Unsupervised Detection of Lesion in Brain MRI using Context Encoding AutoEncoders

Giuseppe Paolini
Politecnico di Milano
giuseppe.paolini@polimi.it

Giacomo Boracchi
Politecnico di Milano
giacomo.boracchi@polimi.it

Matteo Matteucci
Politecnico di Milano
matteo.matteucci@polimi.it

Abstract

Today machine learning is widely used in the medical field to improve and facilitate the work of radiologists in visualizing or segmenting brain lesions. But machine learning models need to be trained to perform specific functions such as classification or segmentation. Collecting and processing huge amounts of data consisting of annotations made by doctors by hand is very time-consuming and expensive. Moreover, for however much data one can collect, it is really difficult to describe the totality of possible lesions and abnormalities within the human brain. To solve this problem, machine learning models are trained in an unsupervised manner on unlabeled datasets of images of patients with no abnormalities so that the models learn the normal distribution and no longer recognize tumors specifically. In this research, various models were used to find lesions or abnormalities within brain MRIs. Specifically, Convolutional AutoEncoders (CAEs), Variational AutoEncoders (VAEs) and Vector Quantized Variational AutoEncoders (VQ-VAEs) were used. The autoencoders were trained with context encoding (CE) logic in order to obtain a more calibrated and interpretable reconstruction error. The BrainMetShare dataset containing images of 156 patients with lesions from 2 mm to more than 4 cm was used in this paper to train and test the models. BrainMetShare includes for each patient 4 different 3D sequences that vary according to the MRI acquisition method: (T1 spin-echo pre-contrast, T1 spin-echo post-contrast, T1 gradient-echo post-contrast) and the corresponding doctor segmentation.

1. Introduction

The main goal of anomaly detection is to identify examples that deviate from what is typical or expected. Anomaly detection in images is considered a fast growing area of research with numerous applications in different fields such as video surveillance, agriculture and industry. In addition, since machine learning models have achieved extraordinary results in anomaly detection, many researchers have also begun to use neural networks in the medical field for detection of lesions, tumors, or cancer in brain images. Medical images, specifically magnetic resonance imaging (MRI), provide the necessary observations for this purpose. Radiologists use these technologies in decision-making processes to reduce their workload and to increase the accuracy of theri diagnoses. Neural networks require a large amount of data to be trained on in order to optimize their numerous parameters. This is crucial to avoid overfitting and to keep model generalization high. Such necessity is a problem since aggregating and accumulating high quality medical images with associated labels is very expensive and time consuming. In addition, obtaining all images of each specific abnormality to completely describe its distribution is nearly impossible. Due to these complications, supervised learning methods are not often used. In fact, most researchers train their models on unlabeled datasets in an unsupervised manner, with the purpose of learning the probability distribution of healthy data. Once obtained, it is used to identify anomalous data.

The main contribution of this paper are:

- Replications of the papers [9], [2] using their 4 models that are Variational AutoEncoders (VAE) and the two Context Encoding Variational Autoencoders (CE-VAE).
- Extensions of papers [9], [2] by applying their concept of Denoising AutoaEncoders (DAE) to two other types of AutoEncoders: Convolutional AutoEncoder and Vector Quantized Variational Autoencoders (VQ-VAE)

2. Related work

Over the past few years, most anomaly detection techniques on brain tumors have been based on supervised methods using handmade features. However, collecting images of lesions with their respective segmentations is very

expensive and time consuming because it requires the radiologist to check each abnormal image by contouring it pixel by pixel. In addition, obtaining all images of each specific abnormality to fully describe its distribution is nearly impossible. All this makes supervised anomaly detection methods not robust and lacking in generalization. Nowadays, most popular technologies use semi-supervised learning methods. These models are trained on healthy images of patients thus learning the distribution of in-liers. In this way when an image with lesion is given as input to the model the latter will be reconstructed with a larger error than a healthy image thus being able to distinguish out-liers and inliers. Recently, GANs [8] neural networks have been much used in anomaly detection of brain tumors. They attempt to learn the distribution of in-liers by training two neural networks, discriminator and generator, in parallel. The generator tries to reconstruct realistic images of healthy brains, and the discriminator tries to figure out whether the image it analyzes is a generated image or an original image. Once the training is finished, theoretically, the generator will be so good at generating synthetic images that it will be able to fool the discriminator. During the inference step, only the generator is used, which should have learned well the distribution of healthy images. Variational Auto-Encoder (VAE)s [3] are also widely used for anomaly detection, such as GAN networks they try to approximate the distribution of healthy data by variational inference. They consist of an encoder that approximates the posterior distribution in the latent space and a decoder that reconstructs the image. Some researchers have recently been working on improving GANs and VAEs [4, 1]. For example Schlegl et al. proposed AnoGAN to detect outliers in the learned feature space of the GAN [6], then, they also presented fast AnoGAN that can efficiently map query images onto the latent space [5].

3. Proposed approach

These are the various AutoEncoders used in this paper:

3.1. AutoEncoder (AE)

An AutoEndoer is trained to reconstruct the input image from latent space z. It consists of two parts, the encoder and the decoder. The encoder takes as input the image x and outputs a vector of reduced size called latent space $z = enc(x)$, the decoder receives as input the latent space z and outputs the reconstructed input image $\hat{x} = dec(z)$. In Convolutional AutoEncoders, the encoder and decoder consist of convolutional layers, and by identifying the parameters of the encoder and decoder with θ and λ, respectively, we can summarize the training phase with the following formula:

Usually, the reconstruction error is given by the mean-squared error $\mathcal{L}_{rec}(x, \hat{x}) = \mathcal{L}_{MSE}(x, \hat{x}) = ||x - \hat{x}||^2$. The main characteristic of AutoEncoder is to be able to map the input data into the latent space containing the main features

$$\min_{\theta,\gamma} \sum_{x} L_{rec}\,(x, \hat{x}),\ \text{with}\ \hat{x} = dec_{\gamma}(enc_{\theta}(x)).$$

and from there be able to reconstruct an output image as faithful as possible to the input image.

3.2. Variational Autoencoder (VAE)

Variational Autoencoders belongs to the class of generative models which use a probabilistic way to represent observations in the latent space. In contrast to the AE encoder, the VAE describes the probability distribution for each variable in the latent space. It outputs two vectors: a vector of means of the data samples and a vector of standard deviations. Encoding is generated through these distributions. The network is trained by optimising the evidence lower bound (ELBO) which is defined as follows:

$$\log p(x) \geq L = -D_{KL}(q(z \mid x) \| p(z)) + E_{q(z|x)}[\log p(x \mid z)]$$

$P(z)$ is the priority distribution of the latent variable, $q(z|x)$ is the approximate inference model and $p(x|z)$ is the generative model. Both $q(z|x)$ and $p(x|z)$ are diagonal normal distributions. By maximising the $ELBO$, the probability distribution approximates the true distribution of the data and allows us to obtain a probability estimation for a sample of data.

3.3. Denoising Autoencoder (DAE)

Denoising AutoEncoders are trained on a data sample to which noise has been added. Usually this noise is additive Gaussian noise, i.e $\widetilde{x} = x + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$.

In this way, the AutoEncoder learns to reconstruct the original image from the perturbed image without noise, resulting in a much more robust model that is invariant to the perturbation of the input data. **Context Encoders** are a special type of DAE where instead of using additive Gaussian noise, quadrangular patches are used to mask various areas of the input data.

3.4. Vector Quantize Variational Autoencoder (VQ-VAE)

The VQ-VAE is also part of the generative models and extends the standard autoencoder by adding a discrete component to the network called codebook. The latter is used to quantize the $prior(x)$ of the AutoEncoder in a discrete manner and is composed of K vectors each of dimension D. The posterior in VQ-VAE is defined as:

The embedding in the codebook closest to the output of the encoder is given as input to the decoder. The decoder then reconstructs the distribution $p(x|q(x))$.

$$q(z = k|x) = \begin{cases} 1 & \text{for } k = \text{argmin}_j \|z_e(x) - e_j\|_2, \\ 0 & \text{otherwise} \end{cases},$$

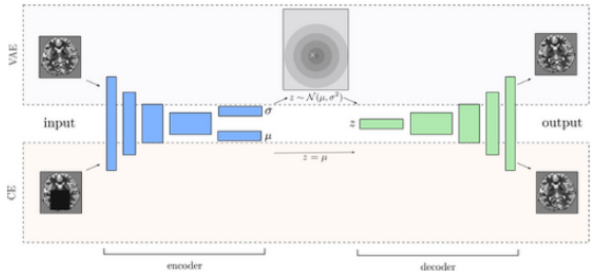
$$L = log(x \mid z_q(x)) + \| sg[z_e(x)] - e \|_2^2 + \| sg[e] - z_e(x) \|_2^2$$

The loss function VQ-VAE is then defined as:

The first term is the MSE between the input image and reconstructed image. The second term indicates how much the codebook vectors differ from the vector generated by the encoder and is used to bring the codebook vectors as close as possible to the encoder output. The third term is similar to the second but is intended to solve the inverse problem, which is for the encoder to generate vectors as close as possible to the nearest codebook.

3.5. CE-VAE

The authors of [9] and [2] combined the Variational AutoEncoder with the Context Encoding AutoEncoder to identify lesions in brain MRI.



As shown in the figure, the idea is to feed the model with two images: the original image and the same image with the addition of a square patch of random size. During training, the model minimises the following loss function::

$$L_{ceVAE} = L_{KL}(f_\mu(x), f_\sigma(x)^2) + L_{rec_{VAE}}(x, g(z)) + L_{rec_{CE}}(x, g(f_\mu(\tilde{x}))),$$

Where the first term is the KL-loss, the second term is the reconstruction error between the original input image and the reconstructed image, and finally the third term is the reconstruction error between the perturbed image and its reconstruction.

The combination of CE and VAE has a regularising effect that prevents the posterior from collapsing and better represents the semantics in the input data.

3.6. CE-AE and CE-VQ-VAE

The experiment performed in this paper was to combine the Context-Encoding part with a Convolutional AutoEncoder and the Vector Quantized Variational Autoencoder [7]. To train the new models, the reconstruction error term between the perturbed input image and the reconstructed image was added respectively to the loss function of the Convolutional AutoEncoder and to the loss function of the Vector Quantized Variational Autoencoder.

Dataset The dataset used for this paper is different from the one used in the papers [9] and [2]. It is called BrainMetShare, it is a dataset containing 256x 256 MRI images of the entire brains of 156 patients with at least one brain metastasis. The mean age of the patients was 63±12 years (range: 29-92 years). 41% of patients had 1 to 3 metastases, 30% had 4 to 10 metastases, and 29% had more than 10 metastases. Lesion sizes ranged from 2 mm to more than 4 cm. The dataset includes for each patient 4 different 3D sequences varying according to the MRI acquisition method: (T1 spin-echo pre-contrast, T1 spin-echo post-contrast, T1 gradient-echo post-contrast , T2 FLAIR. In addition to these 4 types of images, we have for each patient a binary mask which is the segmentation of the tumor. One important thing to say is that the segmentation of the radiologists was done on images T1 gradient-echo post contrast and that the models used in this project were trained on the T2 flair images. For this reason the anomaly maps generated by the models are a little different from the actual tumors in the binary masks created by the radiologists. In addition, all images were skull-stripped to extract the brain tissues.

Preprocessing Before training the various models, the following pre-processing was applied to the images in the dataset: each patient was divided into healthy and diseased brain slices, the intensities of the images were all scaled in the same range of values between [0,255], each image was centred patient by patient with the corresponding lesion masks, the images were cropped from [256,256] to [196x196] removing unnecessary black background pixels for training and thereby decreasing the weight of the various models.

Network architectures For the encoding and decoding networks of the first VAE and AE [9], this work uses fully convolutional networks with five layers 2D-Conv-Layer and 2D-Transposed-Conv-Layer respectively with CoordConv, kernel size 4 and stride 2 and each layer followed by a LeakyReLU non-linearity. The encoder and decoder of the first VAE and AE are symmetrical with feature maps of 16, 64, 256, 512 and a latent variable size of 512, while the second VAE [2] has 64, 128, 256 feature maps and a latent variable size of 256. The VQ-VAE encoder and decoder are also symmetrical with 128, 256, 256 feature maps and a latent embedding space of dimension [K, D] where K is the number of embeddings (512) and D is the dimensionality of

each latent embedding vector (256). The encoder will map the input into a sequence of discrete latent variables, while the decoder will attempt to reconstruct the input from these latent sequences.

Training All models were trained on images without anomalies. More precisely, the dataset was split into a training set of 5450 images, a validation set of 1362 images and a test set of 419 images of which 215 had anomalies and 204 had no anomalies. The models were trained for 200 epochs with a learning rate of 0.0001 and a batch size of 64.

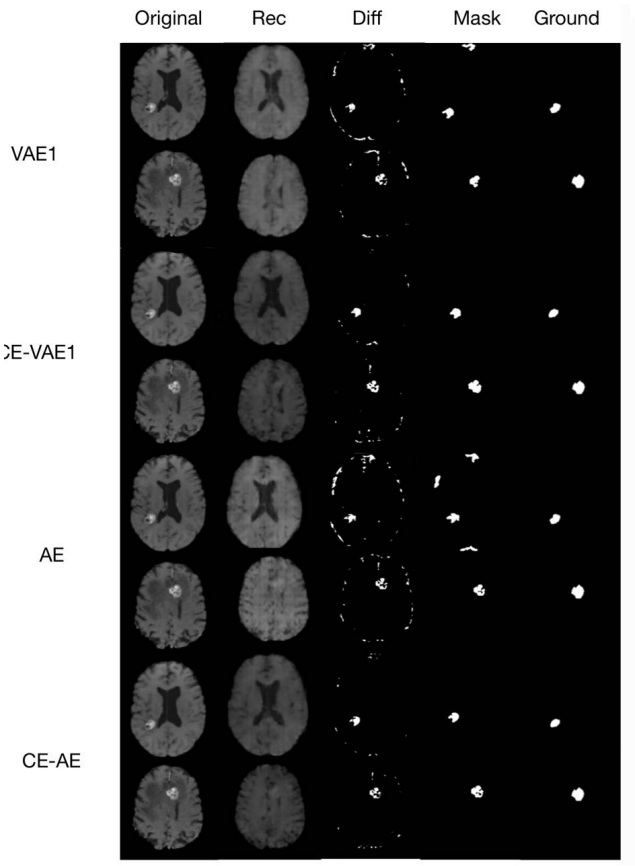
Postprocessing After subtracting the reconstructed image from the input's image, the initial difference image is obtained. It has values between zero and one, where one implies anomaly while zero means no anomaly. A threshold is then applied to the difference image in order to obtain all pixels with a value of 1 or 0. Now, for each anomaly contained in every difference image, the area of the maximum anomaly was calculated, which will then be used to perform further filtering. The following operations were applied to the binary map: morphological closure, binary fill holes, morphological area opening with an adaptive threshold image by image (all the anomalies with an area under the 60% of the area of the largest anomaly in the current image are deleted) and dilation. This results in a binary anomaly mask.

Results and discussion. The Dice Score was used as a metric to evaluate the models.

	Dice Score
VAE1	0.317
CE-VAE2	0.463
AE	0.302
CE-AE	0.405
VQ-VAE	0.169
CE-VQ-VAE	0.083
VAE2	0.302
CE-VAE2	0.326

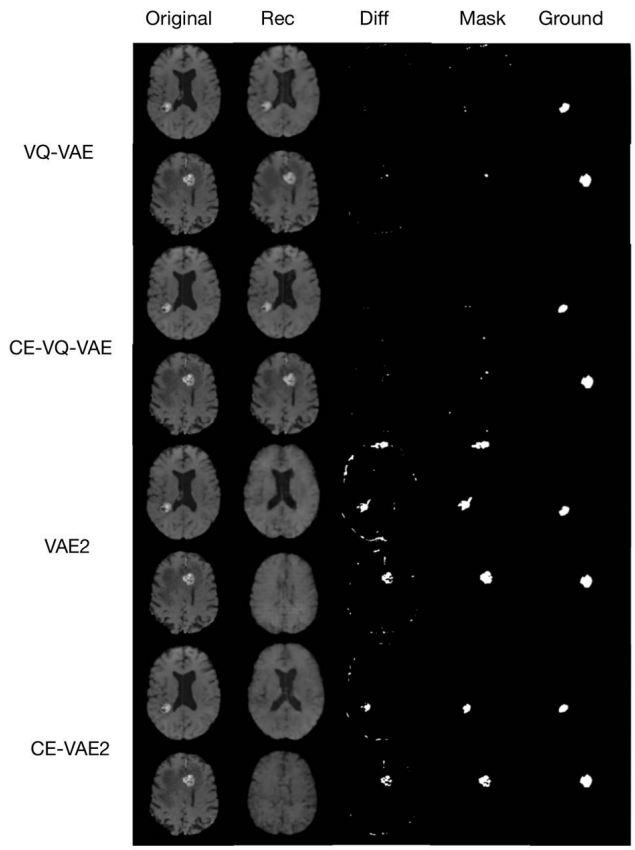
VAE1 and CE-VAE1 [9] are the models with a larger architecture while VAE2 and CE-VAE2 are the models with a smaller architecture [2]. As can be seen from the results, the combination of the context-encoding logic with the various autoencoders increases the Dice Score in almost all models. The figure shows that the dice of the auto-encoders combined with context-encoding is always higher, and the reconstructed images are sharper than the corresponding ones of the auto-encoders without context-encoding part. This, however, does not apply to the VQ-VAE, which, by reconstructing abnormalities in the output images, fails to identify tumours.

Regarding the overall Dice scores of the various models, it is important to remember that the segmentations of the radiologists do not perfectly match the anomalies on the images on which the model is evaluated. In most cases, the pixels of the doctors' masks are always more numerous than the actual pixels perceptible from the images. This is because we remember that while the segmentations were calculated on MRIs acquired by T1 gradient-echo post, the images on which the models are trained are T2 flair.



4. Conclusion

This work aims to demonstrate that the use of Context-Encoding in the loss of the various AutoEncoders as additional term, results in a better reconstruction of the generated image. In the case of anomaly detection, the contribution of Context-Encoding also leads to an improvement in the model's ability to identify anomalies in images such as brain MRIs.



References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan, 2017. 2

[2] S. Chatterjee, A. Sciarra, M. Dünnwald, P. Tummala, S. K. Agrawal, A. Jauhari, A. Kalra, S. Oeltze-Jafra, O. Speck, and A. Nürnberger. Strega: Unsupervised anomaly detection in brain mris using a compact context-encoding variational autoencoder. 2022. 1, 3, 4

[3] D. P. Kingma and M. Welling. Auto-encoding variational bayes, 2013. 2

[4] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015. 2

[5] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth. f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks. *Medical Image Analysis*, 54:30–44, May 2019. 2

[6] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery, 2017. 2

[7] A. van den Oord et al. pytorch-vq-vae. url https://github.com/zalandoresearch/pytorch-vq-vae, 2019. 3

[8] zalandoresearch. Generative adversarial networks. *GitHub repository*, Nov. 2019. 2

[9] D. Zimmerer, S. A. A. Kohl, J. Petersen, F. Isensee, and K. H. Maier-Hein. Context-encoding variational autoencoder for unsupervised anomaly detection, 2018. 1, 3, 4