Comparison of Visual Attention Networks for Semantic Image Segmentation in Reminiscence Therapy

Liane Meßmer* and Christoph Reich†
Institut for Data Science, Cloud Computing and IT-Security,
Furtwangen University of Applied Science
Furtwangen, Germany
Email: {*l.messmer, †christoph.reich}@hs-furtwangen.de

***Abstract*—Due to the steadily increasing age of the entire population, the number of dementia patients is steadily growing. Reminiscence therapy is an important aspect of dementia care. It is crucial to include this area in digitization as well. Modern Reminiscence sessions consist of digital media content specifically tailored to a patient's biographical needs. To enable an automatic selection of this content, the use of Visual Attention Networks for Semantic Image Segmentation is evaluated in this work. A detailed comparison of various Neural Networks is shown, evaluated by Metric for Evaluation of Translation with Explicit Ordering (METEOR) in addition to Billingual Evaluation Study (BLEU) Score. The most promising Visual Attention Network consists of a Xception Network as Encoder and a Gated Recurrent Unit Network as Decoder.**

Keywords–Visual Attention Networks; Image Caption Generation; Dementia Health Care; BLEU; METEOR.

I. INTRODUCTION

Demand-oriented technical solutions can make a valuable contribution to the care of people suffering from dementia - People With Dementia (PWD)s. Their potential is far from being exhausted. Digital media, which are used today, e.g., on tablets in the context of memory care, have considerable potential for the individualization of care offers, which are also becoming increasingly important because of the increasing differentiation of lifestyles in care for the elderly [1].

Reminiscence therapy is used to address the activation process of people with Dementia [2]. However, the identification of suitable content, as well as the design and evaluation of high-quality reminiscence sessions is very labor-intensive and places high qualification demands on care workers. Suitable individual contents for PWDs currently have to be identified "manually" and evaluated in terms of their suitability. In practice, a very limited pool of standard content is therefore often used. Dynamic response to interaction with residents is also not possible with the tools currently available. The individual activation and care required for high-quality, biography-based care (as opposed to mere occupation) therefore remains a major challenge, despite the extensive availability of digital content today.

Semantic segmentation is a well known technique, in the field of computer vision, and the basis to full understanding of a scene. With the popularity of Deep Learning in recent years, many semantic segmentation problems are being addressed with Deep Learning architectures that far outperform other approaches in terms of accuracy and efficiency. Image description models typically consist of an encoder-decoder architecture. Most commonly, Convolutional Neural Networks (CNN)s are used as encoders for image feature extraction and Recurrent Neural Networks (RNN)s are used as decoders for image description modelling [3]. This work analyses the potential of the Convolutional Neural Networks Inceptionv3, VGG16/19-Net, ResNet101 and Xception in combination with the Recurrent Neural Networks Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), concerning their suitability for use in an image search and selection system for people with dementia.

The activation of PWD requires the selection of pictures according to the following picture characteristics such as color, shapes, amount of objects, meaning according to life themes, etc. Ji et al. [4] present a system that allows images to be grouped together based on a domain-specific ontology. Through this approach, a precise image selection, e.g., using a life topic ontology - can be realized. Jaiswal, Liu and Frommholz [5] describe a selection system with user personalization in which "Information Foraging Theory" is applied to explain the suggestions. This allows users to be made "transparent" as to why an image was selected or suggested in the first place.

The dataset used for the model analysis in the context of PWD must be suitable for the task of semantic image description on the one hand, but also fit the life topic ontology of PWD so that suitable content can be identified to activate them. Due to the wide range of different categories in the Microsoft Common Objects in Context (COCO) dataset, many life themes can be covered, such as animals, people or leisure activities. Furthermore, the dataset provides image descriptions that can be used for feature extraction by a CNN, as well as for semantic image description by the RNN. Therefore, the COCO dataset [6] is particularly suitable in the context of semantic segmentation of image content for PWD.

The aim of the work is to compare different VAT architectures for semantic image segmentation, using CNNs and RNNs in the context of people with dementia, and thereby enable

Copyright (c) IARIA, 2023. ISBN: 978-1-61208-950-8

automatic identification, as well as selection of appropriate activation session content from available digital images. The focus is on image features such as the meaning of life themes, colors, shapes and quantities of objects. In particular, an automated, individual and biography-related media selection improves the quality of the session and relieves the caregivers by shortening the preparation time.

This work consists of 7 sections. Section 2 deals with the related work. Semantic Image Segmentation with Visual Attention Networks (VATs) is described in Section 3. In Section 4, the data used for training is presented and explained. The training process, is described in Section 5. Finally, the results are presented in Section 6 and Section 7 describes the conclusion and future work.

II. Related Work

Alm et al. [7] proposed the first project that developed an Application for digital reminiscence therapy was the Computer Interactive and Conversation Aid (CIRCA) Project. This project aimed to support Dementia patients with digital reminiscence sessions. Over the years, it was supplemented by different new technologies, like a specific interface for the interaction with the System [8] or a touch screen computer to enable an easier interaction with the system [9]. Today, CIRCA is an interactive multimedia application, which supports digital pictures, video and music. The latest publication from Astell, Smith, Potter and Preston-Jones [10] was the work "Computer Interactive Reminiscence and Conversation Aid groups - Delivering cognitive stimulation with technology" which demonstrates the effectiveness of CIRCA for group interventions.

The work "Interactive memories - technology-aided reminiscence therapy for people with dementia" from Klein and Uhlig [11] published an approach to support reminiscence caregivers in their work, so that appropriate images for the sessions can be selected more easily, namely by automatic labeling of available content. They used mixed reality user interfaces to help the user explore media artifacts of their individual biography and spark conversations with caregivers and family . Therefore, the Multimedia content of a therapy session has to be identified manually, that is where our work comes in. We evaluate different architectures of Visual Attention Networks to automatically describe images that match the biographical content of a dementia patient to facilitate and improve the quality of Reminiscence sessions.

The network architecture in this paper is based on the approach proposed by Xu et al. [12], in the work "Show, Attend and Tell". They describe the architecture of a Visual Attention Network that uses CNNs as Encoder an RNNs as decoder, with an additional attention layer inside of the RNN network. With the attention layer the network is able to select the focus on specific parts of an image instad of processing the image as a whole.

A comparison of different Visual Attention Network architectures is presented by Ankit, Subasish, Anuveksh and Vinay [13] in the work "Image Captioning and Comparison of Different Encoders". They compare different CNNs as Encoder configuration like Inceptionv3, VGG16, VGG19 and InceptionResNetV2. For Training they use the Flickr8k dataset. The result is that an Inceptionv3 Network performes best as Encoder. They compare the results with BLEU Score, similar to our work, but instead of Flickr dataset, we use prepared COCO dataset with our image class for training and in addition we train different RNN Networks (LSTM and GRU) as Decoder.

There are several metrics which can be used to evaluate automatically generated image descriptions. Common used Metrics are BLEU and METEOR, each metric has well known benefits and blind spots. We have to measure correlation to human judgements and evaluation of the syntax in the generated sentences. To adress these two challenges Cui, Yang, Veit, Huang and Belongie [14] propose a novel learning based discriminative evaluation metric, that is directly trained to distinguish between human and machine-generated captions. They conclude, that the metric could be an effective complementary to the existing rule-based metrics.

III. Semantic Image Segmentation with Visual Attention Networks

Visual Attention Networks are used to address the Problem of generating image descriptions in the field of full scene understanding. It's not only necessary to predict the objects shown on an image. Furthermore, the model should be able to capture the relationships between different objects on an image to convert them into a natural language from large sets of data [12].

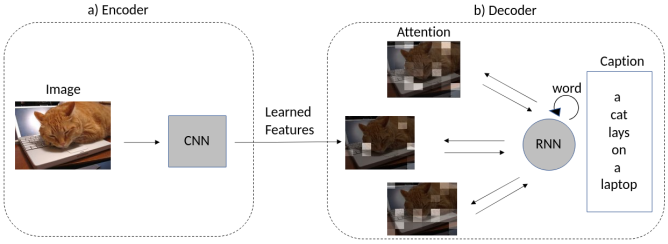


Figure 1. Visual Attention Network Architecture [15]

A VAT consists of an Encoder a) - Decoder b) Architecture as shown in Figure 1. As Encoder a CNN is used to extract the image features into a vectorial representation and a Recurrent Neural Network is used to generate appropriate descriptions for the extracted image features [16][12]. The RNN has a attention layer inside which is based on the functionality of a human visual system. Instead of processing the scene as a whole, the attention is focused on different parts of an image.

IV. Dataset

Reminiscence sessions consist of content that reflects the biographical content of a person with dementia. Media that correspond to the biographical events of a person with dementia can trigger memories which evoke activation's in the patient. These biographical contents are also called life themes and refer to any area of life. Possible examples of life

Copyright (c) IARIA, 2023. ISBN: 978-1-61208-950-8

themes are: Animals, travel, nature, religion, childhood and youth or occupation. The life themes can be represented by different types of media or combinations of them. Examples are pictures, music or videos. The media that correspond to the life themes do not necessarily evoke positive memories in the person with dementia. Therefore, it is important to know negative memories and fears of a patient, so that media content which trigger these emotions can be sorted out for a reminiscence session.

To match the life themes of dementia patients, our dataset contain classes with objects from everyday life, like the classes in MS COCO dataset. The MS COCO dataset [6] has many classes in common with the dementia patients life themes, so this dataset is used for training. For image captioning, each image from the COCO dataset is described with 5 different sentences.

This work primarily targets the description of dog and cat images, so these categories are filtered from the dataset. Dog images have two categories in the area of Reminiscence Therapy: Dog images that activate positive memories in the PWD and dog images that might trigger negative memories. For example dangerous looking dogs, aggressive dogs or snarling dogs are objects that can evoke negative memories. In total, 4114 images of the category "cat" and 4385 images of the category "dog", with a total of 42495 image descriptions, are filtered from the dataset. COCO only contains images with friendly looking dogs but we need angry, fear producing dogs, too. So, the dataset is extended with the category "angry dogs" and filled with new image content, from free image databases. Similar to the caption style of the MS COCO dataset, we labeled each image with 5 description sentences. Every sentence is natural language formulated and contains one or more of the following information: A dog shows or bares his teeth, Number of dogs in the picture, Color of the dogs, Background color, Meadow in background or other objects and toys on the picture.

The number of images in this category amounts to 360 training images with 1800 descriptions. In total, our dataset consists of 8859 images, with 44295 image descriptions. For training, we use a random $80/20$ split on the dataset, to split it into train and validation set.

V. Evaluation and Comparison

This section describes the technologies and parameters used for training. Furthermore, the results are presented, evaluated and compared.

A. Training

For training of the networks, we use the described dataset. The networks are all used with weights pre-trained on ImageNet dataset. We use a fixed length image caption of 9 words per sentence, because performance is poor on long input or output sequences. The Training Vocabulary consists of all words that occur more than tree times in the vocabulary, In total there are 6660 words in the training Vocabulary. Unknown words are provided with the token $< unk >$. The Networks are trained with 100 Epochs.

B. Encoder and Decoder

The use of a CNN as Encoder in a Visual Attention Network Architecture was successfully evaluated by recent works [17][18]. We evaluate different CNN implementations for image feature extraction to determine which is the best in the area of image caption generation for reminiscence sessions. The networks compared and evaluated are: Inceptionv3 [19], ResNet101 [20], VGG16/19 [21] and Xception [22] as Encoder. As Decoder LSTM [23] and GRU [24] networks are used.

C. Metrics for image captioning evaluation

For a formal comparison of the captions generated by the VATs, different metrics are used.

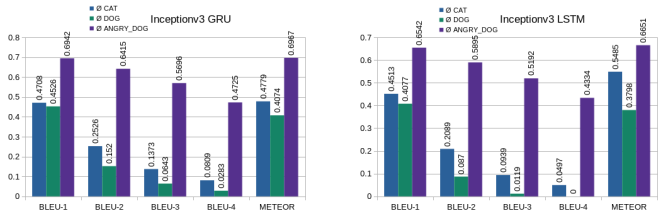
BLEU Score is a metric that can be used to automatically evaluate machine-generated image captures. BLEU is fast, inexpensive and language independent. The metric correlates strongly with the reference captions, as the caption length, word choice and word order are used to calculate the BLEU Score [25].

METEOR is a metric for formal and automatic evaluation of machine-generated captions, too. The metric measures not only precision (accuracy of the match), but also recall (completeness of the match), unlike the BLEU Metric. In this metric, word agreement is not determined using n-grams, but using unigrams, which are grouped into as few chunks as possible, where a chunk is defined as a set of unigrams that are adjacent in the hypothesis and the reference. The metric solve some problems of the BLEU Metric. BLEU measures correlation at the corpus level and METEOR also measures correlation with human judgment at the sentence or segment level [26].

D. Results

The results calculated by the metrics are represented in the following figures. They show respectively the results of the BLEU Score with different n-grams and the results of the METEOR metric. Figure $a)$ represents the results obtained by using the GRU implementation as decoder and $b)$ shows the results of LSTM network as decoder. For evaluation, a dataset, containing 10 images for each class (cat, dog, angry dog), was created. Each image is described with 5 captions per image as reference to calculate the metric scores.

Figure 2 shows the results of Inceptionv3 encoder.



(a) Results Inceptionv3 GRU

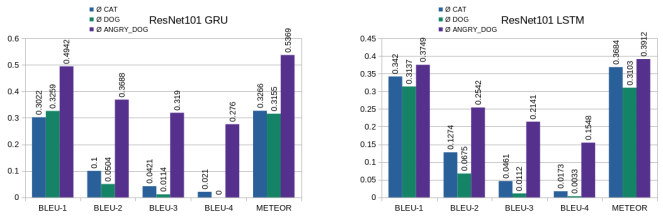
(b) Results Inceptionv3 LSTM

Figure 2. Results Inceptionv3

The results are a little better when using a GRU decoder as opposed to the results of the LSTM decoder, regardless of the metric. The class angry_dog performs best, while BLEU with the use of a 4-gram gives the worst results. However, depending on the class, the results are equally distributed for both the GRU and the LSTM network.

Figure 3 shows the results of the ResNet101 Encoder. This Encoder produces the worst results in comparison to all other Encoder - Decoder combinations. Whereby the GRU decoder provides better results, except for the dog class, calculated with the BLEU score using a 1-gram.

Class dog and angry_dog have nearly the same METEOR value with LSTM Decoder. This phenomenon does not occur with any other network.



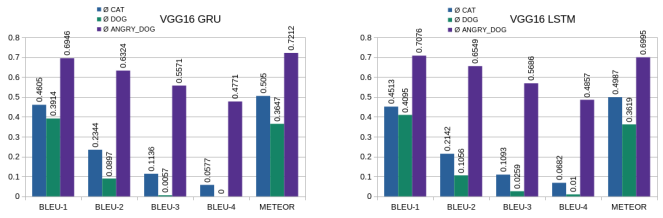
(a) Results ResNet101 GRU (b) Results ResNet101 LSTM

Figure 3. Results ResNet101

The evaluation results of the trained VGG16 Network are shown in Figure 4. With this network architecture, the results are still stable with different decoders. The values of the metrics hardly differ when using the GRU decoder compared to the use of the LSTM encoder.

The results of VGG16 are generally better than those of Inceptionv3 and ResNet101. But in contrast to VGG19 and the Xception rather worse, except for the angry_dog class, trained with LSTM decoder.

This class performs best among all network architectures. There is also an exception in the cat class, which was calculated with the BLEU score and a 4-gram. In this class, the network also performs better than all others used for this evaluation.



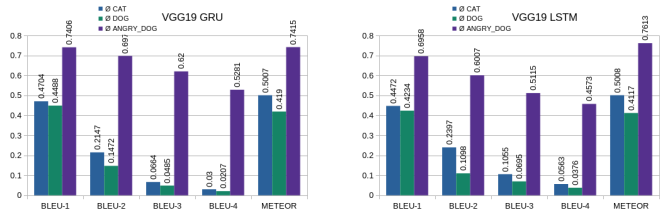
(a) Results VGG16 GRU (b) Results VGG16 LSTM

Figure 4. Results VGG16

Figure 5 shows the results of VGG19 Encoder. For our use case, the deeper VGG network with 19 layers performed better than the VGG network with 16 layers. VGG16 performs better for only one class (which may be due to the natural variance of an RNN network), VGG19 performs better in multiple classes and by using GRU and LSTM decoders.Co

In contrast to all other architectures, BLEU-4 and METEOR give the best results in the angry_dog class, using a GRU decoder. By using a LSTM decoder the values vary and the class Dog calculated by BLEU with a 4-gram performs best, as well as the class angry_dog by calculating the METEOR value.

Both VGG networks have the property that they provide stable results no matter which decoder is used, the results do not vary much. Furthermore, the result values are equally distributed, with all data classes.



(a) Results VGG19 GRU (b) Results VGG19 LSTM

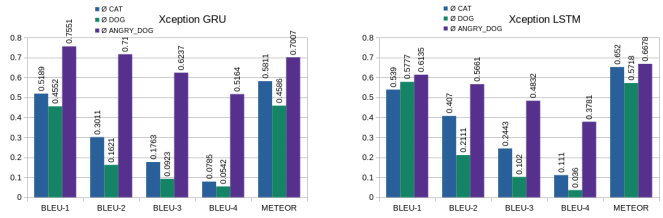
Figure 5. Results VGG19

The last network is the Xception Network. The results are shown in Figure 6. The values of this network are the best, no matter which decoder is used. Both GRU and LSTM results are better than the values from the other networks.

The only exception is within the angry_dog class and LSTM decoder. It does not outperform the VGG16 network by using BLEU score and METEOR metric for result evaluation.

The Xception network is based on the architecture of the Inception network. However, the inception modules are replaced by depth wise separable convolutions followed by point wise convolutions. The number of parameters is the same in both networks, but the Xception network uses the parameters in a more efficient way than the Inception network [22].

As our results show, the extended Inception architecture is successful, since the results are better than those of the Inception network.



(a) Results Xception GRU (b) Results Xception LSTM

Figure 6. Results Xception

For all networks the caption generation for the class ”angry_dog” works best. This is because we created and labeled

the class ourselves. So, the labels used for training fit our use case better than the labels from COCO dataset.

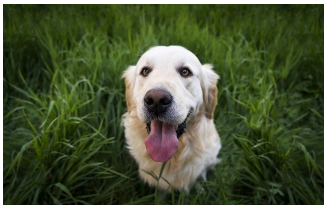
Generally, the results get worse with increasing n-gram value and the METEOR results are quite uniform due to the use of the chunks.

E. Human Evaluation

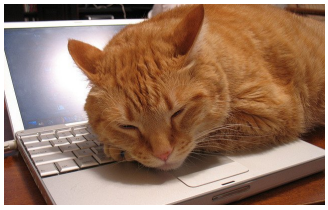
Complementing the evaluation using metrics, we conducted a human evaluation of the results to determine whether the results are good enough to make the trained weights applicable in practice for people with dementia.

In humans, the formal translations exist only in their mind. So, people's pattern translations are preverbal representations, and can be realized with several synonymous expressions when they are verbalized. Therefore, human evaluators may equally evaluate different translation variants as "correct," although their evaluations might differ depending on their emphases [27]. That means, even results with a low BLEU or METEOR value could be rich picture descriptions, since many words and sentence positions mean similar things, even if they are formally different from each other.

We picked one image from each of the ten images that are in a test class and compared the result captures. For this purpose, we collected the best and the worst results generated by the network. The best results are shown in Table I and the worst in Table II. The results refer to Figure 7, where a), b) and c) are respectively selected images from the test classes that were present in the test dataset. The results must refer to the objective opinion of a person and not to a technical evaluation.



(a) Evaluation dog picture



(b) Evaluation cat picture

(c) Evaluation angry dog picture

Figure 7. Example evaluation pictures

The best results based on human evaluation were obtained by the VGG19 and Xception networks. The objects contained in the images were well recognized and correctly described by the networks. Also syntactic criteria are mostly fulfilled by the networks in a sufficient quality. It was observed that the results of the Xception network are more "stable" than the results of the VGG19 network, regardless of which decoder is used. More stable means that the number of the same caption is higher than the number of the same caption generated by other networks, because captions of a RNN can vary, since such a model does not have a fixed number of hidden layers.

TABLE I. HUMAN EVALUATION BEST RESULTS

Picture	Encoder	Caption
a)	Inceptionv3	a white dog wearing a green and holds a
	ResNet101	a dog has a red collar is standing near
	VGG16	a black and white dog
	VGG19	a dog laying in a grassy field looking grass
	Xception	a dog is sitting
b)	Inceptionv3	a cat is curled up asleep lying on a
	ResNet101	a cat is using a laptop keyboard
	VGG16	cat laying on a laptop computer
	VGG19	a orange cat sits on a laptop
	Xception	a cat is laying down on a laptop
c)	Inceptionv3	a black and brown dog looks angry while baring
	ResNet101	a black and brown dog baring his teeth on
	VGG16	an angry looking black and brown dog shows his
	VGG19	an angry looking black and brown dog shows his
	Xception	An angry black and brown dog baring his teeth

TABLE II. HUMAN EVALUATION WORST RESULTS

Picture	Encoder	Caption
a)	Inceptionv3	a cute hair that its mouth resting while wearing
	ResNet101	a dog <unk> in a pink flower and green
	VGG16	a big panting dog
	VGG19	the dog is looking to above his mouth
	Xception	a golden puppy leash standing near some frisbee
b)	Inceptionv3	an orange cat sits resting its head on the
	ResNet101	a cat sleeping half lake in an open suitcase
	VGG16	an orange cat resting it's camera
	VGG19	a cat sleeping half on
	Xception	a close up of a cat sleeping half on
c)	Inceptionv3	a black and brown dog shows his teeth on
	ResNet101	a cat is greeting each other in a chair
	VGG16	a brown dog baring his teeth on green grass
	VGG19	a black and brown dog looks angry while baring
	Xception	an angry looking black and brown dog shows his

The worst results are obtained by ResNet101, the results vary strongly among each other and the network generates many "outlier" captions, which do not fit in any way to the image content that should be described. For example, the caption "a dog has a hat on the beach" or "a small cat is sitting on the ground" are outlier result generated by caption generation with respect to Figure 7 b).

VI. Conclusion and Future Work

This work takes up the basic functionality of Visual Attention Networks and presents their use based on different network configurations intending to automatically describing images in such a way that they can be assigned to the life themes of dementia patients. In this way, reminiscence sessions can be automatically created with biography-related content. By automatically compiling sessions, caregivers are relieved and can invest more time with the patients than in creating the reminiscence session content.

Copyright (c) IARIA, 2023. ISBN: 978-1-61208-950-8

We have figured out which network architecture performs best. For comparison we used the encoder networks Inceptionv3, ResNet101, VGG16/19 and Xception (image feature extraction) in combination with the decoder networks LSTM and GRU (caption generation). For training, we created a dataset specifically suited to dementia patients, which is composed of some classes from COCO dataset and a separate class. By evaluating the networks trained using our dataset, we found that the Xception network in combination with a GRU network produced the best results. Both are evaluated formally and by human.

In the future, the system can be extended with other digital media types, like music or videos. The dataset we use only covers the life theme "animals". To make the reminiscence sessions more valuable, other life themes should be included by extending the training dataset. From a technical point of view, the VAT could be further adjusted by hyperparameter tuning to improve the results and to reduce the number of outlier captions.

REFERENCES

[1] F. Meiland, A. Innes, G. Mountain, L. Robinson, H. van der Roest, A. García-Casal, D. Gove, J. R. Thyrian, S. Evans, R.-M. Dröes, F. Kelly, A. Kurz, D. Casey, D. Szczesniak, T. Dening, M. Craven, M. Span, H. Felzmann, M. Tsolaki, and M. Franco, "Technologies to support community-dwelling persons with dementia: A position paper on issues regarding development, usability, effectiveness and cost-effectiveness, deployment, and ethics," JMIR Rehabil Assist Technol, vol. 4, 01 2017, p. e1.

[2] A. A. Khait and J. Shellman, "Uses of Reminiscence in Dementia Care," Innovation in Aging, vol. 4, no. Supplement_1, 12 2020, pp. 287–287.

[3] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. A. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019, pp. 11 135–11 145.

[4] Z. Ji, W. Yao, H. Pi, W. Lu, J. He, and H. Wang, "A survey of personalised image retrieval and recommendation," in Theoretical Computer Science - 35th National Conference, NCTCS 2017, Wuhan, China, October 14-15, 2017, Proceedings, ser. Communications in Computer and Information Science, vol. 768. Springer, 2017, pp. 233–247.

[5] A. K. Jaiswal, H. Liu, and I. Frommholz, "Effects of foraging in personalized content-based image recommendation," CoRR, vol. abs/1907.00483, 2019.

[6] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft COCO: Common Objects in Context," CoRR, vol. abs/1405.0312, 2014.

[7] N. Alm, A. Astell, M. Ellis, R. Dye, G. Gowans, and J. Campbell, "A cognitive prosthesis and communication support for people with dementia," Neuropsychological Rehabilitation, vol. 14, no. 1-2, 2004, pp. 117–134.

[8] G. Gowans, R. Dye, N. Alm, P. Vaughan, A. Astell, and M. Ellis, "Designing the interface between dementia patients, caregivers and computer-based intervention," The Design Journal, vol. 10, Mar 2007, pp. 12–23.

[9] A. Astell, M. Ellis, L. Bernardi, N. Alm, R. Dye, G. Gowans, and J. Campbell, "Using a touch screen computer to support relationships between people with dementia and caregivers," Interacting with Computers, vol. 22, 07 2010, pp. 267–275.

[10] A. Astell, S. Smith, S. Potter, and E. Preston-Jones, "Computer interactive reminiscence and conversation aid groups—delivering cognitive stimulation with technology," Alzheimer's & Dementia: Translational Research & Clinical Interventions, vol. 4, no. 1, 2018, pp. 481–487.

[11] P. Klein and M. Uhlig, "Interactive memories: Technology-aided reminiscence therapy for people with dementia," in Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments, ser. PETRA '16. New York, NY, USA: Association for Computing Machinery, 2016, pp. 1–2.

[12] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in Proceedings of the 32nd International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 2048–2057.

[13] A. Pal, S. Kar, A. Taneja, and V. Jadoun, "Image captioning and comparison of different encoders," Journal of Physics: Conference Series, vol. 1478, 04 2020, p. 012004.

[14] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, "Learning to evaluate image captioning," pp. 5804–5812, 2018.

[15] L.-M. Meßmer and C. Reich, "Potentials of semantic image segmentation using visual attention networks for people with dementia," pp. 234–252, 2021.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," CoRR, vol. abs/1411.4555, 06 2014, pp. 3156–3164.

[17] J. Aneja, A. Deshpande, and A. Schwing, "Convolutional image captioning," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jun 2018, pp. 5561–5570.

[18] S. Katiyar and S. K. Borgohain, "Analysis of convolutional decoder for image caption generation," CoRR, vol. abs/2103.04914, 2021.

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," pp. 2818–2826, 2016.

[20] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

[21] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015.

[22] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Los Alamitos, CA, USA: IEEE Computer Society, jul 2017, pp. 1800–1807.

[23] M. Phi, "Illustrated Guide to Recurrent Neural Networks - Towards Data Science," Medium, Jun 2020, [Retrieved: Jan, 2022]. [Online]. Available: https://towardsdatascience.com/illustrated-guide-to-recurrent-neural-networks-79e5eb8049c9

[24] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, Mar 2020, p. 132306.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, jul 2002, pp. 311–318.

[26] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Association for Computational Linguistics, jun 2005, pp. 65–72.

[27] H.-Y. Chung, "Automatische Evaluation der Humanübersetzung: BLEU vs. METEOR," Lebende Sprachen, vol. 65, no. 1, Apr 2020, pp. 181–205.

Copyright (c) IARIA, 2023. ISBN: 978-1-61208-950-8