TandemNet: Distilling Knowledge from Medical Images Using Diagnostic Reports as Optional Semantic References

Zizhao Zhang, Pingjun Chen, Manish Sapkota, and Lin Yang(✉)

University of Florida, Gainesville, USA
lin.yang@bme.ufl.edu

Abstract. In this paper, we introduce the semantic knowledge of medical images from their diagnostic reports to provide an inspirational network training and an interpretable prediction mechanism with our proposed novel multimodal neural network, namely TandemNet. Inside TandemNet, a language model is used to represent report text, which cooperates with the image model in a tandem scheme. We propose a novel dual-attention model that facilitates high-level interactions between visual and semantic information and effectively distills useful features for prediction. In the testing stage, TandemNet can make accurate image prediction with an optional report text input. It also interprets its prediction by producing attention on the image and text informative feature pieces, and further generating diagnostic report paragraphs. Based on a pathological bladder cancer images and their diagnostic reports (BCIDR) dataset, sufficient experiments demonstrate that our method effectively learns and integrates knowledge from multimodalities and obtains significantly improved performance than comparing baselines.

1 Introduction

In medical image understanding, convolutional neural networks (CNNs) gradually become the paradigm for various problems [1]. Training CNNs to diagnose medical images primarily follows pure engineering trends in an end-to-end fashion. However, the principles of CNNs during training and testing is difficult to interpret and justify. In clinical practice, domain experts teach learners by explaining findings and observations to make a disease decision rather than leaving learners to find clues from images themselves.

Inspired by this fact, in this paper, we explore the usage of semantic knowledge of medical images from their diagnostic reports to provide explanatory supports for CNN-based image understanding. The proposed network learns to provide interpretable diagnostic predictions in the form of attention and natural language descriptions. The diagnostic report is a common type of medical record in clinics, which is comprised of semantic descriptions about the observations of biological features. Recently, we have witnessed rapid development in

M. Descoteaux et al. (Eds.): MICCAI 2017, Part III, LNCS 10435, pp. 320–328, 2017.
DOI: 10.1007/978-3-319-66179-7_37

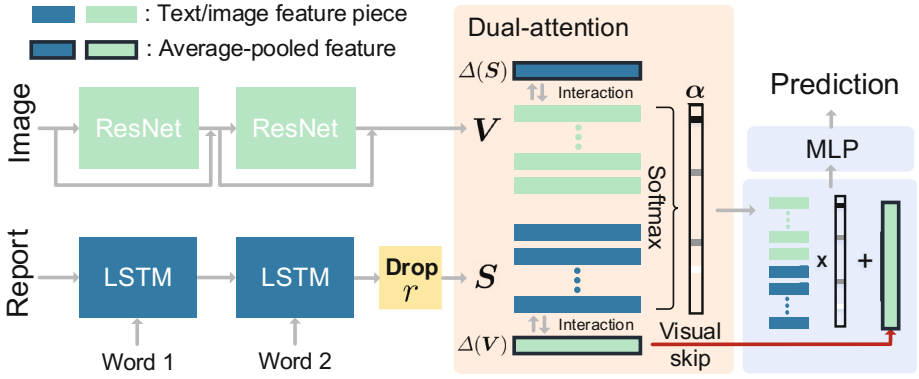

Fig. 1. The illustration of the TandemNet.

multimodal deep learning research [2,3]. We believe the joint study of multimodal data is essential towards intelligent computer-aided diagnosis. However, only a dearth of related work exists [4,5].

To take advantage of the language modality, we propose a multimodal network that jointly learns from medical images and their diagnostic reports. Semantic information is interacted with visual information to improve the image understanding ability by teaching the network to distill informative features. We propose a novel dual-attention model to facilitate such high-level interaction. The training stage uses both images and texts. In the testing stage, our network can take an image and provide accurate prediction with an optional (i.e. with or without) text input. Therefore, the language and image models inside our network cooperate with one another in a tandem scheme to either single(images)- or double(image-text)-drive the prediction process. We refer to our proposed network as TandemNet. Figure 1 illustrates the overall framework.

To validate our method, we cooperate with pathologists and doctors to collect the BCIDR dataset. Sufficient experimental studies on BCIDR demonstrate the advantages of TandemNet. Furthermore, by coupling visual features with the language model and fine-tuning the network using backpropagation through time (BPTT), TandemNet learns to automatically generate diagnostic reports. The rich outputs (i.e. attention and reports) of TandemNet have valuable meanings: providing explanations and justifications for its diagnostic prediction and making this process interpretable to pathologists.

2 Method

CNN for image modeling. We adopt the (new pre-activated) residual network (ResNet) [6] as our image model. The identity mapping in ResNet significantly improves the network generalization ability. There are many architecture variants of ResNet. We adopt the wide ResNet (WRN) [7] which has shown better

performance and higher efficiency with much less layers. It also offers scalability of the network (number of parameters) by adjusting a widen factor (i.e. the channel of feature maps) and depth. We extract the output of the layer before average pooling as our image representation, denoted as $\boldsymbol{V} \in \mathbb{R}^{C \times G}$. The input image size is 224×224, so $G = 14 \times 14$. C depends on the widen factor.

LSTM for language modeling. We adopt Long Short-Term Memory (LSTM) [8] to model diagnostic report sentences. LSTM improves vanilla recurrent neural networks (RNNs) for natural language processing and is also widely-used for multimodal applications such as image captioning [2,9]. It has a sophisticated unit design, which enables long-term dependency and greatly reduces the gradient vanishing problem in RNNs [10]. Given a sequence of words $\{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\}$, LSTM reads the words one at a time and maintains a memory state $\boldsymbol{m}_t \in \mathbb{R}^D$ and a hidden state $\boldsymbol{h}_t \in \mathbb{R}^D$. At each time step, LSTM updates them by

$$\boldsymbol{h}_t, \boldsymbol{m}_t = \text{LSTM}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}, \boldsymbol{m}_{t-1}), \tag{1}$$

where $\boldsymbol{x}_t \in \mathbb{R}^K$ is an input word, which is computed by firstly encoding it as a one-hot vector and then multiplied by a learned word embedding matrix.

The hidden state is a vector encoding of sentences. The treatment of it varies from problems. For example, in image captioning, a multilayer perceptron (MLP) is used to decode it as a predicted word at each time step. In machine translation [11], all hidden states could be used. A medical report is more formal than a natural image caption. It usually describes multiple types of biological features structured by a series of sentences. It is important to represent all feature descriptions but maintain the variety and independence among them. To this end, we extract the hidden state of every feature description (in our implementation, it is achieved by adding a special token at the end of each sentence beforehand and extracting the hidden states at all the placed tokens). In this way, we obtain a text representation matrix $\boldsymbol{S} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_N] \in \mathbb{R}^{D \times N}$ for N types of feature descriptions. This strategy has more advantages: it enables the network to adaptively select useful semantic features and determine respective feature importance to disease labels (as shown in experiments).

Dual-attention model. The attention mechanism [11,12] is an active topic in both computer vision and natural language communities. Briefly, it gives networks the ability to generate attention on parts of the inputs (like visual attention in the brain cortex), which is achieved by computing a context vector with attended information preserved.

Different from most existing approaches that study attention on images or text, given the image representation $\boldsymbol{V}$ and the report representation $\boldsymbol{S}$[1], our dual-attention model can generate attention on important image regions and sentence parts simultaneously. Specifically, we define the attention function f_{att} to compute a piece-wise weight vector $\boldsymbol{\alpha}$ as

$$\boldsymbol{e} = f_{att}(\boldsymbol{V}, \boldsymbol{S}), \quad \boldsymbol{\alpha}_i = \frac{\exp(\boldsymbol{e}_i)}{\sum_i \exp(\boldsymbol{e}_i)}, \tag{2}$$

[1] The two matrices are firstly embedded through a 1×1 convolutional layer with Tanh.

where $\boldsymbol{\alpha} \in \mathbb{R}^{G+N}$ has individual weights for visual and semantic features (i.e. $\boldsymbol{V}$ and $\boldsymbol{S}$). f_{att} is specifically defined as follows:

$$\begin{aligned} \boldsymbol{z}_{s\to v} &= \tanh(\boldsymbol{W}_v \boldsymbol{V} + (\boldsymbol{W}_{s'} \Delta(\boldsymbol{S}))\mathbb{1}_v^T), \\ \boldsymbol{z}_{v\to s} &= \tanh(\boldsymbol{W}_s \boldsymbol{S} + (\boldsymbol{W}_{v'} \Delta(\boldsymbol{V}))\mathbb{1}_s^T), \\ \boldsymbol{e} &= \boldsymbol{w}^T [\boldsymbol{z}_{s\to v}; \boldsymbol{z}_{v\to s}] + \boldsymbol{b}, \end{aligned} \tag{3}$$

where $\boldsymbol{W}_v, \boldsymbol{W}_{v'} \in \mathbb{R}^{M\times C}$ and $\boldsymbol{W}_s, \boldsymbol{W}_{s'} \in \mathbb{R}^{M\times D}$ are parameters to be learned to compute $\boldsymbol{z}_{s\to v} \in \mathbb{R}^{M\times G}$ and $z_{v\to s} \in \mathbb{R}^{M\times N}$, and $\boldsymbol{w}, \boldsymbol{b} \in \mathbb{R}^M$. $\mathbb{1}_v \in \mathbb{R}^G$ and $\mathbb{1}_s \in \mathbb{R}^N$ are vectors with all elements to be one. Δ denotes the global average-pooling operator on the last dimension of $\boldsymbol{V}$ and $\boldsymbol{S}$. $[;]$ denotes the concatenation operator. Finally, we obtain a context vector $\boldsymbol{c} \in \mathbb{R}^M$ by

$$\boldsymbol{c} = \boldsymbol{O}\,\boldsymbol{\alpha} = \sum_{i=1}^{G} \alpha_i \boldsymbol{V}_i + \sum_{j=G+1}^{G+N} \alpha_j \boldsymbol{S}_j, \text{ where } \boldsymbol{O} = [\boldsymbol{V}; \boldsymbol{S}]. \tag{4}$$

In our formulation, the computation of image and text attention is mutually dependent and conducts high-level interactions. The image attention is conditioned on the global text vector $\Delta(\boldsymbol{S})$ and the text attention is conditioned on the global image vector $\Delta(\boldsymbol{V})$. When computing the weight vector $\boldsymbol{\alpha}$, both information contributes through $\boldsymbol{z}_{s\to v}$ and $\boldsymbol{z}_{v\to s}$. We also consider extra configurations: computing two $\boldsymbol{e}$ by two $\boldsymbol{w}$, and then concatenate them to compute $\boldsymbol{\alpha}$ with one softmax or compute two $\boldsymbol{\alpha}$ with two softmax functions. Both configurations underperform ours. We conclude that our configuration is optimal for the visual and semantic information to interact with each other.

Intuitively, our dual-attention mechanism encourages better alignment of visual information with semantic information piecewise, which thereby improves the ability of TandemNet to discriminate useful features for attention computation. We will validate this experimentally.

Prediction module. To improve the model generalization, we propose two effective techniques for the prediction module of the dual-attention model.

(1) *Visual skip-connection.* The probability of a disease label p is computed as

$$p = \mathrm{MLP}(\boldsymbol{c} + \Delta(\boldsymbol{V})). \tag{5}$$

The image feature $\Delta(\boldsymbol{V})$ skips the dual-attention model and is directly added onto $\boldsymbol{c}$ (see Fig. 1). During backpropagation, this skip-connection directly passes gradients for the loss layer to the CNN, which prevents possible gradient vanishing in the dual-attention model from obstructing CNN training.

(2) *Stochastic modality adaptation.* We propose to stochastically "abandon" text information during training. This strategy generalizes TandemNet to make accurate prediction with absent text. Our proposed strategy is inspired by Dropout and the stochastic depth network [13], which are effective for model generalization. Specifically, we define a drop rate r as the probability to remove (zero-out)

Table 1. The quantitative evaluation (averaged on 3 trials). The first block shows standard CNNs so text is irrelevant.

Method	Accuracy (%)	
	w/o text	w/text
WRN16-4	75.4	–
ResNet18-TL	79.4	–
TandemNet-WVS	79.4	85.6
TandemNet	82.4	89.9
TandemNet-TL	84.9	88.6

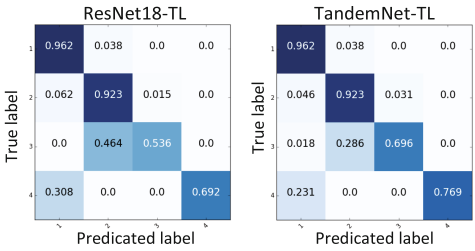

Fig. 2. The confusion matrices of two compared methods ResNet18-TL and TandemNet-TL (w/o text) in Table 1.

the text part $\boldsymbol{S}$ during the entire network training stage. Thus, based to the principle of Dropout, $\boldsymbol{S}$ will be scaled by $1 - r$ if text is given in testing.

The effects of these two techniques are discussed in experiments.

3 Experiments

Dataset. To collect the BCIDR dataset, whole-slide images were taken using a 20X objective from hematoxylin and eosin (H&E) stained sections of bladder tissue extracted from a cohort of 32 patients at risk of a papillary urothelial neoplasm. From these slides, 1,000 500×500 RGB images were extracted randomly close to urothelial regions (each patient's slide yields a slightly different number of images). For each of these images, the pathologist then provided a paragraph describing the disease state. Each paragraph addresses five types of cell appearance features, namely the state of *nuclear pleomorphism*, *cell crowding*, *cell polarity*, *mitosis*, and *prominence of nucleoli* (thus $N = 5$). Then a conclusion is decided for each image-text pair, which is comprised of four classes, i.e. *normal* tissue, *low-grade* (papillary urothelial neoplasm of low malignant potential) carcinoma, *high-grade* carcinoma, and *insufficient information*. Following the same procedure, four doctors (not experts in the bladder cancer) wrote additional four descriptions for each image. They also refer to the pathologist's description to make sure their annotation accuracy. Thus there are five ground-truth reports per image and 5,000 image-text pairs in total. Each report varies in length between 30 and 59 words. We randomly split 20% (6/32) of patients including 1,000 samples as the testing set and the remaining 80% of patients including 4,000 samples (20% as the validation set for model selection) for training. We subtract the data RGB mean and augment through clip, mirror and rotation.

Implementation details. Our implementation is based on Torch7. We use a small WRN with depth $= 16$ and widen-factor $= 4$ (denoted as WRN16-4), resulting in 2.7M parameters and $C = 256$. We use dropout with 0.3 after each convolution. We use $D = 256$ for LSTM, $M = 256$, and $K = 128$. We use

SGD with a learning rate $1e{-}2$ for the CNN (used likewise for standard CNN training for comparison) and Adam with $1e{-}4$ for the dual-attention model, which are multiplied by 0.9 per epoch. We also limit the gradient magnitude of the dual-attention model to 0.1 by normalization [10].

Diagnostic prediction evaluation. Table 1 and Fig. 2 show the quantitative evaluation of TandemNet. For comparison with CNNs, we train a WRN16-4 and also a ResNet18 (has 11M parameters) pre-trained on ImageNet[2]. We found transfer learning is beneficial. To test this effect in TandemNet, we replace WRN16-4 with a pre-trained ResNet18 (TandemNet-TL). As can be observed, TandemNet and TandemNet-TL significantly improve WRN16-4 and ResNet18-TL when only images are provided. We observe TandemNet-TL slightly underperforms TandemNet when text is provided with multiple trails. We hypothesize that it is because fine-tuning a model pre-trained on a complete different natural image domain is relatively hard to get aligned with medical reports in the dual-attention model. From Fig. 2, *high grade* (label id 3) is more likely to be misclassified as *low grade* (2) and some *insufficient information* (4) is confused with *normal* (1).

We analyze the text drop rate in Fig. 3 (left). When the drop rate is low, the model obsessively uses text information, so it achieves low accuracy without text. When the drop rate is high, the text can not be well adapted, resulting in decreased accuracy with or without text. The drop rate of 0.5 performs best and thereby is used in this paper. As illustrated in Fig. 3, we found that the classification of text is easier than images, therefore its accuracy is much higher. However, please note that the primary aim of this paper is to use text information only at the training stage. While at the testing stage, the goal is to accurately classify images without text.

In Eq. (5), one question that may arise is that, when testing without text, whether it is merely $\Delta(\boldsymbol{V})$ from the CNN that produces useful features rather than $\boldsymbol{c}$ from the dual-attention model (since the removal (zero-out) of $\boldsymbol{S}$ could possibly destroy the attention ability). To validate the actual role of $\boldsymbol{c}$, we remove the visual skip-connection and train the model (denoted as TandemNet-WVS

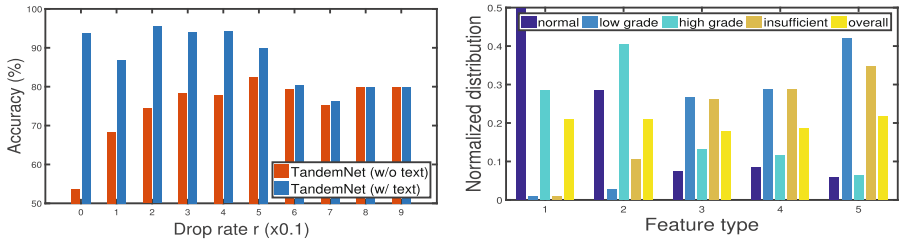

Fig. 3. Left: The accuracy with varying drop rates. Right: The averaged text attention per feature type (and overall) to each disease label. The feature type is specified in the text of dataset introduction (in order).

[2] Provided by https://github.com/facebook/fb.resnet.torch.

in Table 1) and it improves ResNet16-4 by 4% without text. The qualitative evaluation below also validates the effectiveness of the dual-attention model. Additionally, we use the (t-distributed Stochastic Neighbor Embedding) t-SNE dimensionality reduction technique to examine the input of MLP in Fig. 4.

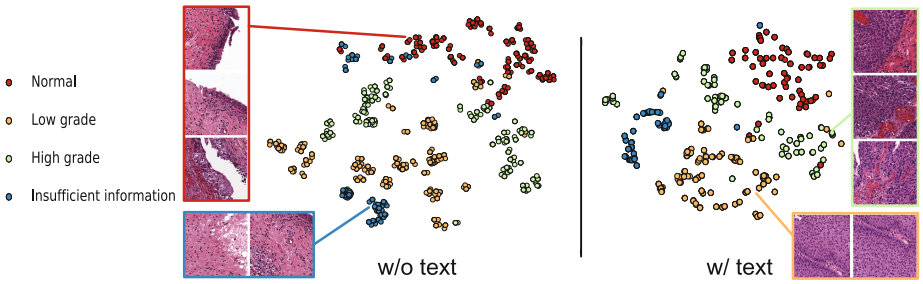

Fig. 4. The t-SNE visualization of the MLP input. Each point is a test sample. The embeddings with text (right) results in better distribution.

Attention analysis. We visualize the attention weights to show how TandemNet captures image and text information to support its prediction (the image attention map is computed by upsampling the $G = 14 \times 14$ weights of $\boldsymbol{\alpha}$ to the image space). To validate the visual attention, without notifying our results beforehand, we ask the pathologist to highlight regions of some test images they think are important. Figure 5 illustrates the performance. Our attention maps show surprisingly high consistency with pathologist's annotations. The attention without text is also fairly promising, although it is less accurate than the results with text. Therefore, we can conclude that TandemNet effectively uses semantic information to improve visual attention and substantially maintains such attention capability though the semantic information is not provided. The text attention is shown in the last column of Fig. 5. We can see that our text attention result is quite selective in only picking up useful semantic features.

Furthermore, the text attention statistics over the dataset provides particular insights into the pathologists' diagnosis. We can investigate which feature contributes the most to which disease label (see Fig. 3 (right)). For example, *nuclear pleomorphism* (feature type 1) shows small effects on the *low-grade* disease label. *cell crowding* (2) has large effects on *high-grade*. We can justify the reason of text attention by closely looking at images of Fig. 5: *high grade* images have obvious high *cell crowding* degree. Moreover, this result strongly demonstrates the successful image-text alignment of our dual-attention model.

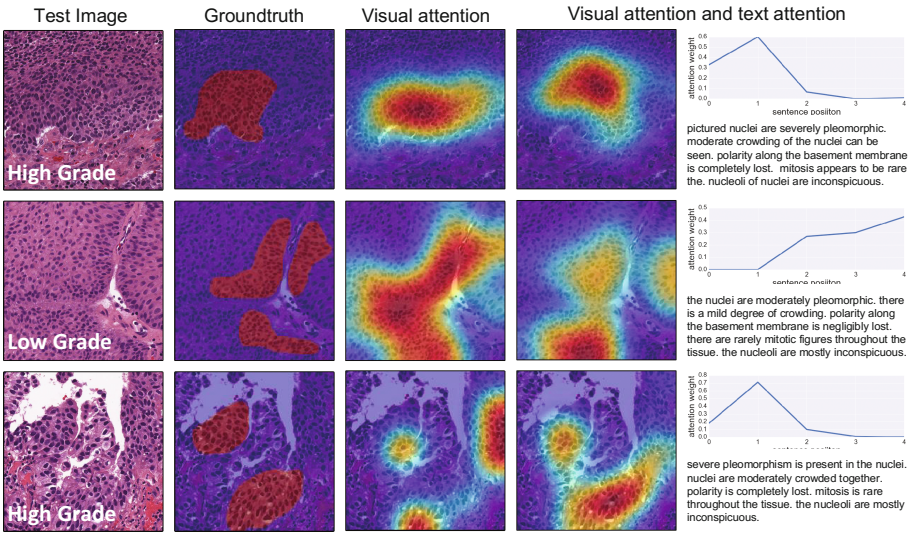

Fig. 5. From left to right: Test images (the bottom shows disease labels), pathologist's annotations, visual attention w/o text. visual attention and corresponding text attention (the bottom shows text inputs). Best viewed in color.

Image report generation. We fine-tune TandemNet using BPTT as an extra supervision and use the visual feature $\Delta(\boldsymbol{V})$ as the input of LSTM at the first time step[3]. We direct readers to [9] about detailed LSTM training for image captioning. Figure 6 shows our promising results compared with pathologist's descriptions. We leave the full report generation task as a future study.

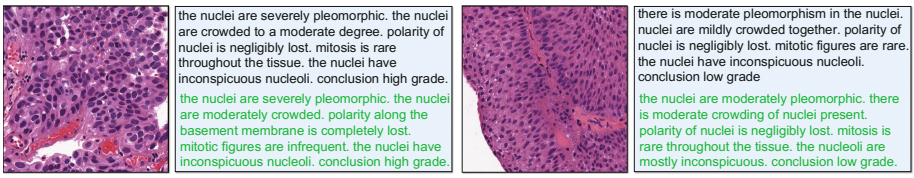

Fig. 6. The pathologist's annotations are in black and the automatic results of TandemNet are in green, which accurately describe the semantic concepts.

4 Conclusion

This paper proposes a novel multimodal network, TandemNet, which can jointly learn from medical images and diagnostic reports and predict in an interpretable

[3] We freeze the CNN for the whole training and the dual-attention model for the first 5 epochs, and then fine-tune with a smaller learning rate, $5e{-}5$.

scheme through a novel dual-attention mechanism. Sufficient and comprehensive experiments on BCIDR demonstrate that TandemNet is favorable for more intelligent computer-aided medical image diagnosis.

References

1. Greenspan, H., van Ginneken, B., Summers, R.M.: Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. TMI **35**(5), 1153–1159 (2016)
2. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR, pp. 3156–3164 (2015)
3. Xu, T., Zhang, H., Huang, X., Zhang, S., Metaxas, D.N.: Multimodal deep learning for cervical dysplasia diagnosis. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 115–123. Springer, Cham (2016). doi:10.1007/978-3-319-46723-8_14
4. Shin, H.C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., Summers, R.M.: Learning to read chest x-rays: recurrent neural cascade model for automated image annotation. In: CVPR, pp. 2497–2506 (2016)
5. Zhang, Z., Xie, Y., Xing, F., Mcgough, M., Yang, L.: MDNet: a semantically and visually interpretable medical image diagnosis network. In: CVPR (2017)
6. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). doi:10.1007/978-3-319-46493-0_38
7. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
9. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: CVPR, pp. 3128–3137 (2015)
10. Pascanu, R., Mikolov, T., Bengio, Y.: On the difficulty of training recurrent neural networks. In: ICML, pp. 1310–1318 (2013)
11. Luong, M.T., Pham, H., Manning, C.D.: Effective approaches to attention-based neural machine translation. In: EMNLP, pp. 1412–1421 (2015)
12. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: ICML, pp. 2048–2057 (2015)
13. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q.: Deep networks with stochastic depth. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 646–661. Springer, Cham (2016). doi:10.1007/978-3-319-46493-0_39