Spatial Semantic-Preserving Latent Space Learning for Accelerated DWI Diagnostic Report Generation

Aydan Gasimova[1⋆], Gavin Seegoolam[1⋆], Liang Chen[1], Paul Bentley[2], Daniel Rueckert[1]

[1] BioMedIA, Department of Computing, Imperial College London, SW7 2AZ
{ag6516,kgs13,lc12,dr}@ic.ac.uk
[2] Department of Brain Sciences, Faculty of Medicine, Imperial College London, SW7 2AZ
{p.bentley}@imperial.ac.uk

Abstract. In light of recent works exploring automated pathological diagnosis, studies have also shown that medical text reports can be generated with varying levels of efficacy. Brain diffusion-weighted MRI (DWI) has been used for the diagnosis of ischaemia in which brain death can follow in immediate hours. It is therefore of the utmost importance to obtain ischaemic brain diagnosis as soon as possible in a clinical setting. Previous studies have shown that MRI acquisition can be accelerated using variable-density Cartesian undersampling methods. In this study, we propose an accelerated DWI acquisition pipeline for the purpose of generating text reports containing diagnostic information. We demonstrate that we can learn a semantic-preserving latent space for minor as well as extremely undersampled MR images capable of achieving promising results on a diagnostic report generation task.

1 Introduction

Patients that have suffered the symptoms of a stroke have a very short time frame in which to be effectively treated; therefore, it is imperative that radiologists determine the cause of the symptoms in order to provide the appropriate treatment. The majority of strokes are caused by cerebral ischaemia, which can be characterised as reduced blood flow to the brain, causing poor oxygenation that can lead to permanent brain cell death. Both computed tomography (CT) and multi-modal magnetic resonance imaging (MRI) are effective in assessing brain ischaemia, but diffusion-weighted MRI (DWI) is particularly advantageous as it provides highest sensitivity to early ischaemic lesions. In comparison to CT, typical DWI has a much longer acquisition time which additionally makes the scans more susceptible to patient motion and subsequent unwanted imaging artefacts. Furthermore, requiring patients to lay dormant without any motion for long periods of time may lead to discomfort. A well-explored approach for

⋆ both authors contributed equally to this study

accelerating scan-time is through *undersampling* whereby fewer scanner measurements are taken, violating the Nyquist-Shannon sampling theorem and thus introducing aliasing artefacts into the reconstruction of the image. Several studies are focused on the dealiasing of such images, validating undersampled MRI as an accepted acceleration technique [17, 14, 13, 4, 27, 27, 5, 15].

Assessing the quality of the MR image reconstruction is typically focused on calculating similarity metrics such as peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) index between the dealiased reconstruction and the fully-sampled image [20]. This does not, however, guarantee the retention of pathological features necessary for a diagnosis, especially at more aggressive acceleration rates. Therefore, a complimentary way of reviewing extremely accelerated images is through the use of real-time diagnostic tasks such as segmentation and classification [16]. In our study, we explore the automated generation of radiological text reports containing relevant diagnostic and contextual information. The logging of diagnostic reports generated by qualified radiologists is standard hospital protocol. As a result, datasets for studies involving automated text report generation can be acquired directly from hospital archives. In contrast, segmentation and classification tasks require non-standard time-consuming manual annotations. In addition, DWI diagnostic reports typically detail contextual information as well as the presence/absence of an acute lesion, such as anatomical location and severity of the lesion, and being able to auto-generate them will additionally expedite the process of identifying and documenting acute ischaemia.

To this end we have developed a pipeline that 1) learns an implicit context-preserving manifold of brain DWIs that captures both spatial and pathological information, 2) enforces a latent code for the accelerated DWIs that performs in a similar fashion to the fully-sampled images 3) utilises these accelerated brain DWI image representations to learn to automatically generate reports using a recurrent neural network. To our knowledge, this is the first demonstration of deep latent space learning for the retention of semantic feature information required for accelerated report generation, and the first demonstration of learning to auto-generate reports from brain DWI images.

2 Previous work

Latent space learning of accelerated MRI Previous work has shown the use of deep latent space learning for performing tasks such as segmentation and reconstruction in the context accelerated MRI [27, 16]. Accelerated MRI data acquisition is centred around the ability to reconstruct image data in a typically ill-conditioned inverse regression problem. However, certain tasks will only require certain parts of information from the sensor space, called 'k-space'. For example, approximate motion estimation from cardiac cine MRI can be performed with acceleration rates of 51.2 [17]. [16] shows that cardiac segmentation can be performed by a single line acquisition in k-space. Inspired by this we explore the use of deep latent space learning for learning diagnostically-relevant

contextual image embeddings. Whilst [16] shows that deep latent space learning provides a manifold that can be robust to different undersampling patterns, they also show that at extreme acceleration rates, deep latent space learning can outperform conventional approaches.

Radiology report generation Learning to automate report generation for radiological images has thus far been heavily influenced by image captioning models formulated as an encoder-decoder machine translation problem. In image captioning, image representations are extracted from a pre-trained convolutional neural network (the encoder) and passed as inputs alongside captions to a sequence-learning decoder by, for instance, mapping the word and image representations to the same feature space [10, 19]. Such a framework was used by [18] to predict structured medical subject heading (MeSH®) annotations for chest X-ray images.

More recently, learning to attend to spatial visual features has been shown to be effective in image captioning [23] and medical report generation [26, 7, 24, 25]. Using structured reports in a dual-attention framework, Zhang et al.[26] were able to improve features used for classifying histopathology images. The co-attention network of Jing et al. [7] is fed visual as well as semantic features in order to provide high-level semantic information to the text-generation task. Xue et al.[24] break down the task of report generation into subtasks of generating one sentence at a time where each succeeding sentence is conditioned on image features and previous sentences. Yuan et al.[25] also demonstrate the benefit of learning radiology-related features from an initial classification task and go a step further by learning features from multi-view 2-D images (chest X-rays) by introducing a cross-view consistency loss.

The accelerated acquisition of brain DWI has been previously studied in the context of image reconstruction [11, 22, 21, 2]. However, in our study, we explore its use for automated text report generation. We demonstrate how the latent space learned by the accelerated reconstruction network captures both spatial semantic and pathology information required in order to learn to generate reports.

3 Method

Our study accelerates DWI acquisition through aggressive variable-density Cartesian undersampling as has been studied in several previous works such as [17, 16]. In our study, we start with attempting a zero-fill reconstruction whereby the lines in k-spaces that are not acquired are filled with zeros. An example of a fully sampled image and a corresponding undersampled, zero-filled image reconstruction is shown in Figure 2. For all acceleration rates, we always sample the two most central lines in k-space whilst the other lines are acquired following a Gaussian distribution centred at the point of highest energy in k-space. During training, undersampling masks are generated on the fly and images are also augmented with additional rotations and translations.

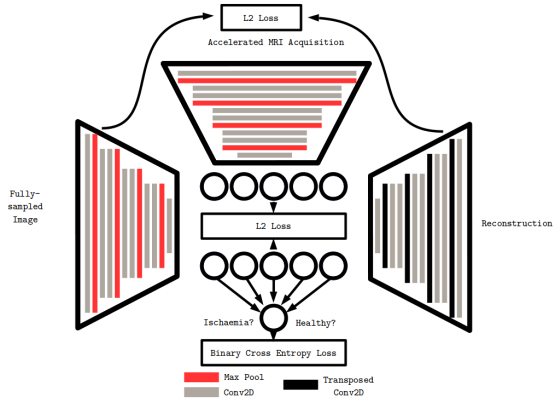

Fig. 1. An autoencoder is trained to reconstruct the fully-sampled image through an L2 loss. The latent space is conditioned to encode pathological information by performing a classification of ischaemia, trained with a binary cross-entropy loss. The latent space encoding learned at the bottleneck is used as a training target for the encoding branch which only sees the accelerated image.

3.1 Latent space learning

In our approach, we use an autoencoder network that takes as input the original fully-sampled DWI brain MRI. The purpose of this is to learn a latent space at the bottleneck that contains spatial and contextual information that may be useful for a text report generator. In particular, we manipulate the embedding manifold toward one more suitable for text report generation by introducing an ischaemia-classification loss as a regulariser. This loss can be summarised by equation (1) where an Adam optimiser with learning rate 1.0×10^{-5}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ was used.

$$L(x, y) = ||D(E(x)) - x||_2^2 - \gamma(y \log C(E(x)) + (1 - y) \log(1 - C(E(x)))), \quad (1)$$

where E, D and C are the encoder, decoder and classifier networks (from figure 1) respectively, x is our fully-sampled image, y is a binary classification label for ischaemia and $\gamma = 8000$. We can measure the performance of the latent space learnt as a combination of reconstruction error (in particular of the ischaemia) and of the classification error.

Along side this, we use a structurally-identical encoding branch to learn a latent space for the accelerated MRI acquisition. We use the approach of performing a zero-fill reconstruction whereby after convolutional layers can be used to identify aliasing artefacts as directly relevant image features themselves. The latent space is trained against the bottleneck of the autoencoder using an L2 loss and another Adam optimizer with the same optimizer parameters. This is summarised in Figure 1 and in equation (2). Note, for each acceleration rate used in our study, a unique encoder is learned to generate the required latent

space. An advantage of deep latent space learning is that we can train the specific encoder associated with different acceleration rates towards the same manifold which avoids the need for retraining of the text report generator model.

$$L(x, x_{\text{acc}}) = ||E(x) - x_{\text{acc}}||_2^2, \tag{2}$$

where x_{acc} is our accelerated, aliased image and E_{acc} is our encoding branch for the accelerated images.

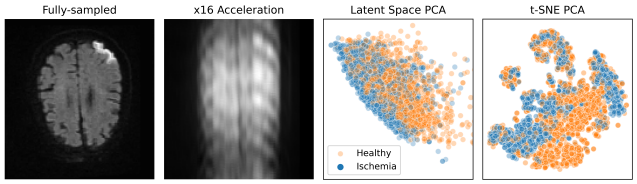

Fig. 2. Left to right: (1) An example of a brain with ischaemia (2) The corresponding x16 accelerated image is zero-fill reconstructed from k-space using a 2D Fourier Transform. Note that this image is infected with heavy aliasing artefacts. (3) A projection of the first two principle components in a PCA analysis of the latent space. Some clustering can be seen (4) a t-SNE projection of the latent space showing clear clustering.

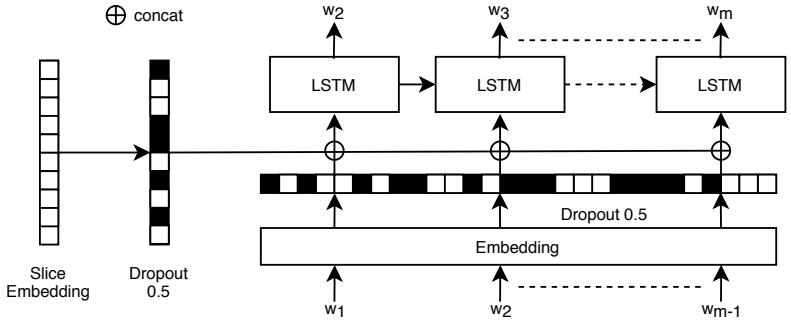

Fig. 3. Clinical report generation model from accelerated image latent space embeddings.

3.2 Report generation model

We use a report generation model based on [3] where the report word sequence is modelled using the Long Short-Term Memory (LSTM)[6], and conditioned

on image embeddings at each time step through concatenation at the input to the LSTM. At each time step, the input, output and forget gates control how much of the previous time steps is propagated through to the output. For an input embedding sequence $\{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^D$, the internal hidden state $h_t \in \mathbb{R}^h$ and memory state $m_t \in \mathbb{R}^m$ are updated as follows:

$$\begin{aligned} h_t &= f_t \odot h_{t-1} + i_t \odot \tanh(W^{(hx)} x_t + W^{(hm)} m_{t-1}) \\ m_t &= o_t \odot \tanh(h_t) \end{aligned} \tag{3}$$

where $x_t \in \mathbb{R}^D$ is the concatenation of the latent space image embedding and word embedding at time step t, $W^{(hx)}$ and $W^{(hm)}$ are the trainable weight parameters, and i_t, o_t and f_t are the input, output and forget gates respectively. The model architecture is illustrated in Figure 3. We additionally add Dropout layers after image and word embeddings to force the model to condition on both thus regularising training.

4 Experiments

The Data The dataset consists of 1226 DWI scans and corresponding radiological reports of acute stroke patients. All the images and reports were fully anonymised and ethical approval was granted by Imperial College Joint Regulatory Office. The scans were pre-processed according to the steps outlined in [1]: images were resampled into uniform pixel size of 1.6×1.6mm, and pixel intensities were normalised to zero mean and unit variance. The number of slices per image varies between 7 and 52, and the slice dimensions are 128×128.

Each report contains between 1 and 2 sentences summarising the presence or absence of the pathology, a visual description, and its location within the brain. In addition, each exam is assigned a diagnostic label as part of hospital protocol: 54% were diagnosed 'no acute infarct', 46% were diagnosed 'acute infarct'. The remaining, which made up a total of <1% and included diagnoses such as 'unknown', 'haematoma', 'tumour', were removed for the purpose of training. Processing was done on the reports to remove words outside the 99th percentile, exams with empty reports were removed, leaving a total of 1104 exams, total vocab length 1021, mean words per exam 10.8, std. 6.3.

In order to simplify the problem, we created a 2D dataset of acute and non-acute (normal) slices from these images. For the acute set, we used the brain ischemia segmentation network developed by Chen et al.[1] to segment the images labelled with acute ischemia, thresholded at 0.8, and selected slices where the total area of ischemia was >10 pixels. For the normal set, we sampled slices from the non-acute labelled images according to the same axial plane distribution as the acute set.

Experimental settings Reports were padded with 'start' and 'end' tokens to length 19 (mean + 1std. + 'start' + 'end'). The word embedding layer maps one-hot encoded word embeddings into a 256 dimensional space. The LSTM

hidden state is also set to dim 256, and the LSTM units are unrolled up to 19 time steps. We train the model on non-accelerated latent embeddings and their associated reports by minimising the categorical cross-entropy loss over the generated words. All models are trained with batch size 128, using Adam optimisation [8], learning rate=0.0001 for 300 epochs.

Results Inference was performed by first sampling from the LSTM using a 'start' token concatenated with the accelerated embeddings, and consequently appending the output word embedding to the input and sampling until an 'end' token was reached. The quality of the generated reports was evaluated by measuring BLUE [12] and ROUGE [9] scores averaged over all the reports, which are a form of n-gram precision commonly used for evaluating image captioning as they maintain high correlation with human judgement. We observe that the both the BLEU and ROUGE scores decrease with increasingly accelerated images, as expected. We note that there is a significant reduction in performance between the x4 and x8 accelerated images possibly due to some contextual information not being captured by the latent space.

We also assess the sampled reports qualitatively in Figure 4. We observe no major grammatical errors for all accelerations, an no major content errors for lower accelerations with x2 and x4 correctly identifying the presence/absence of ischemia as well as the location. Note: the last example shows a text report that was ischemic but was classified as healthy. This is likely to have confused the latent code for this example resulting in poor text report generations.

Table 1. BLEU1,2,3,4-gram and ROUGE1 f1, precision (P) and recall (R) metric comparisons on increasingly accelerated image embeddings.

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1 F1	ROUGE-1 P	ROUGE-1 R
Acc.×1	38.12	27.26	20.28	15.59	47.10	52.89	44.96
Acc.×2	34.07	23.31	15.55	11.57	44.00	51.86	40.68
Acc.×4	31.36	19.42	12.29	8.31	41.17	48.09	38.80
Acc.×8	21.32	10.37	5.06	2.55	29.53	32.92	29.52
Acc.×64	21.58	11.11	4.97	2.35	30.39	35.10	29.07

5 Conclusion and future work

We demonstrate how a latent space capturing pathalogical and spatial information can be learned from accelerated brain DWI images and subsequently used to train a diagnostic report generation network with promising results. In future works, we wish to explore radial undersampling trajectories for DWI brain imaging which are expected to provide improved diagnostic embeddings.

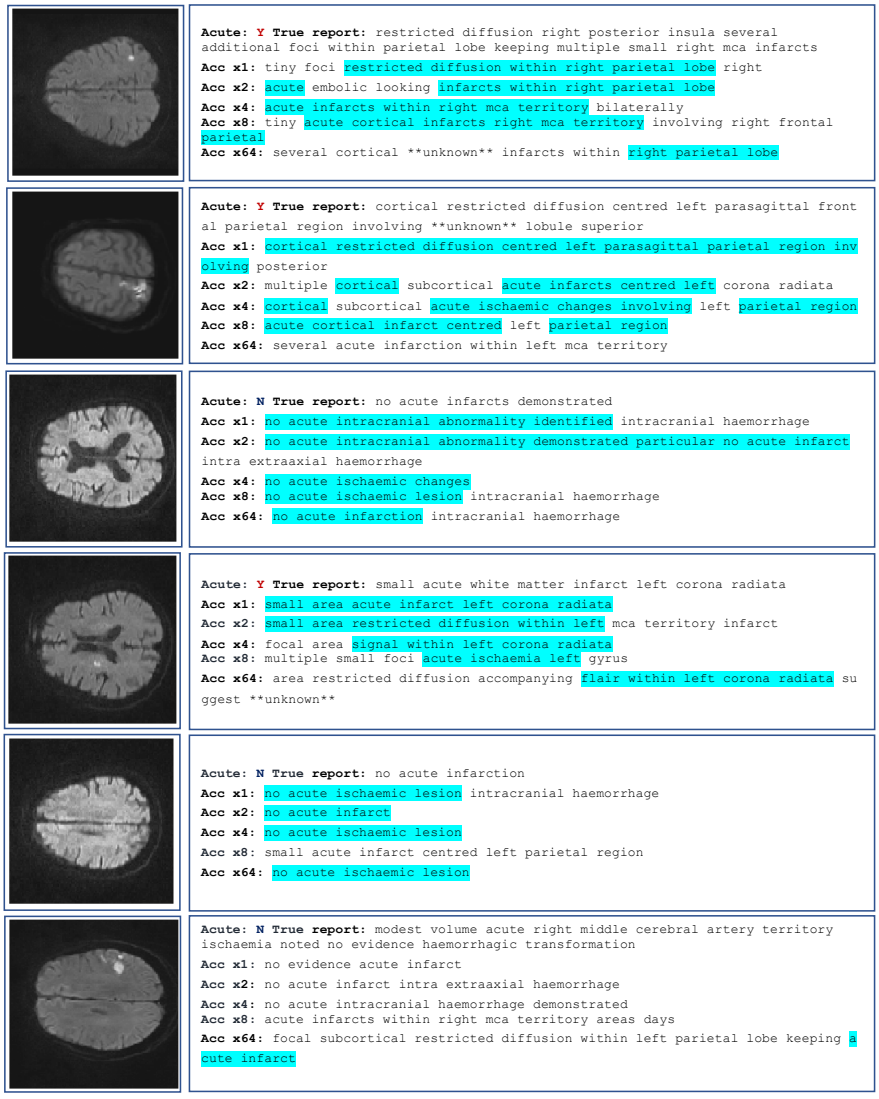

Fig. 4. Sample brain slices and associated reports generated from non-accelerated and increasingly accelerated image embeddings. Correctly identified pathology (acute/non-acute) and spatial contexts are highlighted in blue.

References

1. Chen, L., Bentley, P., Rueckert, D.: Fully automatic acute ischemic lesion segmentation in dwi using convolutional neural networks. NeuroImage: Clinical **15**, 633–643 (2017)
2. Ciritsis, A., Rossi, C., Marcon, M., Van, V.D.P., Boss, A.: Accelerated diffusion-weighted imaging for lymph node assessment in the pelvis applying simultaneous multislice acquisition: a healthy volunteer study. Medicine **97**(32) (2018)
3. Gasimova, A.: Automated enriched medical concept generation for chest x-ray images. In: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support, pp. 83–92. Springer (2019)
4. Griswold, M.A., Jakob, P.M., Heidemann, R.M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., Haase, A.: Generalized autocalibrating partially parallel acquisitions (grappa). Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **47**(6), 1202–1210 (2002)
5. Hammernik, K., Klatzer, T., Kobler, E., Recht, M.P., Sodickson, D.K., Pock, T., Knoll, F.: Learning a variational network for reconstruction of accelerated mri data. Magnetic resonance in medicine **79**(6), 3055–3071 (2018)
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
7. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586 (2018)
8. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
9. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. pp. 150–157 (2003)
10. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). ICLR (2015)
11. Merrem, A., Hofer, S., Voit, D., Merboldt, K.D., Klosowski, J., Untenberger, M., Fleischhammer, J., Frahm, J., et al.: Rapid diffusion-weighted magnetic resonance imaging of the brain without susceptibility artifacts: Single-shot steam with radial undersampling and iterative reconstruction. Investigative radiology **52**(7), 428–433 (2017)
12. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
13. Pruessmann, K.P., Weiger, M., Scheidegger, M.B., Boesiger, P.: Sense: sensitivity encoding for fast mri. Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine **42**(5), 952–962 (1999)
14. Qin, C., Schlemper, J., Caballero, J., Price, A.N., Hajnal, J.V., Rueckert, D.: Convolutional recurrent neural networks for dynamic mr image reconstruction. IEEE transactions on medical imaging **38**(1), 280–290 (2018)
15. Schlemper, J., Caballero, J., Hajnal, J.V., Price, A.N., Rueckert, D.: A deep cascade of convolutional neural networks for dynamic mr image reconstruction. IEEE transactions on Medical Imaging **37**(2), 491–503 (2017)

16. Schlemper, J., Oktay, O., Bai, W., Castro, D.C., Duan, J., Qin, C., Hajnal, J.V., Rueckert, D.: Cardiac mr segmentation from undersampled k-space using deep latent representation learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 259–267. Springer (2018)
17. Seegoolam, G., Schlemper, J., Qin, C., Price, A., Hajnal, J., Rueckert, D.: Exploiting motion for deep learning reconstruction of extremely-undersampled dynamic mri. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 704–712. Springer (2019)
18. 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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2497–2506 (2016)
19. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 3156–3164. IEEE (2015)
20. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
21. Weiss, J., Martirosian, P., Taron, J., Othman, A.E., Kuestner, T., Erb, M., Bedke, J., Bamberg, F., Nikolaou, K., Notohamiprodjo, M.: Feasibility of accelerated simultaneous multislice diffusion-weighted mri of the prostate. Journal of Magnetic Resonance Imaging **46**(5), 1507–1515 (2017)
22. Wu, W., Miller, K.L.: Image formation in diffusion mri: a review of recent technical developments. Journal of Magnetic Resonance Imaging **46**(3), 646–662 (2017)
23. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
24. Xue, Y., Xu, T., Long, L.R., Xue, Z., Antani, S., Thoma, G.R., Huang, X.: Multimodal recurrent model with attention for automated radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 457–466. Springer (2018)
25. Yuan, J., Liao, H., Luo, R., Luo, J.: Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 721–729. Springer (2019)
26. Zhang, Z., Xie, Y., Xing, F., McGough, M., Yang, L.: Mdnet: A semantically and visually interpretable medical image diagnosis network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6428–6436 (2017)
27. Zhu, B., Liu, J.Z., Cauley, S.F., Rosen, B.R., Rosen, M.S.: Image reconstruction by domain-transform manifold learning. Nature **555**(7697), 487–492 (2018)