A SELF-SUPERVISED PRE-TRAINING FRAMEWORK FOR VISION-BASED SEIZURE CLASSIFICATION

Jen-Cheng Hou[1]*, Aileen McGonigal*[2]*, Fabrice Bartolomei*[2]*, Monique Thonnat*[1]

[1]University of Côte d'Azur, INRIA
[2]Aix-Marseille University

ABSTRACT

Seizure events feature temporary abnormalities in muscle control or movements. They are usually caused by excessive neuronal activities in the brain, and are called epileptic seizures (ES). Nevertheless, not all seizures are epileptic in origin. Some are caused by psychological reasons, and such type of seizures are called psychogenic non-epileptic seizures (PNES). We propose a method to classify ES and PNES based on clinical signs in the seizure videos. In particular, inspired by BERT, we propose a Transformer-based framework that pre-trains on large unlabeled clinical videos, and then we fine-tune the pre-trained model for seizure classification with a minimum modification. We conduct a leave-one-subject-out (LOSO) validation on our dataset. The F1-score and accuracy are 0.82 and 0.75, respectively. To our knowledge, the proposed approach is the first attempt to use large unannotated data and learn useful representations for downstream tasks in the field of video based seizure analysis.

Index Terms— Seizure Video Analysis, Self-Supervised Learning, Computer Vision, Deep Learning

1. INTRODUCTION

Epilepsy is one of the most prevalent neurological disorders, affecting nearly 1% of the population worldwide. It is characterized by recurrent seizures, which are caused by abnormal, excessive neuronal activity in the brain [1]. Nevertheless, not all seizures are epileptic in origin. Some are caused by psychological reasons, and such type of seizures are called psychogenic non-epileptic seizures (PNES), which are not associated with an epileptic discharge. To determine if a seizure is caused by epileptic discharges, Video-EEG monitoring is used to check the existence of simultaneous culprit brain EEG rhythms during the seizure. Despite the different cause of epileptic seizures (ES) and PNES, these two types of seizure could be similar in terms of the semiology, i.e. the clinical signs. Even for experienced neurologists, it could be challenging sometimes for them to correctly distinguish them. In addition, the evaluation could be subject to inter-observer variability. Hence, a computer-aided diagnosis is naturally considered as a way to improve the quality of the assessment.

Deep learning is a promising approach to analyze seizure videos given its ability to tackle different complex problems, such as computer vision [2], speech recognition [3], and natural language processing (NLP) [4]. Nevertheless, deep learning usually requires a large volume of annotated data for training, and in medical domains, large labeled data is usually costly to get. Inspired by BERT [5], a self-supervised learning (SSL) framework using large unlabeled data to learn useful features for downstream tasks, we investigate if such paradigm can be applied into vision-based seizure analysis.

Specifically, we collect voluminous unlabeled data for pre-training a Transformer model [6], and fine-tune the pre-trained model for seizure classification (ES v.s. PNES). The unlabeled data for pre-training are clinical videos without labels, and we call them as "contextual videos". The contextual videos are recorded in the EEG-Video monitoring unit, as like the labeled seizure videos. They contain daily behaviors of patients and possibly other associated people in the unit. The videos are expected to provide visual information of the context where seizures are recorded. Such data is easily accessible and thus suitable for SSL-based pre-training.

We pre-train the encoder of Transformer with a denoising objective, where the input video is corrupted and the model aims to recover the visual features of the original frame sequence. The learning objective is inspired by BART [7], a denoising sequence-to-sequence model for NLP. In our study, after pre-trained on the contextual videos, the model is fine-tuned for classification, where the input video is not corrupted. To our knowledge, the proposed approach is the first attempt to use large unannotated data for learning useful representations for downstream tasks in the field of video based seizure analysis. We consider it an inevitable trend for medical applications, as large-scale annotations for medical data are usually difficult to obtain.

2. RELATED WORK

Although the incentive to use deep learning for seizure video analysis given its capability in computer vision is natural, studies on the topic are still very limited compared to those using EEG signals. Karácsony et al. [8] use pre-trained

spatiotemporal features along with a long short-term memory (LSTM) classifier to distinguish temporal lobe epilepsy (TLE) and frontal lobe epilepsy (FLE). Ahmedt-Aristizabal et. al. [9, 10] propose multi-modal approaches to classify patients with mesial temporal lobe (MTLE) and extra-temporal lobe (ETLE) epilepsy, with a focus on facial expressions and pose dynamics. The approaches are based on convolutional neural networks (CNN) and LSTM. Achilles et. al [11] use CNN to analyze infrared and depth videos for epilepsy classification, however the method does not leverage temporal consistency. To distinguish ES from PNES, Ahmedt-Aristizabal et. al. [12] and Hou et al. [13] feed both appearance and keypoints (e.g. joint locations) into neural nets. Nevertheless, with the inclusion of joint information, the performance may be vulnerable to conditions where joint estimation are poorly performed, which is not rare due to the frequent occlusion occurred in seizure videos. In this study, we attempt to use only appearance for comparable results on the same task.

Recent research utilizing large data via pre-training Transformer for medical applications are mainly related to vision-and-language (VL) learning. Moon et al. [14] pre-trains a Transformer-based model on aligned X-ray images and associated reports for learning joint VL representations in the medical domain. The downstream tasks include both comprehension and generation tasks. Li et al. [15] and Monajatipoor et al. [16] use off-the-shelf pre-trained VL models, whose VL data for pre-training are not medical ones, to directly learn multimodal representation on radiographs and associated reports.

Our work introduces the Transformer-based pre-training paradigm into vision-based seizure analysis, and we consider it a promising way to benefit from large unannotated data in the field.

3. APPROACH

3.1. Data collection

3.1.1. The contextual video dataset

Here we introduce the clinical videos used for pre-training. Each patient stays one week in the epilepsy monitoring units (EMUs) of Marseille University Hospital where both Video and EEG monitoring are performed. So there are hundreds of hours of video recordings for each patient. If seizures occurs during the session, the medical staff will identify and extract the video segments afterward, and then save them in the database of the hospital. As for parts where no seizure events are involved, these recordings will be erased weeks later, because they could be bulky for storage yet not informative in terms of medical viewpoints.

Nevertheless, in terms of deep learning, these 'meaningless' seizure-free videos might be useful. The reason is that they can provide the visual information of the surroundings/environments of how seizures are captured. In addition, we can have a large quantity of them, and the mainstream deep learning/machine learning models usually favors big data. Given that, we intentionally collected more than 1000 hours seizure-free videos, and we call them as 'contextual videos' in this research. The behavior in these contextual videos can be as diverse and natural as those in daily routines, such as eating, sleeping, chatting with their families, and interaction with clinicians. The recording conditions include both daytime and night. Some selected samples are shown in Fig. 1.

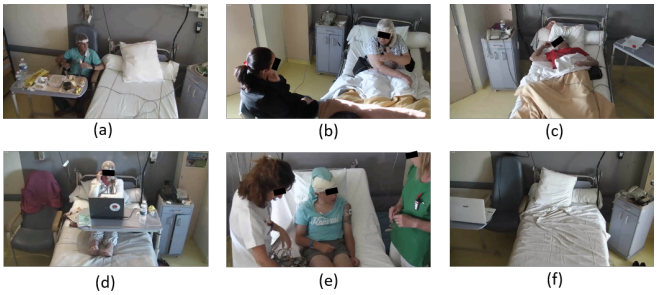

Fig. 1. The contextual videos used for pre-training cover the daily behaviors of patients in the Video-EEG monitoring unit, except for the onset seizure events. They include (a) eating food, (b) interaction with their family, (c) sleeping, (d) using laptops/smartphones, (e) being checked by the clinical staff. The empty settings are possibly recorded if the patients leave the room, like (f). Night conditions are also included.

Type	ES	PNES
Number of patients	52	29
Number of seizures	235	48
Average seizure duration [sec]	45.4	52.9
Min. seizure duration [sec]	7.5	12.1
Max. seizure duration [sec]	150.2	119.4

Table 1. Some statistics of our seizure video dataset.

3.1.2. The seizure video dataset

Our seizure videos consist of ES and PNES. After segmenting out each seizure for a new video clip, we saved the new clip as its original format. We then converted the trimmed clips into image sequences for each clip at a frame rate of 25 frames per second, while keeping the resolution unchanged. We resize the aspect ratio until the frames were fed into the developed models. Table. 1 shows some statistics of our seizure video dataset.

3.2. Pre-training and fine-tuning the model

Inspired by BART [7], another Transformer-based SSL model for NLP, which corrupts input text with an arbitrary noising

function and makes Transformer to reconstruct the original text, we include this concept of denoising objective into our model in the pre-training phase, as shown in Fig. 2. From the contextual video dataset $\mathbb{D}_c$, for each video $V_c \in \mathbb{D}_c$, we have ordered image sequence as $V_c = (m_c^1, \ldots, m_c^K)$. Two noising functions are applied on V_c. We change the sequence ordering by permuting V_c, and then randomly mask out some frames, resulting in a noised version of V_c, denoted as $\widetilde{V_c}$. The pre-training objective is to regress the Transformer output of each frame in $\widetilde{V_c}$ to the visual features of V_c. The L2 regression loss is formulated as:

$$L(\theta) = E_{v_c \sim \mathbb{D}_c} \sum_{i=1}^{K} \|h_\theta(\widetilde{v_c}^{(i)}) - r(v_c^{(i)})\|_2^2 \quad (1)$$

Where θ is the trainable parameters of the Transformer, and its output is expressed as h_θ. We take ResNet-152 [17] as our CNN backbone to generate visual features. The ResNet-152 is pre-trained on ImageNet [18], and we remove the last classification layer to generate a 2048-d feature. We denote it as r as the function for frame descriptor.

After pre-training the Transformer on the contextual video dataset $\mathbb{D}_c$ with the defined objective loss as equation 1, we add a fully-connected layer (FC) on top of our pre-trained Transformer for classification, as shown in Fig. 3. Then fine-tune the whole model on the target dataset $\mathbb{D}_s$, which contains seizure videos for seizure type classification. For each seizure video $V_s \in \mathbb{D}_s$, we have an uncorrupted image sequence as input to Transformer as $V_s = (m_s^1, \ldots, m_s^N)$, with the corresponding binary seizure type labels $y_s \in \mathbb{L}$. In the fine-tuning phase, the seizure classification task is optimized based on the standard binary cross-entropy loss as

$$\begin{aligned} L_{CE} = E_{v_s \sim \mathbb{D}_s}(y_s \cdot log(Softmax(FC(h_\theta(v_s)))) + \\ (1 - y_s) \cdot log(1 - Softmax(FC(h_\theta(v_s))))) \end{aligned} \quad (2)$$

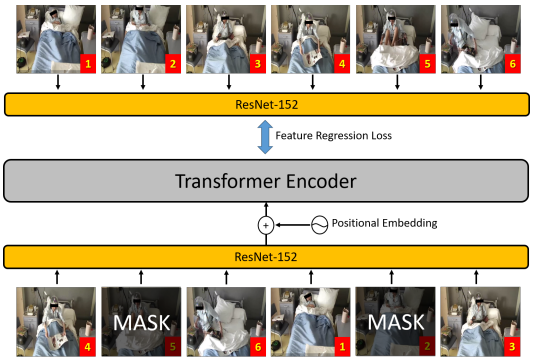

Fig. 2. SSL-based pretraining on contextual videos: The input sequence is the "noised" version of the target sequence, where random frames are masked out and permutation is applied. We pretrain the encoder of Transformer to reconstruct the corresponding visual features.

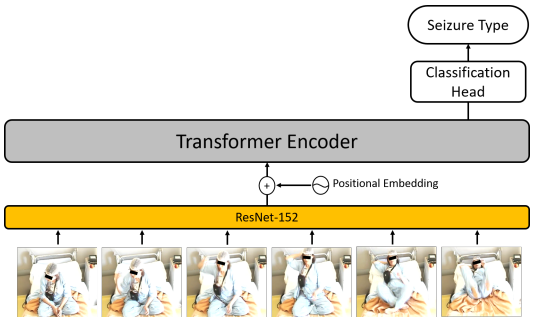

Fig. 3. Finetuning phase for seizure type classification: In the fine-tuning phase, an uncorrupted seizure video sequence is fed into the pretrained model. A classification head, i.e. fully-connected layer, is added on top of the pretrained model for the classification task.

4. EXPERIMENTATION

In this section, we give the details of the implementation of the experimentation.

Dataset and pre-processing

For pre-training the Transformer model, there are about 13k 10-second clips in the contextual video dataset $\mathbb{D}_c$, resulting in a total 36 hours of clip duration. We convert the clip into image sequence at 25 fps. We resize the frame into a 128 $\times$ 171 dimension, and while generating the training mini-batch, a random crop of 112 $\times$ 112 is applied on the frames.

In the fine-tuning step for seizure type classification, seizure dataset $\mathbb{D}_s$ is used. $\mathbb{D}_s$ covers all the seizure videos. In other words, $\mathbb{D}_s$ contains 283 trimmed seizure videos, and among them, 235 videos belong to ES, and 48 videos are PNES. A total of 81 patients are involved, in which the ES and PNES class has 52 and 29 patients, respectively. The length of seizure videos ranges from 7 seconds to 150 seconds. After converting the seizure videos into image sequence, we detect the target patient with a SSD detector [19] pre-trained on our seizure video dataset, as shown in Fig 4. The Intersection over Union (IoU) is 0.89. The cropped region is then resized to a 128 $\times$ 171 dimension, and a center crop of 112 $\times$ 112 is applied on the frames. Normalization of image tensors are implemented by subtracting the mean and divided by the standard deviation across each channel.

Specification of the Transformer

Regarding the specification of the Transformer used in this work, the number of attention head h is 8. Model dimension d_{model} is set as 1024. The maximum position is set as 256. The number of encoder layer is 6. The number of total trainable parameters is about 78M.

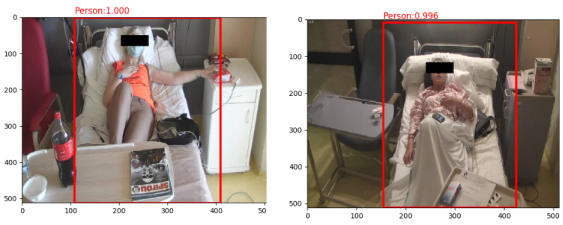

Fig. 4. Patients in image frames are detected and cropped before feeding into the pre-trained Transformer model for the downstream seizure classification task.

Corrupt the input for pre-training

The video sequence as input for the Transformer while pre-training is corrupted, in terms of frame ordering and information masking. The input length of the model is set as 256. A span of consecutive 30 frames are randomly selected and relocated for frame permutation. As for frame masking, following BERT [5], we replace 15% of frames with visual `MASK` tokens. The visual `MASK` and `PAD` tokens are tensors with the shape of image tensors but filled with fixed values of -1.0 and 0.0, respectively.

Training details

We pre-train the Transformer for 60 epochs. The initial learning rate is 0.01, with a linearly decreased scheduler. Weight decay for pre-training is set as 0.0001. The pre-training process takes 670 gpu-hours (roughly 14 hours on 48 V100 gpus across 6 nodes). We adopt AdamW [20] as the optimizer. As for fine-tuning the Transformer, we train it for 50 epochs. Batch size is 16. Except for setting the initial learning rate as 0.005, other training settings are the same as those in the pre-training phase. We test the whole videos by temporally averaging the predictive results. A dropout rate of 0.5 in the final classifier layer is set. In addition, to mitigate the imbalanced dataset, a class weight (reciprocal of the number of video clips per class) is added in the cross entropy loss. The implementation of the Transformer model is based on the Huggingface library [21].

Experimental results

We perform a leave-one-subject-out (LOSO) validation. The F1-score and the accuracy are 0.82 and 0.75, respectively. As shown in Table 2, our results are comparable to other state-of-the-art seizure classification tasks on distinguishing ES from PNES. In particular, the dataset used in [13] is the subset of the one used in this study. We gain a performance boost while including more seizure videos. This indicates our proposed Transformer-based pre-training approach can learn robust and generalizable features for the downstream task. The video-wise confusion matrix is shown in Table 3.

Method	Patients /Videos	Performance
A.-Aristizaba et al. (2019) [12]	35/50	Accuracy (Landmark-based/Region-based): 0.68/0.79
Hou et al. (2021) [13]	34/61	F1-score: 0.76 Accuracy: 0.72
Ours	81/283	F1-score: 0.82 Accuracy: 0.75

Table 2. Comparison of deep learning-based seizure classification task on distinguishing ES from PNES. The number of patients and seizure videos in our study are the largest among the related works. Our experimental results are shown to be comparable to those works.

	(predicted) ES	(predicted) PNES
(true) ES	181	54
(true) PNES	15	33

Table 3. Confusion matrix of the video-wise classification results by leave-one-subject-out validation.

5. CONCLUSION

In this study, we propose a Transformer-based self-supervised pre-training framework for learning features suitable for the downstream task, i.e. classifying ES and PNES videos. The paradigm aligns with the research direction of self-supervised pre-training that takes advantage of large unannotated data and learns useful representations from it for downstream tasks. This may be especially favored for medical applications where data annotations are usually costly. In our work, a Transformer-based model is pre-trained on a large volume of contextual videos with denoising pre-training objectives. By simply fine-tuning the pre-trained model with a minimum model modification, the experimental classification results can compete with methods from other state-of-the-art works for the same task. To our knowledge, this is the first deep learning work exploiting large unlabeled data for facilitating vision-based seizure analysis. We hope our study can inspire the research community regarding seizure video analysis to rethink how we can benefit from large unannotated data.

6. ACKNOWLEDGEMENTS

This work was granted access to the HPC resources of IDRIS under the allocation 2020-AD011011721R1 made by GENCI.

7. REFERENCES

[1] R. S. Fisher, W. E. Boas, W. Blume, C. Elger, P. Genton, P. Lee, and J. Engel, "Epileptic seizures and epilepsy: Definitions proposed by the international league against epilepsy (ILAE) and the international bureau for epilepsy (IBE)," *Epilepsia*, vol. 46, no. 4, pp. 470–472, Apr. 2005.

[2] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Computational Intelligence and Neuroscience*, vol. 2018, pp. 1–13, 2018.

[3] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, "Speech recognition using deep neural networks: A systematic review," *IEEE Access*, vol. 7, pp. 19143–19165, 2019.

[4] A. Torfi, R. A. Shirvani, Y. Keneshloo, N. Tavaf, and E. A. Fox, "Natural language processing advancements by deep learning: A survey," *arXiv:2003.01200*, 2020.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, June 2019.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[7] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, July 2020.

[8] T. Karacsony, A. M. Loesch-Biffar, C. Vollmar, S. Noachtar, and J. P. S. Cunha, "A deep learning architecture for epileptic seizure classification based on object and action recognition," in *ICASSP*. May 2020, IEEE.

[9] D. Ahmedt-Aristizabal, C. Fookes, S. Denman, K. Nguyen, T. Fernando, S. Sridharan, and S. Dionisio, "A hierarchical multimodal system for motion analysis in patients with epilepsy," *Epilepsy & Behavior*, vol. 87, pp. 46–58, Oct. 2018.

[10] D. Ahmedt-Aristizabal, K. Nguyen, S. Denman, S. Sridharan, S. Dionisio, and C. Fookes, "Deep motion analysis for epileptic seizure classification," in *EMBC*, July 2018.

[11] F. Achilles, F. Tombari, V. Belagiannis, A. M. Loesch, S. Noachtar, and N. Navab, "Convolutional neural networks for real-time epileptic seizure detection," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 264–269, July 2016.

[12] David Ahmedt-Aristizabal, Simon Denman, Kien Nguyen, Sridha Sridharan, Sasha Dionisio, and Clinton Fookes, "Understanding patients' behavior: Vision-based analysis of seizure disorders," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2583–2591, Nov. 2019.

[13] Jen-Cheng Hou, Aileen McGonigal, Fabrice Bartolomei, and Monique Thonnat, "A multi-stream approach for seizure classification with knowledge distillation," in *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. Nov. 2021, IEEE.

[14] J. H. Moon, H. Lee, W. Shin, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *arXiv:2105.11333*, 2021.

[15] Y. Li, H. Wang, and Y. Luo, "A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports," in *BIBM*. Dec. 2020, IEEE.

[16] M. Monajatipoor, M. Rouhsedaghat, L. H. Li, A. Chien, C. C. J. Kuo, F. Scalzo, and K.-W. Chang, "Berthop: An effective vision-and-language model for chest x-ray disease diagnosis," *arXiv:2108.04938*, 2021.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*. June 2016, IEEE.

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR*, 2009.

[19] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *ECCV*, 2016.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," *arXiv:1910.03771*, 2019.