2016.
18. T. Nagarajan and K. Grauman. Attributes as operators. 2018.
19. A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847, 2016.
20. H. Noh, P. Hongsuck Seo, and B. Han. Image question answering using convolutional neural network with dynamic parameter prediction. In CVPR, 2016.
21. D. Teney, P. Anderson, X. He, and A. van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In CVPR, 2018.
22. A. Miech, I. Laptev, and J. Sivic. Learnable pooling with context gating for video classification. arXiv preprint arXiv:1706.06905, 2017.
23. J. Chen, Y. Shen, J. Gao, J. Liu, and X. Liu. Language-based image editing with recurrent attentive models. In CVPR, 2018.
24. X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In ICCV, 2017.
25. Nam Vo, et. al., "composing text and image for image retrieval – an empirical odyssey", IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019.
26. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In CVPR, 2015.
27. a tutorial into Long Short-Term Memory Recurrent Neural Networks Ralf C. Staudemeyer, Schmalkalden,2019.
28. Analysis of distance metrics in content-based image retrieval using statistical quantized histogram texture features in the DCT domain Author links open overlay panel,Fazal Malik, Baharum Baharudin.
29. A Decade Survey of Content Based Image Retrieval using Deep Learning, Shiv Ram Dubey, Member, IEEER.
30. Lim, J.-H. & Jin, J. S., 2005. A structured learning framework for content-based image indexing and visual query. Multimedia Systems, Volume 10, p. 317–331.
31. Rasiwasia, N., Moreno, P. J. & Vasconcelos, N., 2007. Bridging the Gap: Query by Semantic Example. IEEE TRANSACTIONS ON MULTIMEDIA, Aug, 9(5), pp. 923- 938.
32. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual Question Answering. In ICCV, 2015.
33. X. Guo, H. Wu, Y. Cheng, S. Rennie, and R. S. Feris. Dialog-based interactive image retrieval. arXiv preprint arXiv:1805.00145, 2018.
34. H. Nam, J.-W. Ha, and J. Kim. Dual attention networks for multimodal reasoning and matching. In CVPR, 2017.
35. A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. A simple neural network module for relational reasoning. In NIPS, 2017.
36. J.-H. Kim, S.-W. Lee, D. Kwak, M.-O. Heo, J. Kim, J.-W. Ha, and B.-T. Zhang. Multimodal residual learning for visual qa. In NIPS, 2016