Potentials of Semantic Image Segmentation Using Visual Attention Networks for People with Dementia

Liane Meßmer and Christoph Reich

Hochschule Furtwangen – University of Applied Sciences
l.messmer@hs-furtwangen.de
christoph.reich@hs-furtwangen.de

Abstract. Due to the increasing number of dementia patients, it is time to include the care sector in digitization as well. Digital media, for example, can be used on tablets in memory care and have considerable potential for reminiscence therapy for people with dementia. The time consuming assembly of digital media content has to be automated for the caretakers.
This work analyzes the potentials of semantic image segmentation with Visual Attention Networks for reminiscence therapy sessions. These approaches enable the selection of digital images to satisfy the patients individual experience and biographically. A detailed comparison of various Visual Attention Networks evaluated by the BLEU score is shown. The most promising networks for semantic image segmentation are VGG16 and VGG19.

Keywords: Dementia, Alzheimer, Visual Attention Network, CNN, RNN, LSTM, GRU, Inceptionv3, VGG, ResNet, Semantic Segmentation, Natural Language Processing, Health Care, Reminiscence Therapy, Memory Triggering

1 Introduction

The German Federal Statistical Office examines the population development until the year 2060. They conclude that the percentage of people older than 67 years will increase from 19% in 2018 to 27% in 2060 [1]. As people live longer than ever before and therefore, the number of dementia patients is also increasing. Today, more than 44 Million people worldwide are living with dementia, 1.5 Million of them live in Germany [2]. Non-pharmacological methods are effective in improving the lives of dementia patients. Reminiscence therapy belongs to the non-pharmacological techniques, used to address the activation process of people with Dementia (pwD) [3].

Nowadays, Reminiscence therapy sessions use digital support systems, which consists of digital content such as images, movies or music. These can be applied for example on mobile devices. Images, movies or music is used for life review and to evoke memories in patients [4], [5]. Thus, demand-oriented and technical solutions cause a valuable contribution to the care of pwD. Their potential is far from exhausted. Yet, a well known challenge is the identification of suitable content as well as the design and evaluation of high-quality reminiscence care services are very labor-intensive. Besides, this task places high demands on the caretakers qualifications.

Currently, the reminiscence session content must be identified and evaluated by caretakers, because the content of a reminiscence session should be suitable and individual for pwD. So, in practice, a very limited pool of standard content is often used.

To support the reminiscence therapies the caretaker would select images according to the following image characteristics such as objects, colors, shapes, number of objects or

meaning according to life themes [6]. In particular, automated, individual and biography-related media selection improves the quality of the reminiscence session and reduces the workload of the caretakers by shortening the preparation time. However, this potential of automation relieves the caretaker in terms of reminiscence sessions. It also gives care giving relatives the opportunity to include memory-triggering content in their care.

The approach of this paper is semantic image segmentation, to extract the features of an image. Semantic segmentation is one of the high-level tasks that pave the way to full understanding of a scene [7]. The importance of full scene understanding as a core computer vision problem affirms by the fact, that an increasing number of applications using the knowledge derivation from images. Some of these applications include medical assistance systems or human-computer interactions. With the popularity of Deep Learning, many semantic segmentation problems are addressed with Deep Learning approaches. They far exceed other methods in terms of accuracy and efficiency.

The goal of this work is to analyzes the potentials of semantic image segmentation using different Convolutional Neural Networks (Inceptionv3, VGG16, VGG19 and ResNet101) in combination with a Recurrent Neural Network (LSTM, GRU) related to people with dementia, to generate automatic descriptions from images (also called image captioning).

This work consists of 8 chapters. Chapter 2 deals with the related work. Life themes identified to activate pwD are described chapter 3. In Chapter 4, the concepts of visual attention networks are presented and explained. The data used for training, as well as the training process, are described in Chapters 5 and 6. Finally, the results are presented in Chapter 7 and Chapter 8 describes the conclusion and future work.

2 Related Work

2.1 Reminiscence Aid Systems

The Computer Interactive Reminiscence and Conversation Aid (CIRCA) project, proposed by Astell et al. [8] was the first project that developed an application for digital reminiscence therapy to support people with dementia. Over the years, it was supplemented by different new technologies, like a specific interface for the interaction with the system [9] or a touch screen computer to enable an easier interaction with the system [10]. Today, CIRCA is an interactive multimedia application. The latest publication of the project "Computer Interactive Reminiscence and Conversation Aid groups - Delivering cognitive stimulation with technology" demonstrates the effectiveness of CIRCA for group interventions [11]. The growing process of the project shows the effectiveness of digital assistance systems in the area of reminiscence therapies. The difference with our work is that we do not want to use random content for a reminiscence session, but individual content that fits the biography of a patient. Therefore, our system should be able to describe images, that match the life themes of pwD, automatically.

Carós et al. [12] presented in their work "Automatic Reminiscence Therapy for Dementia" a solution approach to automate reminiscence therapy, which uses artificial intelligence based systems. Their system is called "Elisabot" and consists of a system that uses personal images of users and generates questions about their lives, using Visual Attention Networks (VATs) In our work we compare different architectures of Visual Attention Networks to find the potentials of these Networks and whether their generated image captions fit to pwDs needs. All encoder models are pretrained on the ImageNet dataset.

In the work "Image Captioning and Comparison of Different Encoders" by Pal et al. [13] a comparison is presented of different encoder implementations as they are used in a Visual Attention Network, for the automatic generation of image captions. As encoder, they compare different convolutional neural networks, these are Inception_v3, VGG16, VGG19 and InceptionResNetV2. The result is that the Inceptionv3 Encoder works best. Since the range of the BLEU score is lowest for the model. Similar to our work, they use BLEU score for result evaluation. We use the MS COCO dataset for training and instead of using the InceptionResNetV2 for model comparison, a ResNet-101 is used. Furthermore, this work additionally presents differences between Long-Short-Term-Memory (LSTM) and Gated Recurrent Unit (GRU) decoder .

2.2 Visual Attention Networks

The caption generation in this paper is based on the work "Show, Attend and Tell" proposed by Xu et al. [14]. They describe a mechanism that generates image captions based on Convolutional Neural Network (CNN) encoder and Recurrent Neural Network (RNN) decoder by using an attention layer in the network. The Encoder extracts specific features from an image and generates a set of feature vectors, which were referred to as annotation vectors in this context. The attention layer takes an annotation vector, to focus on a specific part of an image, because every vector is a representation corresponding to a part of an image. This allows the decoder to selectively focus on specific parts of an image by selecting a subset of all annotation vectors [15]. The results are automatic generated, textual image captions in a natural language.

Encoder Recent works have represented the successful deployment of Convolutional Encoder. Therefore, our work is focused on image caption generation using a CNN as Encoder [16], [17]. There are different CNN implementations for image feature extraction, which are compared in this work: VGG16, VGG19, ResNet-101 and Inceptionv3.

Simonyan et al. [18] proposed multiple Versions of VGG Networks in their work "Very Deep Convolutional Networks for Large-Scale Image Recognition". They differ mainly in their depth, i.e. in their number of layers. As the name suggests, one network has 16 layers and the other 19.

In the work "Deep Residual Learning for Image Recognition", the Residual Network (ResNet) Architecture is described from He et al. [19]. The special feature of a ResNet is that each layer in the network consists of several blocks. As the depth of the network increases, the number of operations in a block increases too, but not the total number of layers. Thus, ResNets solve the problem of vanishing gradients as neural networks become deeper.

The architecture of the Inceptionv3 network emerges from the GoogleNet architecture and was proposed from Szegdy et al. [20] in the work "Rethinking the Inception Architecture for Computer Vision". The model is a combination of many ideas that have emerged in recent years. The network consists of several symmetric and asymmetric blocks, which can contain different types of layers. For example, convolutions, average pooling, max pooling, etc. In total, the network has 42 layers [21].

Decoder As decoder, a Recurrent Neural Network (RNN) is used. The first RNN was published in the work "Finding structure in time" from Jeffrey Elman [22]. This network is capable of reading the annotation vectors extracted by the CNN in previous steps. The important features of RNNs are the memory cells. A normal feed forward network has an input, hidden and output layer. The RNN loops the hidden layer to process sequential

data. With this looping mechanism, the RNN allows flowing information from one step to the next step. During training a RNN Model is enrolled (each word acts as layer) and trained with backpropagation trough stochastic gradient descent [23]. RNNs often have a problem known as short term memory. As the number of words increases, so does the depth of the network. The more steps an RNN has to process, the greater the problem of retaining the information from the previous steps. This phenomenon occurs due to the backpropagation used for training the network.

There are two approaches to tackle the problems of RNNs caused by short-Term-Memory: Long-Short-Term-Memory (LSTM) and Gated-Recurrent Unit (GRU). The LSTM Network was proposed from Hochreiter et al. [24] in the work "Long Short-Term Memory" and the GRU Network was proposed from Cho et al. [25] in the work "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". The underlaying structure of these networks is the same as the structure from an RNN extended by a mechanism that can learn long-term dependecies using "gates". LSTMs are using three gates to process the data while GRUs only use one gate for data selection [26].

3 People with Dementia Life Themes

The reminiscence sessions should include content that matches the biographical needs of a pwD. So that an activation in reminiscence sessions can take place [27]. Life themes represent a generalized categorization of life stages or events that play a role in a person's life. The goal of these particular subjects is to evoke memories in pwD, that are associated with a life theme. Life themes can be represented by pictures, videos, music or physical objects. This work focuses on picture retrieval. The following table 1 shows the life themes we identified and used for image feature extraction in this work. As the table shows, there are not only positive activation for pwD, but also negative/fearful issues, which should be avoided. If these negative loaded images are also labeled, they can be discarded before using them in a reminiscence session. Anxiety-producing image content should be screened out before sessions. This prevents a session from having a negative impact on a patient's well-being.

Table 1. Life themes used in reminiscence sessions for pwD

Life Themes		
Travel	Animals	Professions
Hobbies and Activities	Preferences and Habits	Exterior Appearance
Religion	Education	Nature
Childhood and Youth	Home	Tradition and Culture
Literature	Media	Theather
Garden	People	Fears and Disease
Food		

4 Visual Attention Networks

Automatic generation of image descriptions is a difficult task in the field of full scene understanding. The model must transform large disparate sets of data into a natural language. To address this problem, Visual Attention Networks (VATs) are used [14].

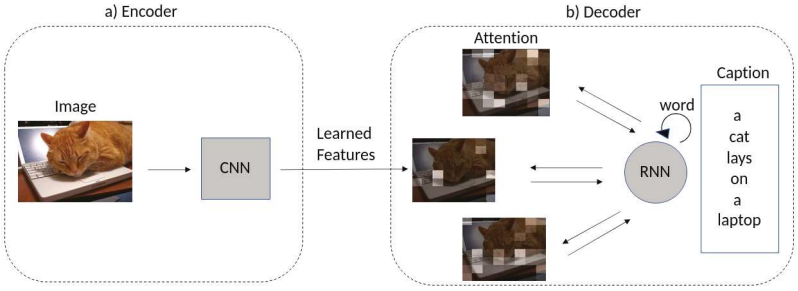

Fig. 1. Visual Attention Network Architecture

A VAT consists of two sub-models, as shown in Figure 1. a) Encoder: First, an image is given as input to the CNN to obtain vectorial representation of input images, on the last hidden layer of the network. b) Decoder: The feature vector is used from the RNN decoder as input to generate sentences that contain the objects and the relationships between them [14],[28]. The visual system of a human has the function to pay attention on different parts of an image instead of processing the scene as a whole [29]. Based on this human attention mechanism, an attention layer is integrated into a visual attention network. In this paper we use soft attention mechanisms for training [30]. The areas selected by attention mechanism are captured by the RNN for further processing. Based on this process suitable textual descriptions are generated.

5 Dataset Used for Training and Evaluation

The dataset used for training should contain everyday objects, which match the life themes from section 3. There are several datasets that contain labeled content for image captioning, for example MS COCO, Pascal VOC or Flickr30k. Since the MS COCO dataset [31] has the most objects in common with the life themes described in section 1, this dataset was used for training in this work. The matching life themes are for example “Animal”, “Person” and “Food” . Each image from the MS COCO Caption dataset is described (labeled) with 5 different sentences. This work primarily targets the description of dogs and cats images, so these categories are filtered from the dataset. Dog images fall into two categories in the context 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. In total, 4298 images of the category “cat” and 4562 images from the categorie “dog”, with a total of 43200 image descriptions, are filtered from the dataset.

Since the MS COCO dataset only contains images with friendly looking dogs, the dataset is extended with the category “Angry Dogs” and filled with our own image

content. Each image is described with 5 description sentences, similar to the the caption style of MS COCO dataset. The number of images in this category amounts to 360 training images with 1800 descriptions. In total, our dataset consists of 9000 images, with 45000 image descriptions. For training, we use a random 80/20 split on the dataset, to split it into train and validation set.

6 Training

This work compares different networks which are used in a Visual Attention Network Architecture. We use ResNet-101, VGG16, VGG19 and Inceptionv3 as encoder and Gated Recurrent Unit (GRU) or Long-Short-Term Memory (LSTM) as decoder, to compare them with each other. The CNNs are all pretrained on the ImageNet [32] dataset. This dataset contains many objects related to the Life themes of people with dementia. Additional we use MS COCO Dataset for training the RNN, as described above.

For training, we use a fixed-length image caption of 9 words per sentence, because the network performes poor on long input or output sequences. This is caused by rolling out the RNN, where each word represents a hidden layer. The longer a caption is, the slower the network is trained. In addition, for our use case of image descriptions for the use in reminiscence therapy, captions of length 9 are sufficient to make sense of the content and at the same time short enough to not impair the performance of the model.

In total, we get 6660 different words as vocabulary, from which we use all words that occur more than three times in the vocabulary. This results in 2206 words. Unknown words, are provided with the token <unk>. We trained the Networks with a batch size of 8 and 100 Epochs.

7 Evaluation and Comparison of the VGG16, VGG19, ResNet-101 and Inceptionv3 Model

For evaluation and comparison, a dataset, containing 10 images for each class in the training dataset (cat, dog, angry dog), was created. These images are with 5 captions per image for result evaluation with BLEU score. In the area of image caption generation it's hard to evaluate resulting captions formally. We only have five reference sentences for an image, but there are much more image descriptions that match the image content. Thus, we decided to evaluate the results formally with BLEU score and verify the results with human evaluation. BLEU is a method for automatic evaluation of machine translation [33], it's quickly, inexpensive, and language-independent. The metric correlates highly with human evaluation, because it measures the closeness of the machine translation to human reference translations by taking translation length, word choice and word order into consideration. For our evaluation we use 2-gram BLEU score. This approach does not check the specific order of all words in the generated caption, but only the adjacent ones and compares them with the reference descriptions.

The following Table 2 shows the average calculated BLEU score, using different encoder-decoder combinations, for (a) Inceptionv3, (b) ResNet101, (c) VGG16 and (d) VGG19. Every category contains 10 images for inference. The generated captions of an RNN are not identical, since such a model has no fixed hidden layer size. They always depend on the generated caption length. Therefore, each model was trained five times (GRU-1, GRU-2,...,GRU5 or LSTM-1, LSTM-2,...,LSTM-5). At the end of the tables, the respective average value (ø) is shown, which is used for result comparison. The scores

in the tables show that ResNet101 behaves the worst. The BLEU scores are below the scores of the other networks, regardless of which model combination was used. Inceptionv3 has the best average BLEU for the category "Cat", combined with a GRU model. The best results are produced by VGG16 and VGG19 models, with VGG19 performing best in combination with a GRU decoder and the VGG16 in combination with an LSTM decoder. The Inceptionv3 model outperforms the VGG16 model in the category "Cat" with a BLEU score of 0.02322 compared to 0.2005. The difference is so small that for all classes in general, the VGG16 model perfomes better.

All models provide the worst results for the category "Dog" and the best for the category "Angry Dog". This is because the images in the angry dog category were self-labeled, specifically tailored to our problem set. The other two categories are labeled with general captions from the COCO dataset. In general, the BLEU scores are stable for each model in each class; no outliers are apparent.

Table 2. Calculated BLEU scores

(a) BLEU score for Inceptionv3

Decoder	Cat	Dog	Angry Dog
GRU-1	0.2377	0.1018	0.6600
GRU-2	0.2629	0.1619	0.6423
GRU-3	0.2793	0.1650	0.6526
GRU-4	0.1701	0.1846	0.5806
GRU-5	0.2114	0.1467	0.6682
ø	**0.2322**	0.1250	0.6407
LSTM-1	0.2511	0.1126	0.5386
LSTM-2	0.1620	0.0702	0.5750
LSTM-3	0.2336	0.0967	0.5910
LSTM-4	0.1718	0.0845	0.6724
LSTM-5	0.1616	0.0759	0.5704
ø	0.1960	0.0879	0.5894

(b) BLEU score for ResNet101

Decoder	Cat	Dog	Angry Dog
GRU-1	0.0875	0.0500	0.3789
GRU-2	0.1889	0.0500	0.2951
GRU-3	0.0634	0.0375	0.3361
GRU-4	0.0375	0.0625	0.3020
GRU-5	0.1000	0.0625	0.3896
ø	0.0954	0.0525	0.3403
LSTM-1	0.1697	0.0611	0.2402
LSTM-2	0.0960	0.0500	0.2339
LSTM-3	0.1625	0.1000	0.2000
LSTM-4	0.1000	0.0723	0.2978
LSTM-5	0.0986	0.0611	0.2216
ø	0.1253	0.0689	0.2387

(c) BLEU score for VGG16

Decoder	Cat	Dog	Angry Dog
GRU-1	0.2154	0.0666	0.5346
GRU-2	0.1785	0.0767	0.5974
GRU-3	0.2236	0.1142	0.6947
GRU-4	0.2339	0.1077	0.5626
GRU-5	0.2378	0.0583	0.6252
ø	0.2178	0.0847	0.6029
LSTM-1	0.1918	0.0875	0.6417
LSTM-2	0.2430	0.0916	0.6165
LSTM-3	0.2085	0.1139	0.6167
LSTM-4	0.1805	0.1111	0.6032
LSTM-5	0.1791	0.0800	0.6990
ø	0.2005	**0.0968**	**0.6354**

(d) BLEU score for VGG19

Decoder	Cat	Dog	Angry Dog
GRU-1	0.1932	0.0951	0.6087
GRU-2	0.1883	0.1468	0.6545
GRU-3	0.2231	0.1571	0.6934
GRU-4	0.2261	0.1303	0.6989
GRU-5	0.1723	0.1105	0.7291
ø	0.2006	**0.1279**	**0.6769**
LSTM-1	0.1684	0.0382	0.5244
LSTM-2	0.2682	0.0454	0.4502
LSTM-3	0.2220	0.0737	0.6332
LSTM-4	0.2316	0.0722	0.5741
LSTM-5	0.1986	0.0737	0.6888
ø	**0.2177**	0.0606	0.5741

The results, calculated by BLEU score are verified manually. The best and worst BLEU scores for eachmodel are taken and the corresponding generated captions were checked. For each model, the best and worst results for all categories are shown in Figure 2. Figure (a) shows the generated captions for cats, (b) shows the captions for dogs and (c) the captions for angry dogs.

Best results:
Inceptionv3, GRU-3: a cat is curled up asleep lying on a
ResNet101, GRU-2: a cat is using a laptop keyboard
VGG16, LSTM-2: cat laying on a laptop computer
VGG19, GRU-4: a orange cat sits on a laptop

Worst results:
Inceptionv3, GRU-4: an orange cat sits resting its head on the
ResNet101, LSTM-2: a cat sleeping half lake in an open suitcase
VGG16, LSTM-5: an orange cat resting it's camera
VGG19, GRU-5: a cat sleeping half on

(a) Resulting Captions "Cat"[31]

Best results:
Inceptionv3, LSTM-1: a white dog wearing a green and holds a
ResNet101, GRU-5: a dog has a red collar is standing near
VGG16, LSTM-3: a black and white dog
VGG19, LSTM-5: a dog laying in a grassy field looking grass

Worst results:
Inceptionv3, GRU-1: 0.1018, a cute hair that its mouth resting while wearing
ResNet101, LSTM-2: 0.0500, a dog <unk> in a pink flower and green
VGG16, GRU-5: 0.0583, a big panting dog
VGG19, GRU-1: 0.6087, the dog is looking to above his mouth

(b) Resulting Captions "Dog"[34]

Best results:
Inceptionv3, GRU-5: a black and brown dog looks angry while baring
ResNet101, GRU-5: a black and brown dog baring his theeth on
VGG16, LSTM-1: an angry looking black and brown dog shows his
VGG19, GRU-5: an angry looking black and brown dog shows his

Worst results:
Inceptionv3, GRU-4: a black and brown dog shows his theeth on
ResNet101, GRU-2: a cat is greeting each other in a chair
VGG16, LSTM-4: a brown dog baring his theeth on green grass
VGG19, GRU-1: a black and brown dog looks angry while baring

(c) Resulting Captions "Angry Dog"[35]

Fig. 2. Resulting Captions

The resulting captions of the category dogs are as good as the BLEU score describes them. Only one bad caption was produced by the ResNet101, GRU-2. The other results coincide with the BLEU score, ResNet101 predictions are generally worse than the others and VGG16, VGG19 generate the best image captions. By comparing the BLEU scores with human evaluation, we came to the conclusion: the better the BLEU score, the better the caption.

8 Conclusion

This work reveals that the use of Visual Attention Networks in the context of reminiscence sessions for dementia patients has significant potential. The result of the comparison from different encoder-decoder combinations is that the use of VGG16-LSTM and VGG19-GRU Models generating promising results. This approach allows activation sessions to be simpler, faster and tailored for a patient's needs. Thereby, higher quality and quantity of reminiscence sessions is created. At the same time, the life of a dementia patient is positively influenced.

In the future, it's important to extend the dataset with more categories, to match more life themes from pwD. In addition, it's possible to extend the system with the ability to describe not only images automatically. There are also music and videos that match the life themes of pwD.

References

1. Bundesamt, S.: 14. koordinierte Bevölkerungsvorausberechnung - Ergebnisse für Deutschland (Jul 2021) [Online; accessed 14. Jul. 2021].
2. Association, A.: Alzheimer's & Dementia Help | Germany | Alzheimer's Association (Jul 2021) [Online; accessed 14. Jul. 2021].
3. Khait, A.A., Shellman, J.: Uses of Reminiscence in Dementia Care. Innovation in Aging **4**(Supplement_1) (12 2020) 287–287
4. Lazar, A., Thompson, H., Demiris, G.: A systematic review of the use of technology for reminiscence therapy. Health Education & Behavior **41**(1_suppl) (2014) 51S–61S PMID: 25274711.
5. Ĺatha, K., Bhandary, P., Tejaswini, S., Sahana, M.: Reminiscence therapy: An overview. Middle East Journal of Age and Ageing **11** (2014)
6. Ji, Z., Yao, W., Pi, H., Lu, W., He, J., Wang, H.: A Survey of Personalised Image Retrieval and Recommendation. In Du, D., Li, L., Zhu, E., He, K., eds.: Theoretical Computer Science, Singapore, Springer Singapore (2017) 233–247
7. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs (2016)
8. Alm, N., Astell, A., Ellis, M., Dye, R., Gowans, G., Campbell, J.: A cognitive prosthesis and communication support for people with dementia. Neuropsychological Rehabilitation **14**(1-2) (2004) 117–134
9. Gowans, G., Dye, R., Alm, N., Vaughan, P., Astell, A., Ellis, M.: Designing the interface between dementia patients, caregivers and computer-based intervention. The Design Journal **10**(1) (2007) 12–23
10. Astell, A.J., Ellis, M.P., Bernardi, L., Alm, N., Dye, R., Gowans, G., Campbell, J.: Using a touch screen computer to support relationships between people with dementia and caregivers. Interacting with Computers **22**(4) (07 2010) 267–275
11. Astell, A.J., Smith, S.K., Potter, S., Preston-Jones, E.: Computer interactive reminiscence and conversation aid groups—delivering cognitive stimulation with technology. Alzheimer's & Dementia: Translational Research & Clinical Interventions **4**(1) (2018) 481–487
12. Caros, M., Garolera, M., Radeva, P., i Nieto, X.G.: Automatic reminiscence therapy for dementia (2021)
13. Pal, A., Kar, S., Taneja, A., Jadoun, V.K.: Image captioning and comparison of different encoders. Journal of Physics: Conference Series **1478** (apr 2020) 012004
14. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. CoRR **abs/1502.03044** (2015)
15. Sarkar, S.: Image Captioning using Attention Mechanism - The Startup - Medium. Medium (Jun 2021)

16. Aneja, J., Deshpande, A., Schwing, A.: Convolutional image captioning (2017)
17. Katiyar, S., Borgohain, S.K.: Analysis of convolutional decoder for image caption generation (2021)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition (2015)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition (2015)
20. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision (2015)
21. Maeda-Gutiérrez, V., Galván-Tejada, C.E., Zanella-Calzada, L.A., Celaya-Padilla, J.M., Galván-Tejada, J.I., Gamboa-Rosales, H., Luna-García, H., Magallanes-Quintanar, R., Guerrero Méndez, C.A., Olvera-Olvera, C.A.: Comparison of convolutional neural network architectures for classification of tomato plant diseases. Applied Sciences **10**(4) (2020)
22. Elman, J.L.: Finding structure in time. Cognitive Science **14**(2) (1990) 179–211
23. Sherstinsky, A.: Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. Physica D: Nonlinear Phenomena **404** (Mar 2020) 132306
24. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8) (11 1997) 1735–1780
25. Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation (2014)
26. Gao, Y., Glowacka, D.: Deep gate recurrent neural network. In Durrant, R.J., Kim, K.E., eds.: Proceedings of The 8th Asian Conference on Machine Learning. Volume 63 of Proceedings of Machine Learning Research., The University of Waikato, Hamilton, New Zealand, PMLR (16–18 Nov 2016) 350–365
27. Huber, S., Berner, R., Uhlig, M., Klein, P., Hurtienne, J.: Tangible objects for reminiscing in dementia care. In: Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction. TEI '19, New York, NY, USA, Association for Computing Machinery (2019) 15–24
28. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator (2015)
29. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Transactions on Image Processing **27**(5) (May 2018) 2368–2378
30. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. CoRR **abs/1511.04119** (2015)
31. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015)
32. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. (2009) 248–255
33. 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. ACL '02, USA, Association for Computational Linguistics (2002) 311–318
34. : Pixabay - Dog Picture (Aug 2021) [Online; accessed 31. Aug. 2021].
35. : iStock - Angry Dog Picture (Aug 2021) [Online; accessed 31. Aug. 2021].