DiVA
http://www.diva-portal.org

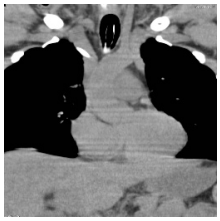
This is the published version of a paper presented at *CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, Sept 9-12, 2019.*

Citation for the original published paper:

Vu, M., Sznitman, R., Nyholm, T., Löfstedt, T. (2019)
Ensemble of Streamlined Bilinear Visual Question Answering Models for the ImageCLEF 2019 Challenge in the Medical Domain
In: Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller (ed.), *CLEF 2019: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*

N.B. When citing this work, cite the original published paper.

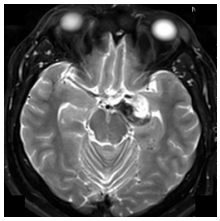
Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:umu:diva-166758



Q:"Is this a contrast or noncontrast CT?"
A: "Noncontrast"

Q:"Is this an MRI image?"
A: "Yes"

Q:"Is this image modality T1, T2, or FLAIR?"
A: "T2"



Q:"What is most alarming about this MRI?"
A: "Dermoid cyst"

Fig. 1: Examples of questions and images and their corresponding answers in the ImageCLEF-VQA-Med 2019 challenge.

that combines the image and question features, and (4) a classifier that uses the combined features to select the most likely answer.

ImageCLEF [8] aims to support the need of the global community of reusable resources for benchmarking the cross-language annotation and retrieval of images. In 2019, ImageCLEF had four main tasks: lifelogging, medicine, nature, and security. With the purpose of providing a "second opinion" for clinicians on complex medical images and offering patients an economical way to monitor their disease status, ImageCLEF organizes a medical domain VQA challenge, called ImageCLEF-VQA-Med [1] (see examples in Figure 1).

In the present work, we describe the model that we developed for the ImageCLEF-VQA-Med 2019 challenge. First, we present a novel fusion scheme for questions and images. Second, we introduce an image preprocessing step that suppresses unwanted distortions to enhance the quality of the ImageCLEF-VQA-Med images before they are fed into a Convolutional Neural Network (CNN) for image feature extraction. Third, we propose to utilize a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model [5] to extract the question features. Last, we present an ensemble of VQA models that gave a large boost in the evaluation metrics on both validation and test sets.

2 Related Work

Since most existing VQA methods use standard embedding models for text [12] and standard CNNs to extract image features [6], the research focus has largely been on fusion strategies that combine information from both input sources [6,10,3]. Recently, attention schemes have also been introduced in VQA models in order to focus the trained models towards question-guided evaluations. The review paper of Kafle *et al.* [9] offers a comprehensive overview of recent VQA models.

An image model is used to extract visual features from the input images. Most recent VQA models use CNNs, often ones that are pre-trained on *e.g.* the ImageNet dataset [13]. Popular choices for the image model includes: VGGNet [15],

GoogLeNet [16], and ResNet [6,10,3]. Multimodal Compact Bilinear (MCB) [6], Multimodal Low-rank Bilinear (MLB) [10], and Multimodal Tucker Fusion for Visual Question Answering (MUTAN) [3] are current VQA methods that employ bilinear transformation to encode image and question. As with these, we used a ResNet-152 model, that was pre-trained on the ImageNet dataset, to extract visual features.

Common models employed to extract question features include Long Short-term Memory (LSTM) [7], Gated Recurrent Units (GRU) [4], and Skip-thought vectors [12]. Skip-thought vectors is a powerful unsupervised encoder-decoder approach that has been used in many recent VQA models [6,10,3]. In the present work, we not only used Skip-thought vectors but also evaluated the use of a pre-trained BERT model [5] to extract question features. The BERT model has obtained state-of-the-art results on a wide variety of NLP tasks recently.

Attention mechanisms have led to breakthroughs in many NLP applications, for example, in neural machine translation [2], and in computer vision, such as in image classification [17]. Propelled by the remarkable success accomplished by attention mechanisms in computer vision and NLP, numerous VQA models have employed attention schemes to improve predictions.

3 Proposed Approach

For the task of VQA [9], we are interested in predicting the most likely answer, $\hat{a}$, given a question, q, about an image, v. The problem can be stated as

$$\hat{a} = \arg\max_{a \in \mathcal{A}} P(a \mid q, v, \Theta), \tag{1}$$

where $\mathcal{A}$ is the set of possible answers and Θ denotes all model parameters.

Figure 2 illustrates the proposed method. It uses pre-trained networks to extracts image and question features (in red and green, respectively), and feed them to a fusion scheme. These features are combined using an attention mechanism [6] (orange) to compute global image features, $\tilde{v}$. We proposed an efficient bilinear transformation that takes two inputs: global image features and global question features, $\tilde{q}$, and yields a single latent feature vector, $\tilde{f}$, that is then linearly mapped to the answer vector (white) to generate the output. The proposed bilinear fusion scheme is further described in the following section.

3.1 Proposed Method

To encode questions and images, we first make use of a multi-glimpse attention mechanism [6] to compute global image features, $\tilde{v} = [\omega_1^T, \ldots, \omega_G^T]^T \in \mathbb{R}^{KG}$, where K denotes the dimensions of the identity core tensor, that is decomposed using Tucker Decomposition in the attention scheme (see [3] for more details), and G is the number of glimpses.

https://github.com/facebook/fb.resnet.torch

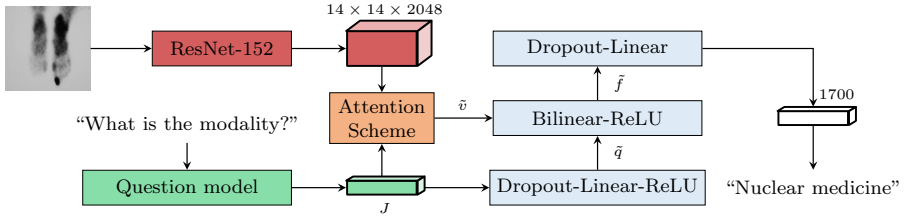


Fig. 2: Proposed method. We used a ResNet-152 model, that was pre-trained on the ImageNet dataset, to extract image features. Skip-thought vectors or a pre-trained BERT is employed to extract question features. These features are passed through an attention mechanism to produce global image features, $\tilde{v}$, while the question features are linearly transformed to obtain global question features, $\tilde{q}$. We then apply the proposed bilinear transformation on these global features to compute output features, $\tilde{f}$, before calculating the output probability vector over the possible answers.

The global question features can be written as

$$\tilde{q} = \text{ReLU}(\mathbf{W}^q q + b^q), \tag{2}$$

where $\tilde{q} \in \mathbb{R}^{KG}$, $q \in \mathbb{R}^J$ are the question features, and $\mathbf{W}^q \in \mathbb{R}^{KG \times J}$ and $b^q \in \mathbb{R}^{KG}$ denote the weight and bias terms, respectively. ReLU is the rectified linear unit activation function.

Given these, the output features of the proposed model are encoded as

$$\tilde{f}_i = \text{ReLU}\left(\sum_{j=1}^{KG}\sum_{k=1}^{KG} \tilde{q}_j w_{ijk}^f \tilde{v}_k + b_i^f\right) = \text{ReLU}\left(\tilde{q}^T \mathbf{W}_i^f \tilde{v} + b_i^f\right), \tag{3}$$

where $\tilde{f} \in \mathbb{R}^K$, $\mathbf{W}_i^f \in \mathbb{R}^{KG \times KG}$ and $b_i^f \in \mathbb{R}$ denote the weight and bias terms in the *bilinear* scheme, respectively.

The probabilities of each target answer over all possible target answers are then written as

$$f = \text{SoftMax}(\mathbf{W}^a \tilde{f} + b^a), \tag{4}$$

where $f \in \mathbb{R}^N$, and $\mathbf{W}^a \in \mathbb{R}^{N \times K}$ and $b^a \in \mathbb{R}^N$ denote the weight and bias terms, respectively.

3.2 Implementation Details and Training

The proposed method, illustrated in Figure 2, contains three different components: an image model (see Section 4.1), a question model (see Section 4.2), and the proposed fusion with an attention mechanism model. The implementation and training details of the latter one is discussed below.

To implement the attention mechanism, we followed the description in [3]. We used the Adam optimizer [11] with a learning rate of 0.0001, a batch size of 128 and used a dropout rate of 0.5 for all linear and bilinear layers. We trained the proposed model for 100 epochs on an Nvidia GTX 1080 Ti GPU, the training time for the whole network with the attention scheme was around 1.5 hours.

4 Ensemble of Multiple Models

We employed ensemble learning to build a *committee* from a collection of trained VQA models, each casts a weighted vote for the predicted answer, in order to use the wisdom of the crowd to produce better predictions.

4.1 Image Model

We preprocessed and augmented the images before passing them through the pre-trained ResNet-152 model to extract image features.

To remove unwanted outer areas (text and/or background) from an image, we applied the following sequence of image processing techniques:

(1) Normalize the intensities of the input image to 0-255.
(2) Apply Otsu's method to binarize the normalized image using a threshold of 5.
(3) Apply an open operation on the thresholded image with a rectangular structuring element of size 40×40.
(4) Fill the holes of the binary image.
(5) Remove all connected components, except the two largest ones.
(6) Compute a bounding-box of the foreground.
(7) Crop the image to the bounding box.
(8) Apply an open operation with a rectangular structuring element of size 50×50.
(9) Crop the normalized image to the enlarged bounding box.
(10) Multiply the results from steps (8) and (9) to obtain a cropped image.
(11) Resize the cropped image to 448×448.
(12) Z-normalize the resized image.

Data augmentation was applied on the pre-processed dataset before the images were sent to the network to improve the generalization. We used two types of data augmentation: (i) rotate the image by a randomly selected number of degrees from the range $[-20, 20]$, and (ii) randomly scale the image size using a scaling factor in the range $[0.9, 1.1]$.

4.2 Question Model

We evaluated the use of Skip-thought vectors and a pre-trained BERT model for extracting the question features. These features were then used in the VQA models (see Table 2).

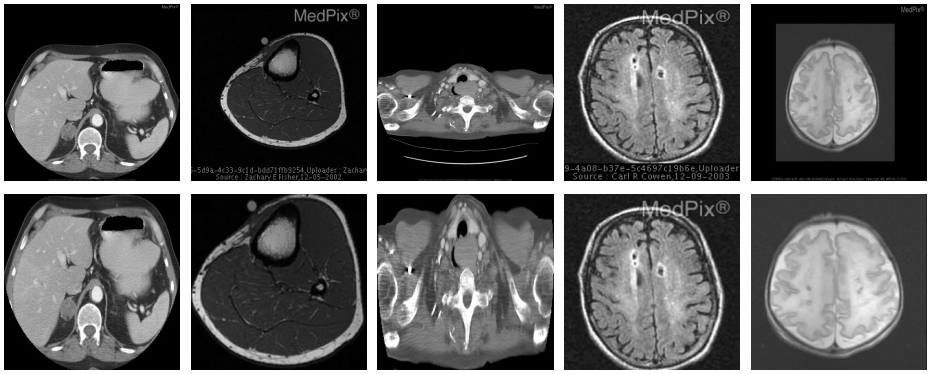


Fig. 3: Example images passed through the pre-processing pipeline.

We used the same preprocessing techniques for the questions as was used in [3,6]. These were: (i) removing the punctuation marks, and (ii) converting to lower-case.

To overcome the challenge of seeing new words in the medical domain, we employed a Word2Vec model trained on the Google News dataset that includes three million words vectors. We then used a linear regression model without regularization to map the Word2Vec to the Skip-thought embedding space [12]. This enabled our Skip-thought vectors to generate 2,400-dimensional word features.

BERT is a deep bidirectional transformer encoder that has obtained new state-of-the-art results on multiple NLP tasks [5]. The BERT model was pre-trained for general-purpose *language understanding* on a large text corpus, called WikiText-103, on two unique tasks: Masked Language Model (MLM) and Next Sentence Prediction (NSP) [5]. We employed two pre-trained BERT models: (i) *bert-base-multilingual-uncased*, and (ii) *bert-base-multilingual-cased*, to extract question features. Of each pre-trained model, we used a feature-based approach by generating ELMo-like [14] pre-trained contextual representations using two methods: (i) Second-to-Last Hidden (768-dimensional), and (ii) Concat Last Four Hidden (3,072-dimensional) (see Table 7 in [5] for details).

4.3 Fusion Model with Attention Mechanism

In addition to the proposed model, the ensemble contained MLB and MUTAN models, for which we used freely available PyTorch code.

We integrated ten different MLB models [10] in the ensemble (see Table 2). To prevent overfitting, we reduced the dimensions of the identity core tensor, K,

https://github.com/Cadene/skip-thought.torch/tree/master/pytorch
https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/
https://github.com/huggingface/pytorch-pretrained-BERT
https://github.com/Cadene/vqa.pytorch

to 64, 100 and 200 (the original value was $K = 1{,}200$, see [10]). Furthermore, we replaced all hyperbolic tangent (tanh) activations by ReLU activation functions.

We employed five different versions of the MUTAN architecture [3] in the ensemble model (see Table 2). All hyper-parameters were set as in [3]. As with the MLB model, all hyperbolic tangent activations were replaced by ReLU activation functions.

Both MLB and MUTAN were trained to minimize the categorical cross entropy loss using the Adam optimizer with a learning rate of 0.0001 and exponential decay rates of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. As in the proposed model, the batch size was 128 and the model was trained for 100 epochs. As for the proposed method, the training time for both the MLB and the MUTAN models on an Nvidia GTX 1080 Ti GPU was about 1.5 hours.

4.4 Ensemble Model

By varying the pre-trained question models and a few hyper-parameters of the fusion schemes, we trained more than 40 base models separately on the training set. We then evaluated their performance on the validation set to select the top 26 performing models (see Table 2), and built ensemble models using those 26 models.

To generate the outputs for the test set, we trained the 26 aforementioned models on the concatenation of the training and validation sets with the aim of making the networks learn a wider range of answers.

We then used two ensemble techniques: the average,

$$\tilde{a} = \frac{1}{M} \sum_{m=1}^{M} f_m, \tag{5}$$

and the weighted average,

$$\tilde{a}_{\text{weighted}} = \frac{\sum_{m=1}^{M} w_m f_m}{\sum_{m=1}^{M} w_m}, \tag{6}$$

where $\tilde{a}, \tilde{a}_{\text{weighted}} \in \mathbb{R}^N$ are the output probability vectors over the answers, $f_m \in \mathbb{R}^N$ is the answer vector corresponding to model m that was computed by Equation 4. The M is the number of models, and $w_m \in \mathbb{R}$ is the weight corresponding to the performance of the mth model (computed as the mean accuracy over the last 21 epochs on the validation set, as seen in Table 2).

5 Experiments

In this section, we detail the ImageCLEF-VQA-Med dataset, and compare the proposed method to MLB [10] and MUTAN [3] when applied on the validation set. In addition, we discuss the results of the ensemble model on the test set.

Table 1: Mean accuracy (and standard errors) computed from the last 21 epochs on the validation set for MUTAN, MLB (with default hyper-parameters), and the proposed model. K and G are the dimensions of the identity core tensor and the number of glimpses, respectively (see Section 3.1 for details). Note that we used Skip-thought vectors for all models.

Fusion	**Question**	**Activation**	K	G	**Mean**	**SE**
MUTAN [3]	skip-thought	tanh	*n.a.*	2	58.35	0.18
MLB [10]	skip-thought	tanh	100	8	58.96	0.11
	skip-thought	tanh	200	4	58.23	0.16
	skip-thought	tanh	1200	4	58.74	0.12
Proposed	skip-thought	ReLU	64	8	**60.12**	0.17

5.1 Material

The ImageCLEF-VQA-Med data were partitioned in three sets: (i) a training set of 3,200 images with 12,792 Question & Answer (QA) pairs, (ii) a validation set of 500 images with 2,000 QA pairs, and (iii) a test set of 500 images with 500 questions. Different from previous challenges, the organizers of the ImageCLEF-VQA-Med 2019 categorized the questions in four groups: Modality, Plane, Organ System, and Abnormality. The task was to answer the questions about the medical images in the test set as correctly as possible.

The evaluation metrics were: (i) strict accuracy, defined as the percentage of correctly classified predictions, and (ii) Bilingual Evaluation Understudy (BLEU) score that computes the similarity between n-grams of the ground truth answers and the corresponding predictions.

5.2 Results and Discussion

Table 1 compares the performance of the proposed method to MLB [10] and MUTAN [3], while Table 2 shows the mean and standard error of the accuracies of the last 21 epochs of the 26 best-performing methods on the validation set. From Table 1 and Table 2 we see that: (1) the proposed method performs better than state-of-the-art methods on the ImageCLEF-VQA-Med dataset, (2) *bert-base-multilingual-uncased* gives better question representations than *bert-base-multilingual-cased* does, and (3) the question features extracted by the pre-trained BERT models are as good as those produced by the Skip-thought vectors.

There are two possible explanations to why the proposed model outperforms MLB and MUTAN. First, the ReLU overcomes the vanishing gradient problem that hyperbolic tangent activation functions suffers from. It thus allows the proposed model to learn faster and therefore it may perform better. Second, by using the bilinear transformation instead of an inner product operation to produce the global question and image features, that are used in MLB and MUTAN, the proposed method considers every possible combination of elements

Table 2: Mean accuracy (and standard errors) computed from the last 21 epochs for the 26 best-performing models on the validation set. The first column indicates the fusion scheme with attention mechanism that was used. The second column contains the question models that was used to extract the question features. Here, "bert-uncased" and "bert-cased" are *bert-base-multilingual-uncased* and *bert-base-multilingual-cased*, respectively (see Section 4.2). J denotes the dimension of the question feature space, while K and G are the dimensions of the identity core tensor and the number of glimpses, respectively (see Section 3.1). Note that we used ReLU activation functions for all models.

Fusion	**Question**	J	K	G	**Mean**	**SE**
MUTAN [3]	bert-cased	768	*n.a.*	2	57.75	0.18
	bert-uncased	768	*n.a.*	2	58.35	0.18
	bert-cased	3072	*n.a.*	2	58.42	0.17
	bert-uncased	3072	*n.a.*	2	58.88	0.21
	skip-thought	2400	*n.a.*	2	59.64	0.20
MLB [10]	bert-cased	768	200	4	57.57	0.15
	bert-cased	3072	200	4	58.45	0.11
	bert-cased	768	100	8	58.56	0.15
	bert-cased	3072	100	8	58.73	0.12
	bert-uncased	3072	100	8	58.74	0.19
	bert-uncased	3072	200	4	59.15	0.16
	bert-uncased	768	200	4	59.45	0.19
	skip-thought	2400	200	4	59.90	0.11
	skip-thought	2400	100	8	60.02	0.10
	bert-uncased	768	100	8	60.09	0.22
Proposed	bert-uncased	3072	200	4	58.83	0.15
	bert-cased	3072	200	4	58.97	0.12
	bert-uncased	3072	100	8	59.10	0.21
	bert-cased	768	200	4	59.12	0.15
	bert-cased	3072	100	8	59.33	0.18
	skip-thought	2400	200	4	59.62	0.13
	bert-cased	768	100	8	59.63	0.24
	bert-uncased	768	200	4	59.72	0.17
	skip-thought	2400	100	8	59.85	0.13
	bert-uncased	768	100	8	60.09	0.22
	skip-thought	2400	64	8	**60.12**	0.17

from two aforementioned features, and thus become more capable of learning a larger range of answers.

Table 3 presents the accuracy and BLEU scores of the ensemble models on the validation and test sets. We selected the top 6 performing ensemble models on the validation set and used those to make predictions to submit to the evaluation server. As can be seen in Table 3, the ensemble of 11 proposed models resulted

Table 3: Results of the ensemble models on the validation and test sets in the ImageCLEF-VQA-Med 2019. # denotes the number of base models used in the ensemble, while "Run" is the submission ID on the leaderboard. "skip-thought", "bert-768", and "bert-3072" denote the ensembles of base models, that use different types of question models: Skip-thought vectors, 768-dimensional BERT and 3072-dimensional BERT, respectively. Our results won the 3rd place at ImageCLEF-VQA-Med 2019 without using any additional training data.

Ensemble	**Description**	**#**	**Validation**	**Test**	**Run**
n.a.	single best	1	60.50 (62.62)	60.60 (62.30)	26843
Weighted	bert-3072	10	60.85 (63.21)		
	all models	26	61.15 (63.48)		
	skip-thought	6	61.25 (63.39)		
	bert-768	10	61.40 (63.69)	61.20 (63.10)	26880
	proposed	11	61.55 (63.87)	61.20 (63.20)	27196
Average	bert-3072	10	61.30 (63.86)		
	bert-768	10	61.40 (63.73)		
	skip-thought	6	61.55 (63.80)	61.40 (63.30)	27197
	all models	26	61.35 (63.61)	61.40 (63.30)	26863
	proposed	11	**61.60 (63.89)**	**61.60 (63.40)**	27195

in a 1 % improvement on strict accuracy, which is consistent with the literature results of using ensembles.

Our best performing model, that achieved the strict accuracy of 61.60 and the BLEU score of 63.89 on the validation set, was the ensemble of 11 proposed models (see Table 2). This ensemble model also performed the best on the test set (61.60 accuracy and 63.89 BLEU score), and won 3rd place in the ImageCLEF-VQA-Med 2019 challenge without using additional training data.

6 Conclusion

We have presented a novel fusion scheme for the VQA task. The proposed approach was shown to perform better than current methods in the ImageCLEF-VQA-Med 2019 challenge. In addition, we introduced an image preprocessing pipeline and utilized a pre-trained BERT model [5] to extract question features for further processing. Last, we presented an ensemble method that boosted the performance.

References

1. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In: CLEF2019 Working Notes. CEUR Workshop Proceedings (CEUR-

WS.org), ISSN 1613-0073, http://ceur-ws.org/Vol-2380/, Lugano, Switzerland (September 9-12 2019)
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473 (2014)
3. Ben-younes, H., Cadene, R., Cord, M., Thome, N.: MUTAN: Multimodal Tucker fusion for visual question answering. In: ICCV. p. 3 (2017)
4. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
6. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint:1606.01847 (2016)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. Ionescu, B., Müller, H., Péteri, R., Cid, Y.D., Liauchuk, V., Kovalev, V., Klimuk, D., Tarasau, A., Abacha, A.B., Hasan, S.A., Datla, V., Liu, J., Demner-Fushman, D., Dang-Nguyen, D.T., Piras, L., Riegler, M., Tran, M.T., Lux, M., Gurrin, C., Pelka, O., Friedrich, C.M., de Herrera, A.G.S., Garcia, N., Kavallieratou, E., del Blanco, C.R., Rodríguez, C.C., Vasillopoulos, N., Karampidis, K., Chamberlain, J., Clark, A., Campello, A.: ImageCLEF 2019: Multimedia retrieval in medicine, lifelogging, security and nature. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 10th International Conference of the CLEF Association (CLEF 2019), LNCS Lecture Notes in Computer Science, Springer, Lugano, Switzerland (September 9-12 2019)
9. Kafle, K., Kanan, C.: Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding **163**, 3–20 (2017)
10. Kim, J.H., On, K.W., Lim, W., Kim, J., Ha, J.W., Zhang, B.T.: Hadamard product for low-rank bilinear pooling. arXiv preprint: 1610.04325 (2016)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: NIPS. pp. 3294–3302 (2015)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
14. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., et al.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2015)
17. Xiao, T., Xu, Y., Yang, K., Zhang, J., Peng, Y., Zhang, Z.: The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 842–850 (2015)