INTERPRETABILITY USING RECONSTRUCTION OF CAPSULE NETWORKS

Dominik VRANAY, Mykhailo RUZMETOV, Peter SINČÁK
Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics,
Technical University of Košice, Letná 9, 042 00 Košice, Slovak Republic, Tel. +421 55 602 5101
E-mail: dominik.vranay@tuke.sk

ABSTRACT

This paper evaluates the effectiveness of different decoder architectures in enhancing the reconstruction quality of Capsule Neural Networks (CapsNets), which impacts model interpretability. We compared linear, convolutional, and residual decoders to assess their performance in improving CapsNet reconstructions. Our experiments revealed that the Conditional Variational Autoencoder Capsule Network (CVAECapOSR) achieved the best reconstruction quality on the CIFAR-10 dataset, while the residual decoder outperformed others on the Brain Tumor MRI dataset. These findings highlight how improved decoder architectures can generate reconstructions of better quality, which can enhance changes by deforming output capsules, thereby making the feature extraction and classification processes within CapsNets more transparent and interpretable. Additionally, we evaluated the computational efficiency and scalability of each decoder, providing insights into their practical deployment in real-world applications such as medical diagnostics and autonomous driving.

Keywords: *Capsule Neural Networks, Model Explainability, Reconstruction Mechanism, Decoder Architectures, Explainable Artificial Intelligence*

1. INTRODUCTION

In recent years, model explainability has become a critical concern within the field of deep learning, particularly in applications where decision-making impacts human lives, such as healthcare, autonomous driving, and security monitoring. Traditional deep learning models, including Convolutional Neural Networks (CNNs) and densely connected networks, often act as "black boxes", making it challenging to interpret their decision-making processes. This lack of transparency poses significant risks in high-stakes environments, where understanding the basis for a model's decisions is crucial [5].

Capsule Neural Networks (CapsNets), introduced by Geoffrey Hinton and his team [21], offer a promising alternative to traditional architectures by maintaining the hierarchical spatial relationships between features. CapsNets utilize groups of neurons, known as capsules, which activate for various properties of an entity, allowing the network to learn more meaningful representations of data. One of the key features of CapsNets that can potentially enhance their interpretability is the reconstruction mechanism. By improving the quality of the reconstruction, it becomes easier to visualize and understand what features the network has learned and how it processes input data [9].

The primary objective of this paper is to explore how different decoder architectures can improve the reconstruction capabilities of CapsNets, thereby indirectly enhancing their interpretability. By examining the interactions between the parameters of the capsules and their ability to reconstruct input data, the study aims to demonstrate how advancements in reconstruction techniques can make the models more transparent and reliable. This work is particularly relevant for applications in medical diagnostics and autonomous driving, where high-quality reconstructions can provide crucial insights into the model's decision-making process.

Furthermore, this paper addresses the computational efficiency and scalability of the proposed decoder architectures, which are critical for practical deployment in real-world applications. By comparing different decoder architectures—linear, convolutional, and residual—this study provides insights into their performance trade-offs, particularly in terms of reconstruction quality, computational requirements, and scalability.

Additionally, we perform a comprehensive evaluation of our approach by comparing it with state-of-the-art methods in both CapsNets and reconstruction techniques. This comparison highlights the unique contributions of our method and situates it within the broader landscape of research in explainable AI (XAI). The inclusion of recent works from top conferences and journals further strengthens the context and relevance of our findings. The specific goals of the paper are:

Enhancing Reconstruction Quality: To investigate how different reconstruction mechanisms in CapsNets contribute to improved reconstruction quality. This involves detailing the reconstruction process, comparing it with similar concepts, and demonstrating how dimension perturbations within capsules affect the network's outputs.

Comparative Analysis of Reconstruction Methods: To review existing methods for reconstruction, including autoencoder-based architectures, and compare their performance with traditional CapsNet reconstructions.

Implementation of Reconstruction Networks: To design and implement various reconstruction networks using CapsNets, detailing the base architecture, technologies, and concepts underlying these networks.

Evaluation of Reconstruction Performance: To evaluate the performance of different reconstruction approaches using selected metrics, providing quantitative and qualitative comparisons to determine the effectiveness of each method in improving model transparency and decision-making reliability.

By addressing these goals, the paper seeks to contribute to the growing body of research on model interpretability in

deep learning, offering insights into how improving reconstruction techniques in CapsNets can lead to more transparent and reliable artificial intelligence systems. This study not only enhances our understanding of CapsNets' reconstruction capabilities but also provides practical guidelines for improving the transparency of deep learning models in real-world applications.

2. LITERATURE REVIEW

This chapter explores the fundamentals and recent advances in Capsule Networks and surveys various neural network reconstruction techniques to enhance performance, interoperability, and explainability in deep learning.

2.1. Capsule Networks and Their Variants

CapsNets, introduced by Sabour et al. [21], address traditional CNN limitations by using capsules—groups of neurons outputting vectors that encode feature probabilities and properties like pose and scale. A dynamic routing algorithm updates coupling coefficients between capsules based on output agreement, enhancing robustness to transformations. CapsNets showed superior performance on the MNIST dataset, particularly with rotated and overlapping digits [21]. See Fig. 1 for a simple CapsNet architecture schematic.

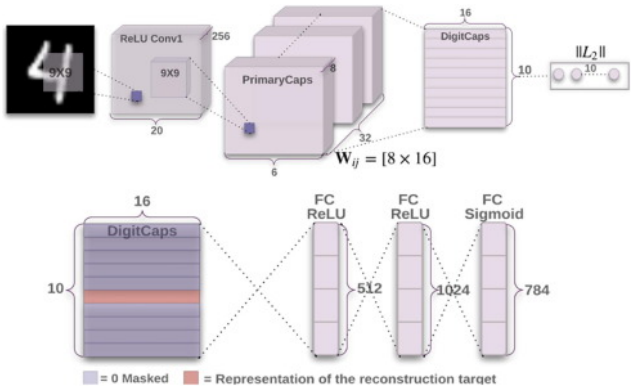

Fig. 1 A simple CapsNet architecture on MNIST dataset [21]

Hinton et al. [10] extended CapsNets with Matrix Capsules, using the Expectation-Maximization (EM) algorithm for routing to better model complex feature relationships. Another variant, Capsule Networks with Inverted Dot-Product Attention Routing, improves routing efficiency using an attention mechanism [25]. DeepCaps, a deeper CapsNet architecture, captures intricate hierarchical relationships, enhancing performance on complex tasks and datasets [20]. CapsNets have also improved object segmentation and detection, especially in challenging scenarios [15].

2.2. Reconstruction Methods in Neural Networks

Reconstruction techniques in neural networks enhance interpretability, training, and applications like image generation and super-resolution.

Autoencoders, the foundational reconstruction-based network, consist of an encoder compressing the input into a latent space and a decoder reconstructing the input. Reconstruction loss, measured by mean squared error (MSE) [2] or binary cross-entropy [4], ensures essential features are captured [12]. See Fig. 2 for a simple autoencoder schematic.

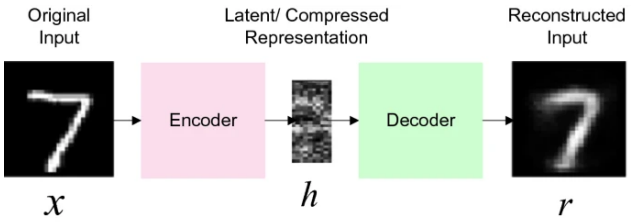

Fig. 2 A simple autoencoder schematic [19]

Variational Autoencoders (VAEs) encode input data into a probabilistic latent space, combining reconstruction loss with a regularization term, typically Kullback-Leibler divergence [14], for more meaningful representations [11]. Generative Adversarial Networks (GANs) consist of a generator creating synthetic data and a discriminator distinguishing real from fake data, leading to high-quality reconstructions [6]. Super-resolution networks like SRCNN [29] and ESRGAN [27] reconstruct high-resolution images from low-resolution inputs using perceptual loss functions for visually appealing results.

In recent years, Vision Transformers (ViTs) have also been employed for image reconstruction tasks, leveraging their ability to model long-range dependencies and capture global context. Unlike traditional CNNs, ViTs process images as sequences of patches, allowing the transformer architecture to learn complex patterns and relationships across the entire image. When used as decoders, ViTs can effectively reconstruct high-quality images by leveraging self-attention mechanisms to focus on relevant features and maintain spatial coherence. This approach has shown promising results in generating detailed and accurate reconstructions, particularly in scenarios where capturing intricate structures and global context is crucial [1, 26].

In CapsNets, reconstruction mechanisms enhance interpretability and regularize training by capturing relevant input features. The original CapsNet includes a reconstruction network using MSE loss to focus on key features [17, 21].

2.3. Explainable Artificial Intelligence (XAI)

XAI aims to make the decision-making processes of complex models more transparent and interpretable.

Saliency maps, like Grad-CAM (Gradient-weighted Class Activation Mapping), visualize influential input regions by highlighting areas contributing most to the predicted class. Grad-CAM generates visual explanations by computing gradients of a target class, producing class-discriminative visualizations without altering the network architecture [23]. Layer-wise Relevance Propagation (LRP)

explains decisions by propagating prediction scores backwards, assigning relevance scores to input pixels based on their contribution. LRP helps understand model decisions at a granular level [3].

Reconstruction-based methods enhance interpretability by reconstructing input data from intermediate representations within the network. This approach provides a tangible means to visualize what the model has learned at different stages of processing, thereby offering insights into the features that influence its predictions [28].

In CapsNets, the reconstruction mechanism is particularly valuable for interpretability, especially through the deformation of the output capsules. CapsNets use capsules to encode various properties of an entity, such as pose, texture, and other attributes. When the output capsules are deformed, they influence the reconstruction process by altering these encoded properties, thereby providing a visual representation of how the network interprets and transforms input features. By examining these deformations, one can gain insights into which features are deemed important by the network and how they contribute to the final prediction.

3. USED ARCHITECTURES

In this chapter, we explore the architectures of networks used in our experiments.

3.1. Simple CapsNet Encoder Architecture

The CapsNet encoder architecture in this paper is structured to extract features from the input data and organize them into capsules, which are then refined through several stages.

The encoder begins with a series of convolutional layers. These layers process the input data to extract various low-level features such as edges, textures, and simple shapes. Each convolutional layer applies a set of filters to the input, enhancing specific patterns and passing the result to the next layer for further processing. This hierarchical feature extraction helps in capturing the essential characteristics of the input.

The output from the final convolutional layer is reshaped into primary capsules. Each primary capsule represents a small group of neurons that encapsulate different properties of the detected features. The primary capsule layer organizes these features into vectors, where each vector's length represents the probability of the feature's presence, and its orientation encodes the spatial information.

The primary capsules are passed through several routing layers, which are designed to refine and aggregate the capsule outputs. The first routing layer takes the primary capsules and routes them to higher-level capsules based on the agreement between their predictions. This routing process continues through intermediate layers, each time refining the capsules and focusing on higher-level patterns and more complex features. The final routing layer produces capsules that represent entire objects or classes, with each capsule corresponding to a specific class and encapsulating the high-level features needed for classification.

3.2. Decoder Architectures

This study explores three different decoder architectures: Linear Decoder, Convolutional Decoder, and Residual Decoder. Each decoder aims to reconstruct the original images from the capsule outputs by focusing on the capsule that corresponds to the identified class.

Linear Decoder: The linear decoder utilizes a series of fully connected layers to reconstruct the original image. Starting from the capsule output, the data is gradually transformed through these layers, increasing in dimensionality until it matches the original image dimensions. This step-by-step transformation ensures that the decoder can effectively rebuild the input image from the high-level features encoded in the capsules.

Convolutional Decoder: The convolutional decoder employs transposed convolutional layers, also known as deconvolutional layers, to upsample the capsule outputs. These layers work by reversing the process of standard convolution, gradually increasing the size of the feature maps and reconstructing the spatial structure of the original image. Each transposed convolution is followed by a ReLU activation to introduce non-linearity, and the final layer uses a Sigmoid activation to ensure the output values are within the range [0, 1] [16].

Residual Decoder: The residual decoder incorporates residual blocks, which are designed to maintain the integrity of the data as it is transformed. Each residual block includes convolutional layers and non-linear activations, along with skip connections that allow the data to bypass certain transformations. This helps preserve important features and enables the decoder to produce more accurate reconstructions. The feature map is progressively upsampled through a series of transposed convolutional layers, similar to the convolutional decoder, ensuring the reconstructed image is detailed and faithful to the original [7, 8].

3.3. Conditional Variational Capsule Network

The Conditional Variational Capsule Network (CVAE-CapOSR) integrates Conditional Variational Autoencoders (CVAEs) with Capsule Networks (CapsNets) to enhance open-set recognition tasks. This architecture leverages the strengths of both models to handle complex and diverse visual data more effectively than traditional CapsNets [7].

CapsNets excel in preserving hierarchical relationships in data, making them suitable for spatial and temporal tasks. However, they often struggle with high variability in data, especially with images that have diverse backgrounds and varying object appearances. To address this, CVAEs enhance data reconstruction capabilities by modelling data distributions within the latent space.

CVAEs extend traditional VAEs by incorporating conditional variables into the generative process, using additional label information during encoding and decoding. This enables CVAEs to produce outputs based on specific conditions, making them more adaptable to new and unseen data categories [22, 24].

In CVAECapOSR, the encoder transforms input data into a latent space conditioned on class labels, shaping the latent space to reflect the attributes of the given labels. The

decoder then generates data from this conditioned latent space, ensuring outputs adhere to the specified conditions, which is crucial for handling new, unseen data categories in open set recognition tasks.

The encoder architecture first employs the deep convolutional neural network ResNet34 for initial feature extraction. ResNet34 processes the input image to produce a complex feature map and intermediate outputs, which are later utilized in the decoding stage.

The decoder reconstructs the input image from the encoded latent representation using a combination of convolutional, residual and transposed convolutional layers. It incorporates lateral connections, which reintegrate spatial details from intermediate encoder layers, improving the detail and accuracy of the reconstructed image. However, these lateral connections can reduce interpretability by blending features from different stages, obscuring the contributions of individual layers to the final output. Overall architecture can be seen in Fig. 3.

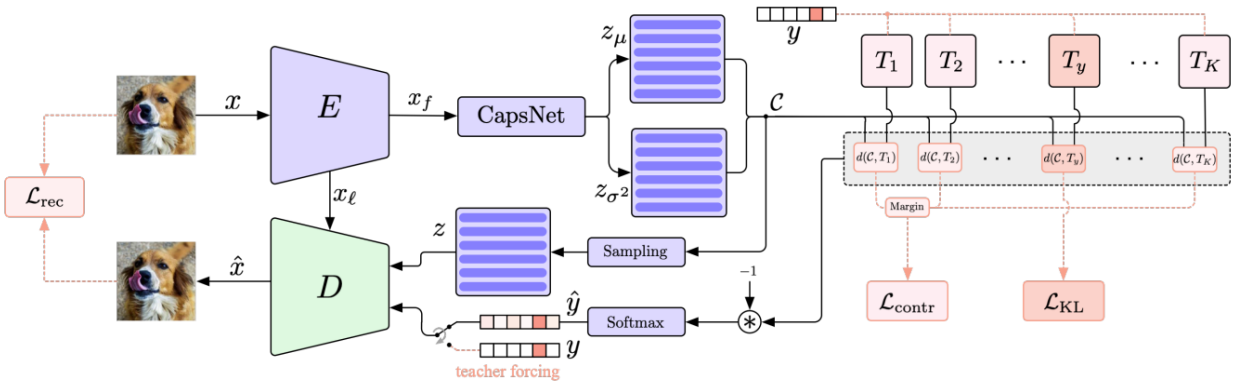

Fig. 3 Graphical representation of the network's architecture [24]

The loss function in CVAECapOSR includes both the reconstruction loss and the KL divergence term, which ensures that the learned latent space distribution is close to the true distribution while maintaining the ability to generate conditioned outputs.

By combining the hierarchical data structuring of CapsNets with the generative capabilities of CVAEs, the CVAECapOSR model excels in open set recognition tasks. It can effectively handle new and diverse categories of data by leveraging conditional information during the generation process, thus offering robust performance in complex visual recognition scenarios.

3.4. Background Capsules

Background capsules are often used and integrated into the network in a manner similar to class/output capsules, ensuring that each additional capsule has the same number of dimensions. Their primary connection is to the reconstruction loss, allowing them to focus on capturing features crucial for reconstruction, such as background details, rather than features necessary for classification. This separation enables the network to accurately reconstruct the background without interference from classification tasks. The integration of additional background capsules also into the CVAECapOSR model significantly enhances the quality of reconstructions. This improvement is especially notable when lateral connections are not utilized.

4. EXPERIMENTS

In this section, the experiments are described and the results are discussed.

4.1. Experimental Setup

The models were trained on an Nvidia RTX A4500 GPU with 20 GB of memory. To ensure reproducibility, a random seed was set to 42. The training configuration varied slightly depending on the model and dataset used. Specifically, the traditional CapsNet was trained with a learning rate of 1×10^{-3}, while the CVAECapOSR required a lower learning rate of 5×10^{-5} due to instability at higher rates. The importance of the reconstruction loss relative to the classification loss was balanced by setting the reconstruction loss coefficient to 1.0.

The batch size for most datasets was set to 32, except for the Brain Tumor MRI dataset, which had a batch size of 16 due to the higher resolution of the images causing GPU memory overflow with larger batches. The models were trained using the Adam optimizer with different learning rate schedulers: ExponentialLR for the traditional CapsNet and ReduceLROnPlateau for the CVAECapOSR. Data augmentation techniques, such as rotations and random crops, were employed to enhance generalization. The training duration was 50 epochs for most datasets and 100 epochs for the Brain Tumor MRI dataset.

Two datasets were utilized to evaluate the quality of the reconstructions across various data types:

CIFAR-10: This dataset includes RGB images with diverse and complex features, providing a more challenging

test for the reconstruction capabilities [13].

Brain Tumor MRI Dataset: High-resolution MRI images of brain tumours were used to test the models on medical imaging data [18].

The performance of the models was assessed using three primary metrics:

MSE: This measures the average of the squares of the errors, which is the difference between the original and the reconstructed images.

Structural Similarity Index (SSIM): This index measures the similarity between two images, focusing on changes in structural information.

Peak Signal-to-Noise Ratio (PSNR): This metric measures the ratio between the maximum possible power of a signal and the power of corrupting noise, expressed in decibels (dB). Higher PSNR values indicate better image quality.

These metrics provided a comprehensive evaluation of the reconstruction performance, capturing both pixel-wise accuracy (MSE), perceptual quality (SSIM), and signal quality (PSNR).

4.2. Effect of Reconstruction Capsule Number on Reconstruction Performance

We assessed the impact of varying the number of reconstruction capsules (0, 1, and 5) on the reconstruction quality using the CIFAR-10 dataset with the CVAECapOSR architecture.

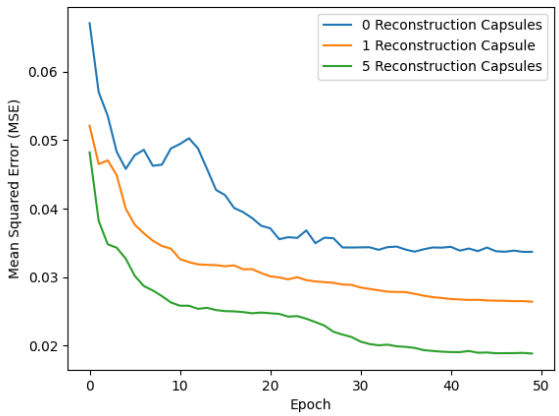

Fig. 4 Comparison of 0, 1, and 5 reconstruction capsules

Figure 4 depicts the mean squared error over 50 epochs for configurations with 0, 1, and 5 reconstruction capsules. Adding more reconstruction capsules consistently improved the reconstruction quality, as evidenced by the lower MSE values. The configuration with 5 reconstruction capsules achieved the best performance, followed by the configuration with 1 reconstruction capsule, and finally, the configuration with 0 reconstruction capsules.

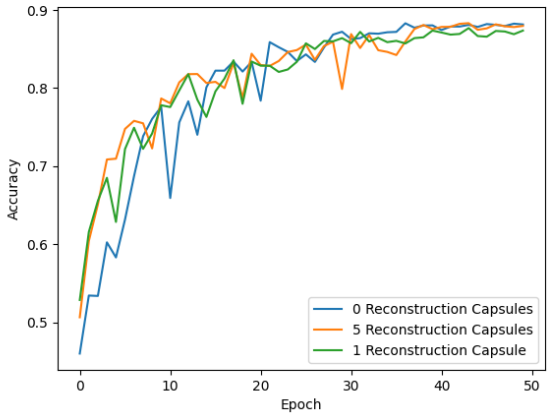

Fig. 5 Accuracy comparison of configurations with 0, 1, and 5 reconstruction capsules over 50 epochs

Figure 5 shows the accuracy behaviour of the network over 50 epochs with 0, 1, and 5 reconstruction capsules. The final accuracies are 0.8815 for 0 capsules, 0.8737 for 1 capsule, and 0.8801 for 5 capsules. These results suggest that the number of reconstruction capsules does not directly influence overall accuracy.

Fig. 6 Example reconstructions with the original image, 0, 1, and 5 reconstruction capsules, respectively

Figure 6 provides a visual comparison of the reconstructions from the CIFAR-10 dataset for each configuration. The images reconstructed with 5 capsules are visually more accurate and detailed than those with fewer capsules.

4.3. Evaluation of Different Architectures

We evaluated the performance of the CVAECapOSR and three different decoder architectures used with a traditional CapsNet: Residual Decoder, Linear Decoder, and Convolutional Decoder.

Table 1 Reconstruction performance metrics for different architectures on the CIFAR-10 dataset

Architecture	MSE	SSIM	PSNR
Residual Decoder	0.0330	**0.4040**	15.213
Linear Decoder	0.0375	0.2741	14.732
Convolutional Decoder	0.0383	0.2908	14.495
CVAECapOSR	**0.0273**	0.2568	**16.194**

Table 1 summarizes the reconstruction performance metrics (MSE, SSIM, PSNR) for different architectures on the CIFAR-10 dataset. The CVAECapOSR's low MSE indicates the best reconstruction accuracy, although its SSIM score suggests it may not preserve perceptual image quality as effectively as the Residual Decoder.

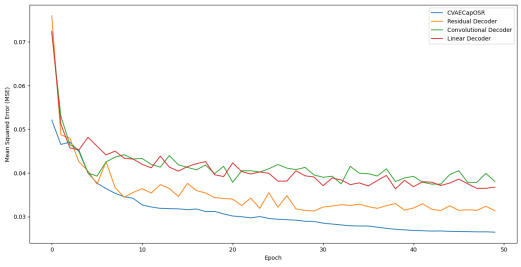

Fig. 7 MSE over 50 epochs of training for CVAECapOSR, Residual Decoder, Convolutional Decoder, and Linear Decoder on the CIFAR-10 dataset

Figure 7 illustrates the MSE over 50 epochs for CVAECapOSR and the three different decoders. The CVAECapOSR consistently achieves lower MSE values compared to the other decoders throughout the training process.

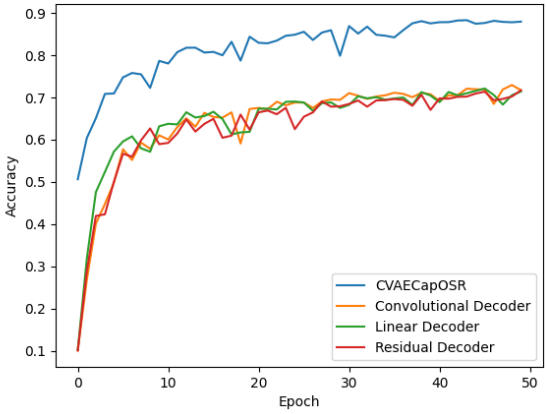

Fig. 8 Accuracy comparison of CVAECapOSR, Residual Decoder, Linear Decoder, and Convolutional Decoder over 50 epochs

Figure 8 shows the accuracy behaviour of the network over 50 epochs with the CVAECapOSR and three decoder architectures. The final accuracies are 0.8737 for CVAECapOSR, 0.717 for Residual Decoder, 0.7142 for Linear Decoder, and 0.7173 for Convolutional Decoder. These results indicate that while CVAECapOSR consistently outperforms the other architectures in terms of accuracy, the type of decoder itself does not have a significant influence on accuracy values.

4.4. Performance on High-Resolution Data

This section compares the performance of the CVAECapOSR and a traditional CapsNet with a Residual Decoder on the Brain Tumor MRI Dataset, which consists of high-resolution grayscale images.

Table 2 Performance metrics for CVAECapOSR and traditional CapsNet with Residual Decoder

Architecture	MSE	SSIM	PSNR
CVAECapOSR	0.0164	0.3897	18.804
Residual Decoder	**0.0109**	**0.4354**	**25.785**

As can be seen from Table 2, the traditional CapsNet with a Residual Decoder achieved better performance across all metrics, indicating it is more effective at minimizing reconstruction error and preserving image quality for high-resolution images.

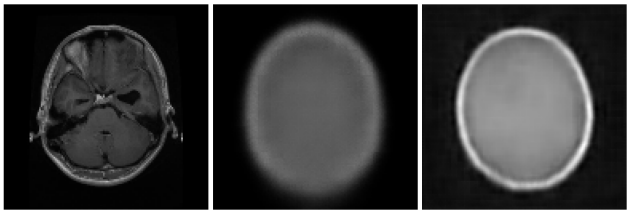

Fig. 9 Example reconstructions with the original image, CVAECapOSR, and the Residual Decoder, respectively

Figure 9 shows reconstructed images from the Brain Tumor MRI Dataset for both networks, illustrating that the Residual Decoder produces more accurate and visually appealing reconstructions compared to the CVAECapOSR.

5. CONCLUSION

CapsNets offer promising advancements in artificial intelligence, particularly in enhancing model interpretability through improved reconstruction mechanisms. This study explored various decoder architectures—linear, convolutional, and residual—to assess their effectiveness in reconstructing input images. The results indicate that CapsNets, through their unique capsule structures, have the potential to produce high-quality reconstructions, which in turn aid in their interpretability by making the internal processing more transparent.

Our experiments demonstrated that all decoder architectures were able to perform their classification objectives effectively, with the CVAECapOSR model achieving the best results by leveraging a pretrained ResNet as the backbone. The linear decoder, while straightforward, is the least scalable due to its exponential increase in parameters with larger fully connected layers, making it less suitable for high-dimensional data.

The convolutional and residual decoders offered a better balance between scalability and reconstruction performance. The convolutional decoder, with its parameter-efficient architecture, maintained good reconstruction quality, making it a practical choice for many applications. The residual decoder, although slightly more computationally expensive and slower, achieved superior reconstruction

quality, making it ideal for scenarios where interpretability is critical.

CVAECapOSR demonstrated the best overall performance in terms of reconstruction quality and interpretability, despite its higher computational demands. This architecture remains feasible for larger datasets, provided that sufficient computational resources are available.

Improved reconstruction quality directly enhances interpretability by providing clearer and more detailed visualizations of what the network has learned. By examining the reconstructed images, one can infer how the network processes and transforms input features, thus gaining valuable insights into the decision-making process of the CapsNets.

While CapsNets and their reconstruction capabilities show significant potential in enhancing model interpretability, the choice of decoder architecture must consider trade-offs between reconstruction quality, scalability, and computational efficiency. Future work should focus on optimizing these decoders to improve performance and reduce computational demands. Additionally, integrating reconstruction-based interpretability with other explainable AI techniques could offer more comprehensive insights into model behaviour, fostering greater trust and understanding in AI systems.

ACKNOWLEDGEMENT

This work was supported by the Slovak National Science Foundation project "Basic Research of Deep Learning for Image processing - DL4VISION" supported during 2022-2025 under registration code 1/0394/22, additionally also by the LIFEBOTS Exchange project funded from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 824047.

REFERENCES

[1] A. M. ALI, B. BENJDIRA, A. KOUBAA, W. EL-SHAFAI, Z. KHAN, and W. BOULILA. Vision transformers in image restoration: A survey. *Sensors*, 23(5):2385, 2023.

[2] D. M. ALLEN. Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13(3):469–475, 1971.

[3] A. BINDER, G. MONTAVON, S. LAPUSCHKIN, K.-R. MÜLLER, and W. SAMEK. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning–ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*, pages 63–71. Springer, 2016.

[4] D. R. COX. The regression analysis of binary sequences. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 20(2):215–232, 1958.

[5] S. GHOLIZADEH and N. ZHOU. Model explainability in deep learning based natural language processing. *arXiv preprint arXiv:2106.07410*, 2021.

[6] I. J. GOODFELLOW, J. POUGET-ABADIE, M. MIRZA, B. XU, D. WARDE-FARLEY, S. OZAIR, A. COURVILLE, and Y. BENGIO. Generative adversarial networks, 2014.

[7] Y. GUO, G. CAMPORESE, W. YANG, A. SPERDUTI, and L. BALLAN. Conditional variational capsule network for open set recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 103–111, 2021.

[8] K. HE, X. ZHANG, S. REN, and J. SUN. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] G. E. HINTON, A. KRIZHEVSKY, and S. D. WANG. Transforming auto-encoders. In *Artificial Neural Networks and Machine Learning–ICANN 2011: 21st International Conference on Artificial Neural Networks, Espoo, Finland, June 14-17, 2011, Proceedings, Part I 21*, pages 44–51. Springer, 2011.

[10] G. E. HINTON, S. SABOUR, and N. FROSST. Matrix capsules with em routing. In *International conference on learning representations*, 2018.

[11] D. P. KINGMA and M. WELLING. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[12] M. A. KRAMER. Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, 37(2):233–243, 1991.

[13] A. KRIZHEVSKY, G. HINTON, et al. Learning multiple layers of features from tiny images. 2009.

[14] S. KULLBACK and R. A. LEIBLER. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[15] R. LALONDE and U. BAGCI. Capsules for object segmentation. *arXiv preprint arXiv:1804.04241*, 2018.

[16] J. LONG, E. SHELHAMER, and T. DARRELL. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[17] V. MAZZIA, F. SALVETTI, and M. CHIABERGE. Efficient-capsnet: Capsule network with self-attention routing. *Scientific reports*, 11(1):14634, 2021.

[18] M. NICKPARVAR. Brain tumor mri dataset, 2021.

[19] K. PAWAR and V. Z. ATTAR. *Assessment of Autoencoder Architectures for Data Representation*, pages 101–132. Springer International Publishing, Cham, 2020.

[20] J. RAJASEGARAN, V. JAYASUNDARA, S. JAYASEKARA, H. JAYASEKARA, S. SENEVIRATNE, and R. RODRIGO. Deepcaps: Going deeper with capsule networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10725–10733, 2019.

[21] S. SABOUR, N. FROSST, and G. E. HINTON. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

[22] W. J. SCHEIRER, A. De REZENDE ROCHA, A. SAPKOTA, and T. E. BOULT. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[23] R. R. SELVARAJU, M. COGSWELL, A. DAS, R. VEDANTAM, D. PARIKH, and D. BATRA. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[24] K. SOHN, H. LEE, and X. YAN. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28, 2015.

[25] Y.-H. H. TSAI, N. SRIVASTAVA, H. GOH, and R. SALAKHUTDINOV. Capsules with inverted dot-product attention routing. *arXiv preprint arXiv:2002.04764*, 2020.

[26] N. VERMA, D. KAUR, and L. CHAU. Image reconstruction using enhanced vision transformer. *arXiv preprint arXiv:2307.05616*, 2023.

[27] X. WANG, K. YU, S. WU, J. GU, Y. LIU, Ch. DONG, Y. QIAO, and Ch. CHANGE L. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018.

[28] Y. WANG, C. QIN, Y. BAI, Y. XU, X. MA, and Y. FU. Making reconstruction-based method great again for video anomaly detection. In *2022 IEEE International Conference on Data Mining (ICDM)*, pages 1215–1220. IEEE, 2022.

[29] J. YAMANAKA, S. KUWASHIMA, and T. KURITA. Fast and accurate image super resolution by deep cnn with skip connection and network in network. In *Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14-18, 2017, Proceedings, Part II 24*, pages 217–225. Springer, 2017.

Received May 23, 2024, accepted July 15, 2024

BIOGRAPHIES

Dominik Vranay received the MEng. degree in Computer Science from the University of Southampton, Southampton, England in 2021. His research interest includes computer vision and deep learning with a focus on XAI and Capsule networks.

Mykhailo Ruzmetov finished his BSc. degree in 2024 at the Department of Cybernetics and Artificial Intelligence of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice.

Peter Sinčák received the MSc. degree in Intelligent Systems from Technical University of Košice, Košice, Slovakia in 1984 and the PhD degree in electronic engineering from the Academy of Science in Prague, Czechia, in 1992. Since then, he has been an Assistant Professor, an Associate Professor and a Full Professor in the AI branch at the Technical University of Košice. His research interests include the integration of Cloud Computing, Artificial or Computational Intelligence and Robotics. He is the author and co-author of more than 100 papers at conferences, and journals he is the author of 2 monographs and an editor in a number of books in Physica Verlag, IOS press and WorldScientific Publication Houses.