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

Liane-Marina Meßmer, M. Sc. and Prof. Dr. 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

This work is part of the project “Digital Technologies for People with Dementia (DIDEM)”, which is funded by the Carl Zeiss Foundation
www.didem.hs-furtwangen.de

Authors

Liane Meßmer, M. Sc.

- Academic Staff at the institute for Data Science, Cloud Computing and IT-Security, Furtwangen University of Applied Science
- Research area: machine learning

Prof. Dr. Christoph Reich

- Professor at faculty Informatik of Furtwangen University
- Head of the institute for Data Science, Cloud Computing, and IT Security
- Research area: machine learning, distributed systems, IT security

Content

- Technology Supported Reminiscence Therapy
- Visual Attention Networks
- Dataset
- Evaluation
- Results
- Conclusion and Future Work

Technology Supported Reminiscence Therapy

- Reminiscence Therapy is used to adress the activation process of people with Dementia (pwD).
- Today, the session content often consists of digital media on tablets. Suitable content has to be identified “manually” by the caregivers.
- The identification of suitable content is very labor-intensive and dynamic interaction with the residents is not possible.
- Individual activation and care required for high-quality, biography-based sessions therefore remains a major challenge, despite the extensive availability of digital content today.
- To automatically select suitable reminiscence session content, digital media has to be described automatically.
- **> A method for the automated generation of image descriptions is needed.**

Technology Supported Reminiscence Therapy – Life Themes

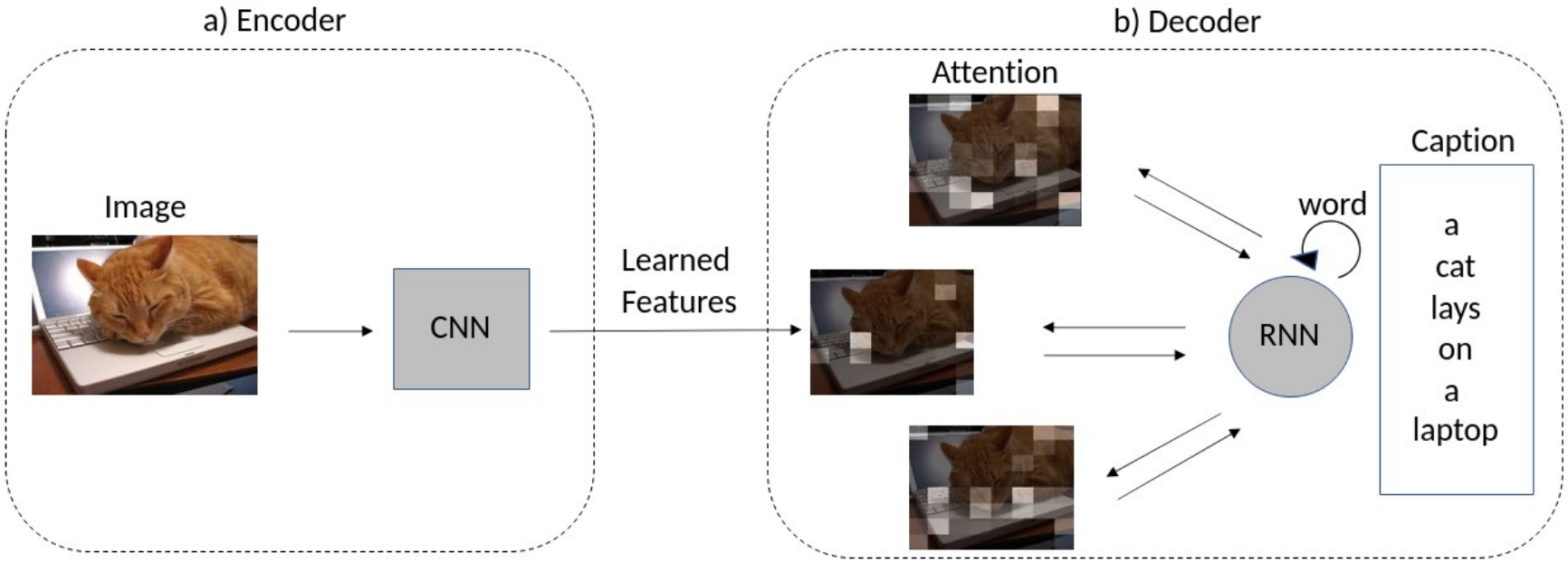
- The reminiscence sessions should include content that match the biographical needs of a pwD, which requires the selection of digital media according to life themes. The focus is on digital images.
- Life themes reflect the biografical content in a dementia patient's life and represent a generalized categorization of life stages or events.
- The goal of these particular subjects is to evoke memories in pwD, which are associated with a life theme.
- Life themes do not only refer to the activation of positive memories, negative memories can also be triggerd. For this purpose, images belonging to the negative life themes should also be described, so that anxiety-producing image content is sorted out before a session.
- **Possible Life themes:** *Animals*, Travel, Hobbies and Activities, Childhood and Youth or Food. The focus is on the description of Animal images.

Visual Attention Network (VAT)

- VATs addresses the problem of generating image descriptions for full scene understanding.
- They consist of an Encoder a) – Decoder b) Architecture.
  - Encoder: Convolutional Neural Network (CNN), used to extract the image featuers into a vectorial representation.
  - Decoder: Recurrent Neural Network (RNN), used to generate appropriate descriptions for the extracted image features.
- The special thing about the Visual Attention Network architecture is that a RNN has an attention layer inside, which is based on the functionality of the human visual system.
- Instead of processing the scene as a whole, the attention is focused on different parts of an image.

Visual Attention Network (VAT)

- **Architecture VAT:** Comprising encoder a) and decoder b)



- **Main objective of this work:** Comparison of different encoder – decoder architectures.
  - **Encoder:** Inceptionv3, ResNet101, VGG16/19, Xception
  - **Decoder:** GRU, LSTM

Dataset

- The focus is on the life theme animals, so a dataset from this area is created.
- The dataset contains: Cats, dogs and angry dogs.
- Classes cat and dog are extacted from the COCO Dataset, which is already labeld for image captioning. Every picture was described by a human with 5 sentences.
- Class angry dog was created from pictures of public databases and labeld by ourselvs with also 5 descriptions per image.
- For example:
  - A brown dog baring his theeth
  - A brown dog looks angry while baring his theeth
  - An angry looking brown dog
  - A brown dog baring his theeth
  - An angry brown dog shows his theeth

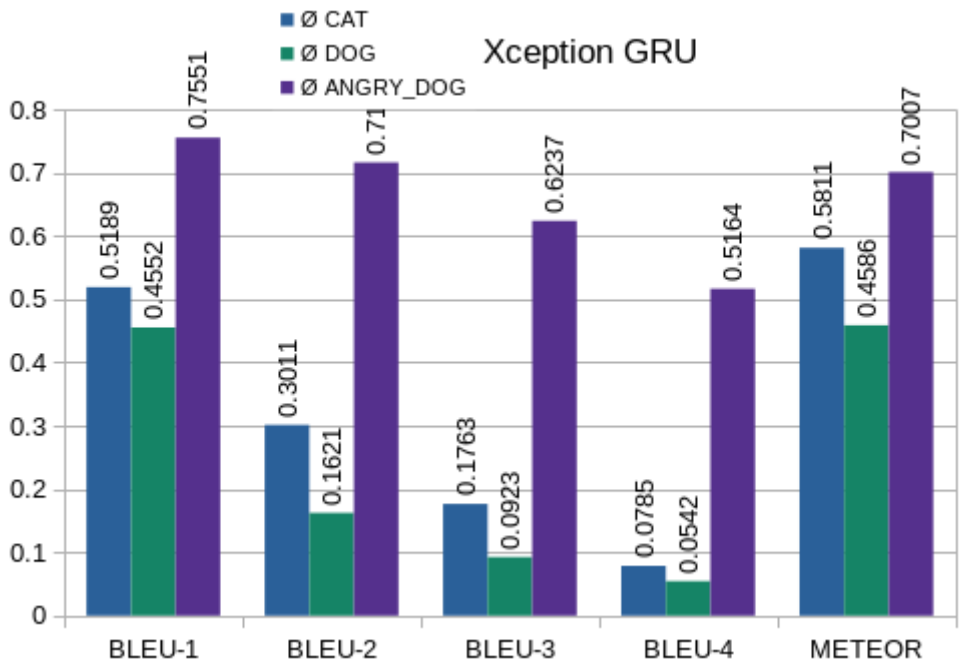
Source: https://pixabay.com/de/photos/hund-b%c3%b6ser-hund-aggressiv-bissig-329280/

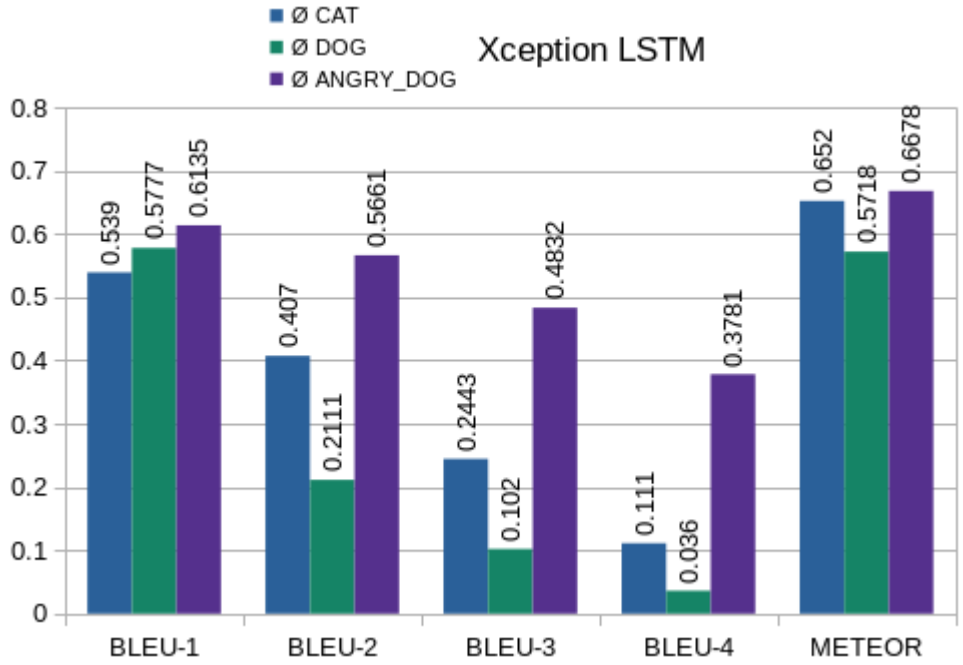
Evaluation by Metric

- BLEU Score and METEOR Metric are used for automated evaluation of machine-generated image captions.
- **BLEU Score**
  - Correlates strongly with the reference captions, as the caption length, word choice and word order are used to calcualte the score.
  - Matching n-grams of words are used to calcualte the result.
- **METEOR Metric**
  - Measures precision (accuracy of the match) and recall (completeness of the match)
  - Uses chunks insted of n-grams
  - **Chunk:** set of unigrams that are adjacent in the hypothesis and the reference
- **Difference:** BLEU measures correlation at the corpus level and METEOR additionaly measures correlation with human judgement at the sentence or segment level

Evaluation by Metric

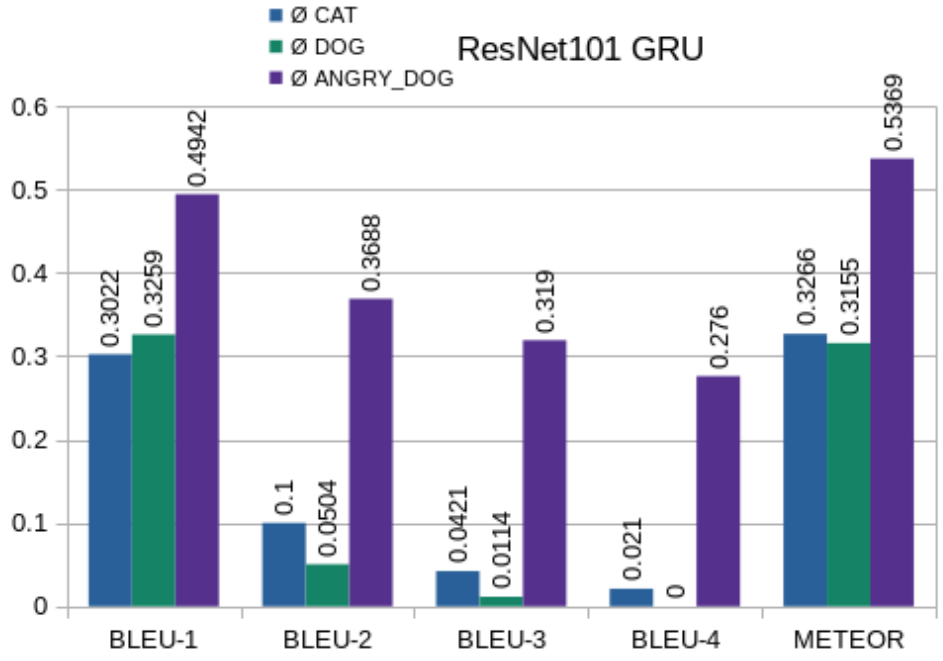
- The results which are calculated by BLEU Score and METEOR Metric are shown in the following figures, where BLEU Score is calculated for different n-grams (n = 1,2,3,4).
- The best results are calculated by Xception as encoder and GRU as decoder.

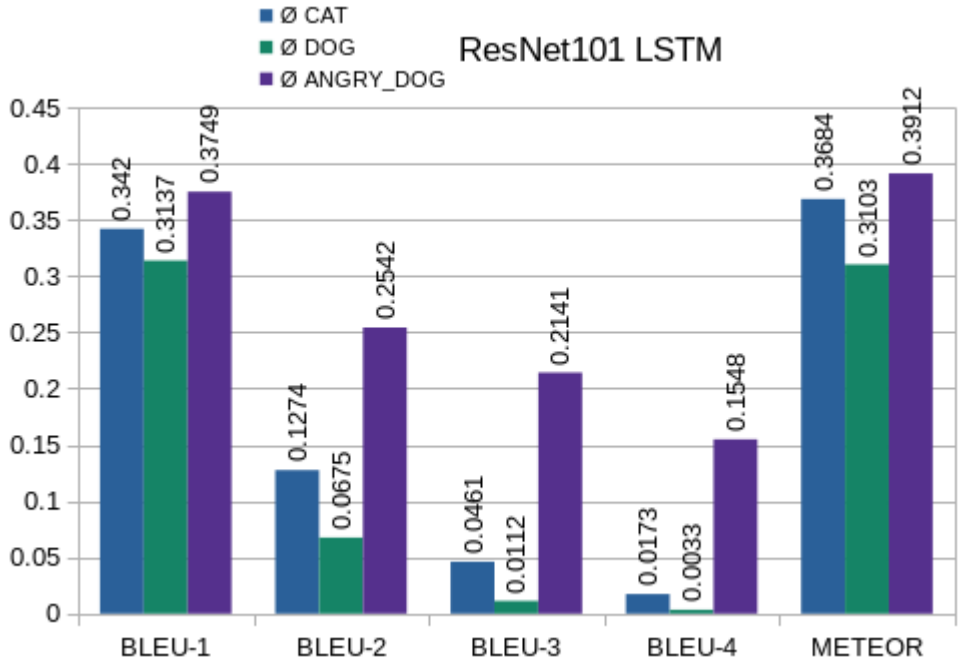




Evaluation by Metric

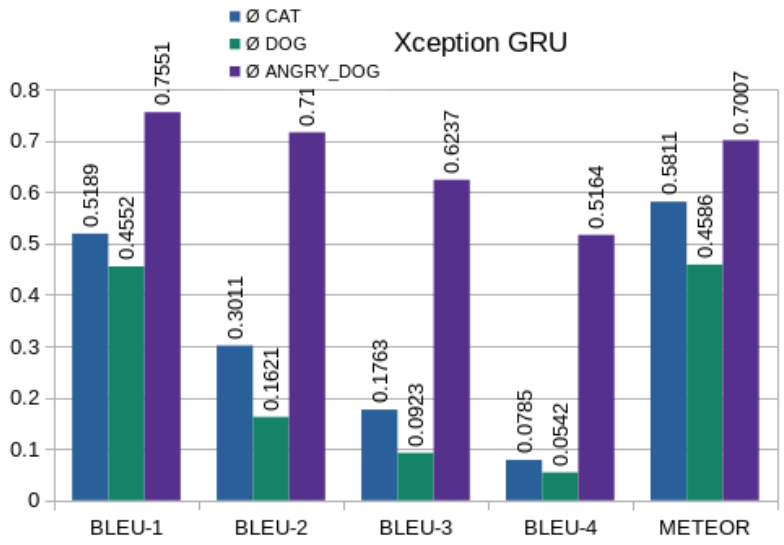
- The worst results are calculated by ResNet101 as encoder and GRU as Decoder.

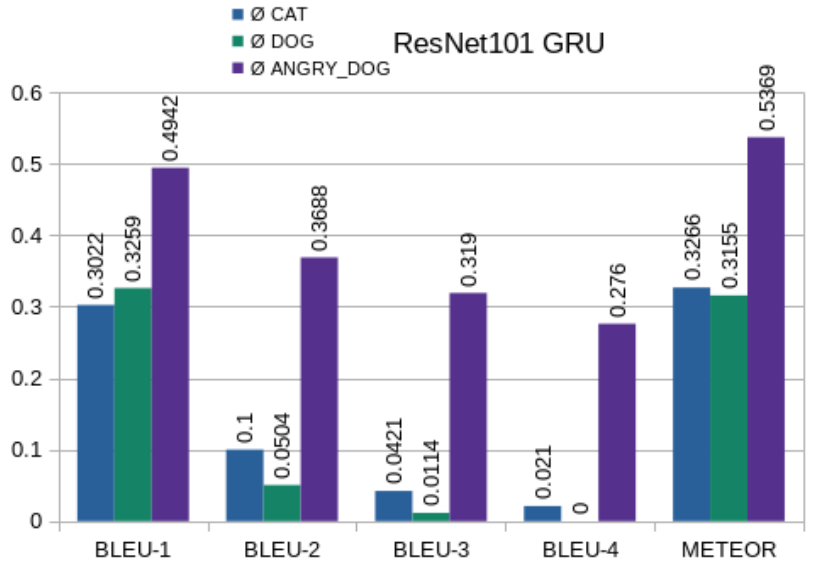




Evaluation by Metric

- Comparison of best and worst results
- In general angry dog class performes best because the reference captions are a better match regarding our use case.
- The use of an Xception network shows a clear improvement over ResNet101. It outperformes ResNet101 in every class.





Evaluation by Human

- Complementing the evaluation using metrics, we conducted a **human evaluation** to determine if the results are good enough in practice for pwD.
- In humans the formal translations exist only in their mind because they are preverbal representations. They can be realized with several synonymous expressions when they are verbalized.
- Therefore, a human evaluator may equally evaluate different translation variants as “correct”, although their evaluations might differ depending on their emphases.
- Even results with low BLEU or METEOR Score could be rich picture decriptions, since many words and sentence positions mean similar things, even if they are formally different from each other.
- For evaluation, we picked one image from the test class, compared to the result caption and collected the best and worst results.

Evaluation by Human – Good Results

(a) Evaluation dog picture

(b) Evaluation cat picture

(c) Evaluation angry dog picture

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 theeth 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 theeth

Evaluation by Human – Bad Results

(a) Evaluation dog picture

(b) Evaluation cat picture

(c) Evaluation angry dog picture

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 theeth on
	ResNet101	a cat is greeting each other in a chair
	VGG16	a brown dog baring his theeth on green grass
	VGG19	a black and brown dog looks angry while baring
	Xception	an angry looking black and brown dog shows his

Results

- **Goal:** Create image descriptions which can be mapped to life themes for pwD.
- Dataset for the life theme “animals” contains images that trigger positive memories in pwD and images that might trigger negative memories.
- Comparison of 5 different enocder models with 2 decoder models.

- **Results:**
- Xception encoder with GRU decoder generates the best results.
- ResNet101 delivers the worst results, no matter which decoder is used.
- Xception network has more stable results than the other networks (less “outliers”).

Conclusion

- This work takes up the basic functionallity of Visual Attention Networks and presents their use based on different network configurations intending to automatically describe images in such a way that they can be assigned to the life themes of demetia patients.
- In this way reminiscence sessions can be automatically created with biography-related content and caregivers are relieved and can invest more time in their patients than in creating the reminiscence session content.
- For comparison, we used Inceptionv3, ResNet101, VGG16/19 and Xception as encoder networks. As decoder we compared GRU and LSTM networks.
- The result was that the Xception – GRU combination produces the best results, whether formally or humanly evaluated.

Future Work

- The system can be extended with other digital media types like music or videos.
- The Dataset we used only covers the life themes “animals”. To make the reminiscnece sessions more valuable, other life themes should be included by extending the training dataset.
- The VAT could be further adjusted by hyperparameter tuning to improve the results and to reduce the number of outlier captions.

Thank you very much for your attention!

For further information, please contact:

Liane-Marina Meßmer and Christoph Reich

{l.messmer, christoph.reich}@hs-furtwangen.de

Furtwangen University
Robert-Gerwig-Platz 1
78120 Furtwangen

Germany

www.furtwangen-university.de
www.didem.hs-furtwangen.de

STUDIEREN AUF HÖCHSTEM NIVEAU