On the unreasonable effectiveness of CNNs

Andreas Hauptmann, *Member, IEEE,* and Jonas Adler

***Abstract*—Deep learning methods using convolutional neural networks (CNN) have been successfully applied to virtually all imaging problems, and particularly in image reconstruction tasks with ill-posed and complicated imaging models. In an attempt to put upper bounds on the capability of baseline CNNs for solving image-to-image problems we applied a widely used standard off-the-shelf network architecture (U-Net) to the "inverse problem" of XOR decryption from noisy data and show acceptable results.**

I. Introduction

An ever-increasing amount of data and emerging methods in deep learning have led to considerable advances in numerous computer vision tasks, such as object detection and classification. The underlying methodology is based on convolutional neural networks (CNN), which can be understood as a multi-layered feature extraction from neighbourhood relations in the input image.

The subclass of methods that we consider here are image-to-image networks and in particular, we are motivated by their immense impact in inverse problems and biomedical imaging applications. This success is partly driven by the availability of large amounts of data and the tendency to open-source this information for other scientists, but also by the ability to learn visually appealing and data specific representations. Consequently, we are now experiencing a transition to utilise tools from data science to harness the potential of large data, which is in stark contrast to classical deterministic algorithms that only operated with few data. For instance, whereas in tomographic imaging meticulously tuned algorithms were developed for reconstructions, we can now simply train a network to recover the important features in a tomographic image, if sufficient data is available.

This transition is marked in parts due to the success of widely established convolutional neural network architectures, such as the U-net which was initially used for semantic segmentation [1], but has since been utilised for various imaging tasks as an essential processing step to improve image quality. For instance, to remove noise in low-dose CT images [2], [3] or artefacts from undersampled magnetic resonance imaging [4], [5], and as essential part in the pipeline for automated diagnosis [6].

However, it is not unusual to propose a new method using different network designs [7], [8], or intertwine learned components with explicit hand-designed operations [9]–[11]. These methods generally provide impressive results, but it is often claimed without computational proof that the problem under consideration is too special for the application of basic network architectures. We believe, this is in part due to the common understanding that the problem has to be sufficiently well-behaved for standard CNNs to be applicable [12]. For example, the use of convolutions would imply that the underlying problem should be translation equivariant[1], we refer to [13] for a discussion in the context of inverse problems. Furthermore, the continuity and almost everywhere differentiability of the neural networks seemingly implies that the functions they approximate should at least be continuous [14].

These somewhat contradictory trains of thought lead us to our question: *are there really any image-to-image tasks that can not be solved reasonably well with a standard CNN and sufficient data?* Our main result is that even for seemingly ridiculous image-to-image problems standard CNNs basically always perform acceptably. In particular we'll show that an off-the-shelf U-Net can be trained to invert XOR encryption, a function which is *everywhere discontinuous* and *not translation equivariant*. The results hold even in the noisy case where standard decryption fails.

II. Experimental setup

We shall consider the inverse problem of inverting XOR encryption. Our forward operator is hence the XOR operation with a fixed byte string and the implementation we use is based on running the advanced encryption standard (AES) [15] in Counter mode (CTR) [16], a block cipher with randomly generated initialisation (iv) and key, here 128 bit long and fixed for all examples. The encryption process first generates a string of bytes using the initialisation and key and then takes the input, represented as a sequence of bytes, applies a bitwise XOR of the input and byte string, and then returns a byte array of the same size, called the ciphertext. The recipient of the ciphertext can then use the key in order to recover the input. For obvious reasons, the process has been designed to be highly discontinuous in the input in order to thwart attackers from reading the text. To make the problem even harder, we also study the case with noisy observations.

As with any standard machine learning method for inverse problems, the recovery of the encrypted images can be formulated as a basic supervised learning problem. That is, given a set of ground-truth images $\{f_i\}$ with corresponding measured data $\{g_i\}$, we then formulate a network Λ_θ with parameters θ to recover the ground-truth from the encrypted image, such

This work was partially supported by the Academy of Finland Project 312123 (Finnish Centre of Excellence in Inverse Modelling and Imaging, 2018–2025) and the CMIC-EPSRC platform grant (EP/M020533/1)

A. Hauptmann is with the Research Unit of Mathematical Sciences; University of Oulu, Oulu, Finland and with the Department of Computer Science; University College London, London, United Kingdom.

J. Adler did this work at the department of mathematics; KTH – Royal Institute of Technology, Stockholm Sweden. He is currently with DeepMind, London, UK.

[1]A function is translation equivariant if translating the input and then applying the function is equivalent to applying the function and then translating the output. All convolutions satisfy this property.

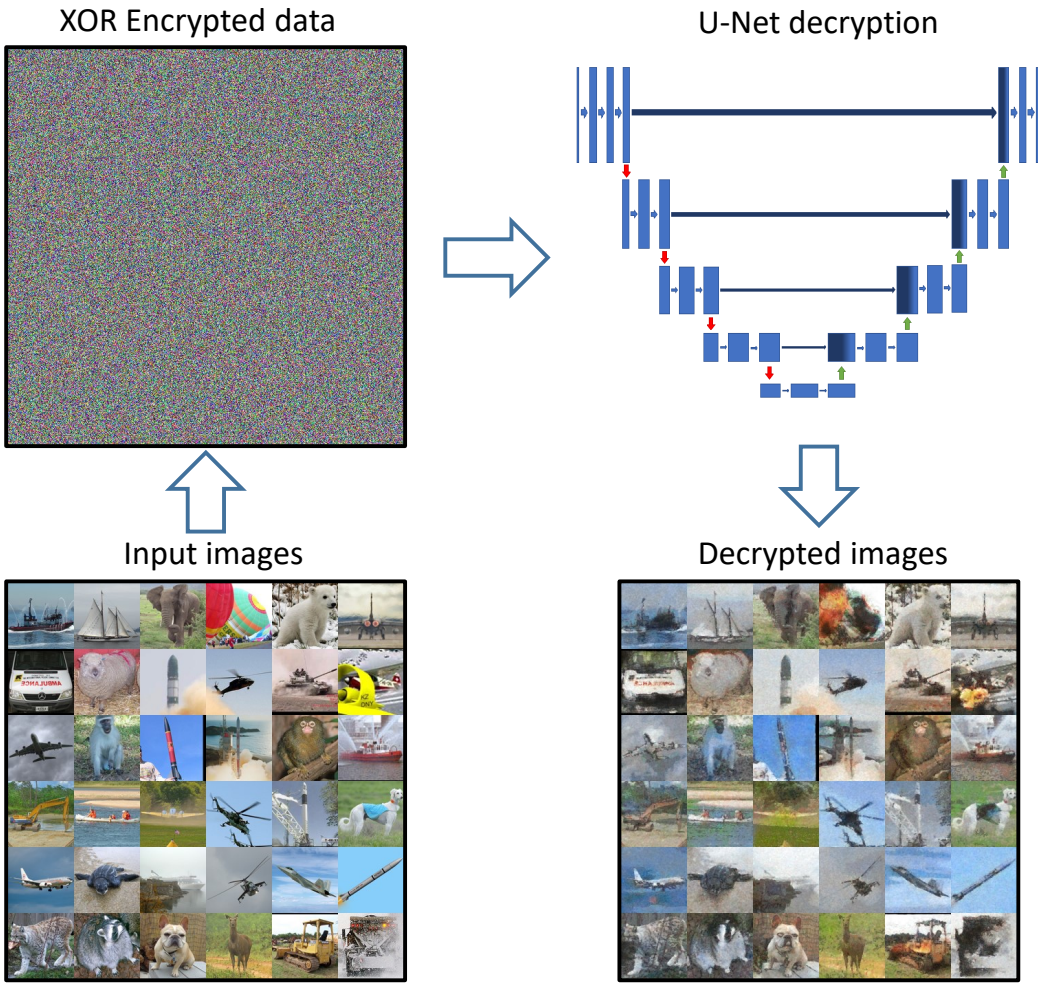

Fig. 1. Experimental setup for XOR encryption of natural images and decryption using a standard off-the-shelf deep convolutional neural network. The network is trained using known cipherimages (top-left images, 3 out of 12 channels shown here) ⇔ plaintext (bottom left) pairs. Reconstructed test images after successful training (bottom right) exhibit a loss of resolution and detail, but retain identifying characteristics.

that $\Lambda_\theta(g_i) \approx f_i$. This corresponds to the standard procedure in image processing, where g_i represents a corrupted image, e.g. obtained from undersampled data and/or under high noise or even, as in our case, from bit-wise encoding. The training is then given as optimisation problem to find an optimal set of parameters θ^* by minimising a loss functions such as

$$\theta^* = \arg\min_\theta \frac{1}{N} \sum_{i=1}^{N} \|\Lambda_\theta(g_i) - f_i\|_2^2. \tag{1}$$

For training, we use the standardised STL-10 dataset [17], a set of 100,000 natural 96x96 pixel colour images (bottom-left of Fig. 1). The images we seek to encode are RGB images, which we normalised to the range $[0, 1]$ and then stored as single precision floating point numbers, three per pixel (one per colour channel). We encrypt the raw byte-representation of the image using AES in CTR mode with the fixed 128 bit key and fixed iv. As we aim to establish an image-to-image problem, we need to reinterpret the ciphertext as an image and make it admissible as input to the U-Net. To achieve this, we first converted the encryption output to `uint8`, then normalised similarly to the ground-truth images and reformatted to a 96×96 image with resulting 12 channels in single precision floats, which we call the call cipherimages, see top-left of Fig. 1 for examples. In the noisy setting we add 1% pixel- and channel-wise Gaussian noise.

We split the data into 90,000 training images and 10,000 images for testing. The corresponding encrypted images were then computed for both sets with the same settings and noise added. A standard U-net architecture as originally proposed [1], with a minor modification for consistency in image size and linearly rescaling the inputs to [0, 1), was used. The training was done by minimising Eq. (1) and performed with standard choices on the hyperparameters, using the Adam optimiser, batch size of 32, learning rate $2{\cdot}10^{-3}$ and 35 epochs.

The XOR forward operator has a closed form inverse and this inverse serves as our decryption baseline. In the noisy case we first round to the closest byte. We also consider another trivial baseline to see if we have actually learned decryption and not just some statistics about the dataset. Here, we note that the minimise of Eq. (1) if g_i is uninformative of f_i is the

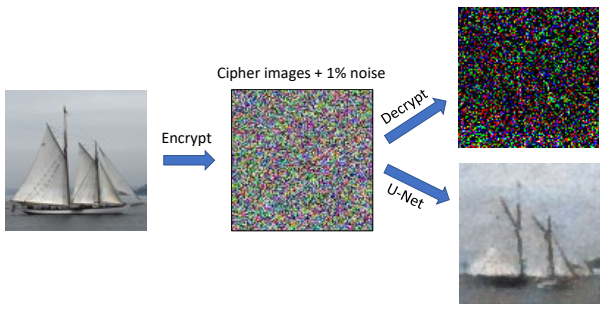

Fig. 2. Reconstruction from corrupted data. We added 1% noise to the cipherimages (Middle, shown 3 out of 12 channels). Resulting reconstructions on the right are with the XOR decryption with known key (top) and learned reconstruction (bottom).

mean of the training set, $\bar{f} = \frac{1}{N}\sum_{i=1}^{N} f_i$ and use this as the reconstruction for any input.

III. Discussion of results

TABLE I
Quantitative values for the recovered test images. Mean values for 10,000 test samples.

	No noise		1% Noise	
	PSNR	SSIM	PSNR	SSIM
U-Net	18.0	0.62	**17.6**	**0.57**
Decryption	$+\infty$	**1.00**	$-\infty$	0.00
Mean of train set	11.8	0.16	11.8	0.16

To our surprise, the network can successfully establish a relation between the cipherimages and the original input images. A set of such reconstructed images by the network from the test set is shown in Fig. 1. As one can see, the reconstructed images suffer from a loss of resolution and some colours are not correctly recovered. Nevertheless, we can say that the reconstruction quality is well beyond our expectation.

This qualitative observation is also supported in terms of quantitative values, as the mean PSNR of all 10,000 reconstructed test images after training was 18dB and a SSIM of 0.62, as shown in Table I. In this noise free setting the baseline decryption method performs perfectly and obtains an infinite PSNR, as expected. In the noisy setting with the learned decryption we observe a minor deterioration in mean PSNR of 0.4dB and 0.05 in SSIM, whereas decryption from the corrupted data is impossible, giving a PSNR of negative infinity. We illustrate this in Fig. 2. Additionally, the simple mean baseline gives notably worse results, indicating that a nontrivial relation has indeed been successfully established.

As for limitations of this study, we note that while our encryption operation is different for each pixel (due to the XOR key being different) and hence not translation equivariant, there is no connection between pixels. We also tried running AES in Cipher Block Chaining (CBC) mode, which introduces a strong dependency, but were unable to learn anything useful in this setting. This was also the case for lower-dimensional training data such as the MNIST dataset.

Finally, we reiterate that a safe encryption uses varying keys as well as a random iv. If these are fixed, the encryption is deterministic and consequently not safe. This work does not show that supervised learning can crack encryption, what it does show is that supervised learning can solve problems with terrible numerical properties.

IV. Conclusions

Our results indicate, that against ones intuition, an off-the-shelf CNN can successfully learn a relation even in this extremely challenging setting. We believe there are at least two important takeaways from this observation that should be kept in mind when solving image-to-image problems:

- With enough data, basic CNNs such as a U-Net, always work on image-to-image problems.
- Training a simple CNN should always be used as a strong baseline for *any* newly proposed method, regardless of how unreasonable it seems.

Acknowledgement

The authors would like to thank Ozan Öktem for suggesting that decryption ought to be an impossible inverse problem for supervised deep learning. We also thank Sebastian Lunz and Olaf Ronneberger for helpful discussions and comments.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[2] Eunhee Kang, Junhong Min, and Jong Chul Ye. A deep convolutional neural network using directional wavelets for low-dose x-ray ct reconstruction. *Medical physics*, 44(10), 2017.

[3] Kyong Hwan Jin, Michael T McCann, Emmanuel Froustey, and Michael Unser. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing*, 26(9):4509–4522, 2017.

[4] Dongwook Lee, Jaejun Yoo, and Jong Chul Ye. Deep residual learning for compressed sensing mri. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 15–18. IEEE, 2017.

[5] Andreas Hauptmann, Simon Arridge, Felix Lucka, Vivek Muthurangu, and Jennifer A Steeden. Real-time cardiovascular mr with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease. *Magnetic resonance in medicine*, 81(2):1143–1156, 2019.

[6] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342, 2018.

[7] Bo Zhu, Jeremiah Z Liu, Stephen F Cauley, Bruce R Rosen, and Matthew S Rosen. Image reconstruction by domain-transform manifold learning. *Nature*, 555(7697):487–492, 2018.

[8] Guang Yang, Simiao Yu, Hao Dong, Greg Slabaugh, Pier Luigi Dragotti, Xujiong Ye, Fangde Liu, Simon Arridge, Jennifer Keegan, Yike Guo, et al. DAGAN: Deep de-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction. *IEEE transactions on medical imaging*, 37(6):1310–1321, 2017.

[9] Kerstin Hammernik, Teresa Klatzer, Erich Kobler, Michael P Recht, Daniel K Sodickson, Thomas Pock, and Florian Knoll. Learning a variational network for reconstruction of accelerated mri data. *Magnetic resonance in medicine*, 79(6):3055–3071, 2018.

[10] Chen Qin, Jo Schlemper, Jose Caballero, Anthony N Price, Joseph V Hajnal, and Daniel Rueckert. Convolutional recurrent neural networks for dynamic mr image reconstruction. *IEEE transactions on medical imaging*, 38(1):280–290, 2018.

[11] Morteza Mardani, Enhao Gong, Joseph Y Cheng, Shreyas S Vasanawala, Greg Zaharchuk, Lei Xing, and John M Pauly. Deep generative adversarial neural networks for compressive sensing mri. *IEEE transactions on medical imaging*, 38(1):167–179, 2018.

[12] Simon Arridge, Peter Maass, Ozan Öktem, and Carola-Bibiane Schönlieb. Solving inverse problems using data-driven models. *Acta Numerica*, 28:1–174, 2019.

[13] Jonas Adler and Ozan Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33(12):124007, 2017.

[14] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[15] James Nechvatal, Elaine Barker, Lawrence Bassham, William Burr, Morris Dworkin, James Foti, and Edward Roback. Report on the development of the advanced encryption standard (aes). *Journal of Research of the National Institute of Standards and Technology*, 106(3):511, 2001.

[16] Whitfield Diffie and Martin E Hellman. Privacy and authentication: An introduction to cryptography. *Proceedings of the IEEE*, 67(3):397–427, 1979.

[17] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223, 2011.