An EEG-based Image Annotation System

Viral Parekh[1], Ramanathan Subramanian[2], Dipanjan Roy[3], and C. V. Jawahar[1]

[1] IIIT Hyderabad, India,
[2] University of Glasgow, Singapore
[3] National Brain Research Centre, Manesar, India

Abstract. The success of deep learning in computer vision has greatly increased the need for annotated image datasets. We propose an EEG (Electroencephalogram)-based image annotation system. While humans can recognize objects in 20-200 milliseconds, the need to manually label images results in a low annotation throughput. Our system employs brain signals captured via a consumer EEG device to achieve an annotation rate of up to 10 images per second. We exploit the P300 event-related potential (ERP) signature to identify target images during a rapid serial visual presentation (RSVP) task. We further perform unsupervised outlier removal to achieve an F1-score of 0.88 on the test set. The proposed system does not depend on category-specific EEG signatures enabling the annotation of any new image category without any model pre-training.

Keywords: EEG, Image annotation, Active learning

1 Introduction

Image annotation is a critical task in computer vision, intended to bridge the **semantic gap** between automated and human understanding via the use of tags and labels. Image annotation is useful for building large-scale retrieval systems, organizing and managing multimedia databases, and for training deep learning models for scene understanding. A trivial way to annotate images is to tag them manually with the relevant labels, but this approach is slow and tedious for huge databases. Therefore, many efforts have been undertaken to address/circumvent this problem. Some methods are completely automatic [1,2,3,4,5], while some others are interactive [6,7,8,9,10]– these approaches have considerably reduced the human effort required for annotation.

Human vision is a very powerful system for object recognition and scene understanding. It is also robust to variations in illumination, scale or pose. We are habitually used to recognizing objects even in cluttered scenes. Humans can identify objects in tens of milliseconds [11,12], but the representation of the perceived information via hand movements or verbal responses for annotation is very slow compared to the processing speed of contemporary digital devices. In this regard, the emerging field of brain-Computer Interfaces (BCI) offers us an innovative way to exploit the power of human brain for data annotation with minimal effort.

Brain-Computer Interfaces rely on various technologies for sensing brain activity such as Electroencephalography (EEG), MEG (Magnetoencephalography), PET (Positron Emission Tomography), SPECT (Single Photon Emission Computed Tomography), fMRI (functional Magnetic Resonance Imaging) and fNIRS (functional near infrared spectroscopy). Among these, EEG provides a high temporal resolution (sampling rate of up to 1 KHz) and adequate spatial resolution (1-2 cm). In this work, we specifically use the portable and easy-to-use consumer grade *Emotiv* EEG device, which enables a minimally intrusive user experience as users perform cognitive tasks, for sensing and recording brain activity. While having these advantages, consumer EEG devices nevertheless suffer from a high signal-to-noise ratio, which makes subsequent data analytics challenging.

In this work, we focus on the annotation of a *pre-selected* object category over the entire image dataset instead of labeling all categories at once. If the images are presented serially in a sequence for annotation, then the task is equivalent to that of *target detection*. Now whenever an image containing a target class instance is observed by the human annotator, an event-related potential (ERP) signature known as **P300** [13] is observed in the EEG data. By examining the EEG signals generated during image presentation, we can discover the images of interest and annotate them accordingly. In this paper, we provide the pipeline and architecture for image annotation via EEG signals.

2 Related work

The use of EEG as an additional modality for computer vision and scene understanding tasks has been explored by a number of works. In [14], EEG signals are used to automate grab cut-based image segmentation. In [15], authors exploit ERP signatures such as P300 for image retrieval. In [16], authors use the N400 ERP to validate tags attached to video content. Emotions from movies and ads are inferred via EEG signals in [17] and [18].

Few studies directly use image category-based EEG signatures for recognizing aspects related to multimedia content as well as users. For example, the authors of [19] use EEG signals to classify images into three object categories– animals, faces and inanimate. In a recent work [20], the authors present how EEG features can be employed for multi-class image classification. Another recent work recognizes user gender from EEG responses to emotional faces [21]. Given the state-of-the-art, the key contributions of our work are we how how (i) the P300 ERP signature can be employed for image annotation; (ii) the model trained for one object category can be directly used for a novel category, and (iii) the image presentation time affects annotation system performance for complex images.

3 System architecture

The proposed image annotation system consists of several components– RSVP generation, EEG data acquisition, EEG pre-processing, (binary) classification and outlier removal. Fig.1 presents an overview of the EEG-based annotation
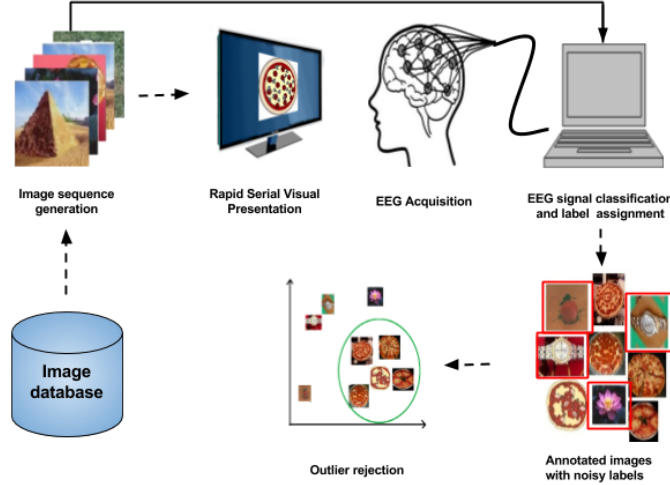
Fig. 1. EEG-based annotation pipeline: An exemplar illustration for the *pizza* object class is presented. Best viewed in color and under zoom.

pipeline. The RSVP generation unit prepares the set of images for viewing, so that a few among those correspond to the target object category. The image sequence is created via random sampling from the whole dataset. A human annotator is then asked to identify the target category images as the sequence is presented rapidly, and the annotator's brain activity is recorded via an EEG headset during the visual recognition task. The compiled EEG data is first pre-processed for artifact removal. Then, the classification unit categorizes the EEG responses into *target* and *non-target* annotations based on P300 patterns. Images classified as *target* are annotated with the target label class. However, this labeling is noisy due to the presence of false positives and imbalance towards the negative (non-target) class. An outlier removal unit finally performs unsupervised dimensionality reduction and clustering to improve the labeling precision.

3.1 Rapid Serial Visual Presentation and Oddball paradigm

Rapid Serial Visual Presentation is popularly used in psychophysical studies, and involves a series of images or other stimuli types being presented to viewers with a speed of around 10 items per second. This paradigm is basically used to examine the characteristics pertaining to visual attention. In RSVP studies, the *oddball* phenomenon [22] is widely used. In the oddball paradigm, a deviant (target) stimulus is infrequently infused into a stream of audio/visual stimuli. For EEG-based annotation, we generated an RSVP sequence by combing a few *target* category images with many *non-target* images via random sampling from the original dataset. Each image in the sequence was then shown to the viewer for 100 ms, and a fixation cross was presented for 2 seconds at the beginning

of the sequence to minimize memory effects and to record resting state brain activity (see Fig.4).

## 3.2 EEG data preprocessing and classification

We used the *Emotiv EPOC* headset to record EEG data. This is a 14 channels (plus CMS/DRL references, P3/P4 locations) Au-plated dry electrode system. For ERP analysis, the Emotiv provides signals comparable to superior lab-grade EEG devices with 32, 64 or 128 channels. The headset uses sequential sampling at 2048 Hz internally which is down-sampled to 128 Hz. The incoming signal is automatically notch filtered at 50 and 60 Hz using a 5th order sinc notch filter. The resolution of the electrical potential is 1.95 μV. The locations for the 14 channels are as per International 10-20 locations as shown in Fig.2.
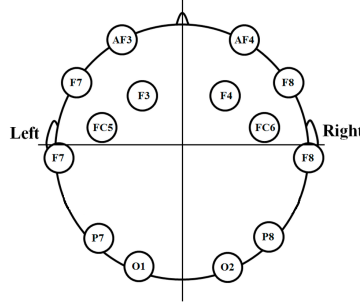
Fig. 2. Sensor configuration: Emotiv electrode locations as per International 10-20 system.

The recorded EEG data is contaminated by various noise undesirable signals that originate from outside the brain. For instance, while recording EEG, one often encounters 50/60Hz power-line noise and artifacts caused by muscle or eye movements. We extracted one second long *epochs* corresponding to each 100 ms long *trial* denoting the presentation of an image, with 128Hz sampling rate. Our EEG preprocessing includes (a) baseline power removal using the 0.5 second pre-stimulus samples, (b) band-pass filtering in 0.1-45Hz frequency range, (c) independent component analysis (ICA) to remove artifacts relating to eye-blinks, and eye and muscle movements. Muscle movement artifacts in EEG are mainly concentrated between 40-100 Hz. While most artifacts are removed upon EEG band-limiting, the remaining are removed manually via inspection of ICA components.

The human brain's response to a stimulus can be measured as a voltage fluctuation resulting from the ionic current within the neurons. The event-related potential is one such measure that is directly related to some motor, cognitive or sensory activation. Out of various ERP components, the P300 signature is
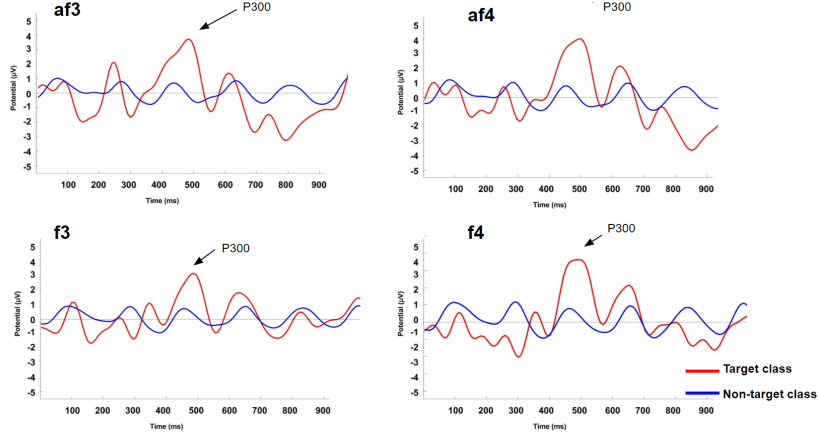
Fig. 3. ERP plots: ERP curves for the Emotiv af3, af4, f3 and f4 channels for *target* (red) and *not-target* (blue) images. P300 signatures are evident for targets but not for non-targets.

commonly elicited in the oddball paradigm where very few targets are mixed with a large number of non-targets. In our experimental setup, we employed a 1:12 ratio for target-to-non-target images. As shown in Fig.3, the P300 ERP signature is observed between 250 to 500 ms post *target* stimulus presentation. Also, the ERP response is significantly different for target and non-target images, and therefore can be exploited for EEG-based image annotation.

We used the Convolutional Neural Network (CNN)-based EEGNet architecture [23] to classify our EEG data based on P300 detection in the RSVP task. The EEGnet architecture consists of only three convolutional layers. All layers use the Exponential Linear Unit (ELU) [24] as nonlinear activation function with parameter $\alpha = 1$. We trained the model using the minibatch gradient descent algorithm with categorical cross-entropy criterion and Adam optimizer [25]. The models were trained on a NVIDIA GEFORCE GTX 1080 Ti GPU, with CUDA 8 and cuDNN v6 using the Pytorch [26] based Braindecode [27] library.

3.3 Outlier removal

We select one category at a time for the annotation task, which results in class imbalance for the RSVP task. The selected object category forms the *target* class, while all other categories collectively form the *non-target* class. Due to this heavy class imbalance and the characteristics of P300 as discussed in Section 5, the false positive rate of the predicted labels is high. Therefore we performed unsupervised outlier removal on the predicted *target* images. Deep learning features have proven advantages over hand-crafted features like SIFT and HoG [28]. We used a pre-trained VGG-19 model [29] to obtain the feature descriptors for the targets.

These feature descriptors provide compact representation of raw images while preserving the information required to distinguish between image classes. Each target image was fed forwarded within the VGG-19 model to obtain the 4096 dimensional feature vectors. Target images need not belong to the image classes on which the model is pre-trained. Then, we perform dimensionality reduction with t-SNE [30] to generate low-dimensional features. The t-SNE algorithm retains the local structure of the data while also revealing some important global structure, and hence it performs better than principal component analysis (PCA) alone.

In our case, we assume that samples from the target class should be close in feature space as compared to non-target samples. By performing a grid search on hyper-parameters, we found that the algorithm works best with perplexity value 20, 50 PCA components and 3-5 output dimensions. Then, we performed k-means clustering for two classes assuming that target class samples will form a cluster distinct from the false positives. Also, since the false positive cluster would contain samples from many categories, the cluster would not be as dense as the target cluster.

4 Protocol design and Experiments

4.1 Datasets

To evaluate the performance of our image annotation system, we used the Caltech101 (CT) [31] and Pascal VOC2012 (PV) [32] datasets. The CT dataset consists of 101 object categories with 40 to 800 images per category. The PV dataset contains a total of 11,530 images from 20 categories, and multiple object categories can be present in one image.

4.2 Experimental setup

We utilized 2500 images for training, and 2500 images for testing. Both these image sets comprised 200 images of a particular target category that we wanted to annotate. All images were resized 512×512 pixels, and images were displayed at 10 Hz frequency in blocks of 100 in order to minimize viewer distraction and fatigue. During the RSVP task, participants were shown a fixation display for 2 seconds at the beginning of each 100 image sequence. Train and test EEG data were captured using an identical experimental setup with the temporal gap of 5 minutes. Target image categories were decided *a priori* before every experiment.

Our study was conducted with five graduate students (5 male, age 24.4 ± 2.1) with 10/20 corrected vision, seated at a distance of 60 cm from the display. A total of three sessions (each involving train and test set) were performed with each participant. To facilitate engagement, viewers were instructed to count the number of target images during the experiment. Target image classes were different for each session, and included categories like *bike*, *pizza*, *panda*, *sofa*, *etc.* Each participant performed two sessions on the CT dataset and one session on the PV dataset.
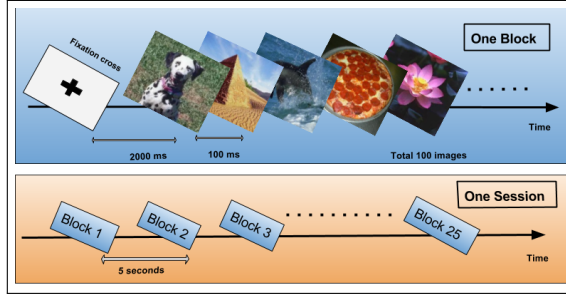
Fig. 4. Experimental protocol: Participants completed two identical sessions (one used for training and the other for test) which were 5 minutes apart. Each session comprised 25 blocks of 100 images, and lasted about six minutes.

5   Results and Discussion

Due to a heavy class imbalance between *target* and *non-target* category images, we use the F1-score to evaluate our annotation results. The F1-score is a popular performance metric used in retrieval studies, and denotes the harmonic mean of the precision and recall scores. All reported results denote the mean F1 achieved with five-fold cross validation.

Table 1. Results synopsis: Annotation performance obtained for the CT and PV datasets across total 15 sessions (5 viewers).

| Dataset	Caltech101	Pascal VOC 2012
Before outliers removal		
F1 score	0.71	0.68
Precision	0.66	0.63
Recall	0.81	0.72
After outliers removal		
F1 score	**0.88**	**0.83**
Precision	0.99	0.97
Recall	0.81	0.72
Target image percentage	8%	8%
Image presentation speed	10 Hz	10 Hz
Number of images in test set	2500	2500

In Table 1, we report the averaged F1 and precision-recall values for the CT and PV datasets across all participants. Note that the precision and F1

scores improve significantly upon outlier removal due to a stark reduction in the number of false positives via feature-based clustering. Overall F1 scores for the PV dataset are lower than for the CT dataset. This can be attributed to the fact that the PV dataset is more complex, as it contains multiple object classes in many images, as compared to CT which contains only one object class per image.

As our annotation system is dependent on viewer ability, its performance is sensitive to human factors. One key factor is the image presentation rate. The image display latency (100 ms) is lower than the P300 response latency ($\approx$ 300 ms) [33]. The rapid image display protocol results in (i) viewers confusing between similar object classes, (ii) viewers unable to fully comprehend visual information from complex images, and (iii) EEG data for consecutive images having significant overlap leading to misclassification.

Therefore, we hypothesized that reducing the image display rate would (a) allow the viewer to better comprehend the visual content (especially for complex images), (b) better delineation of EEG responses, and (c) better manifestation of ERP signatures. These in turn, would improve our annotation performance while marginally reducing the annotation throughput. Fig.5 presents the observed results. Note that a 3% increase in F1-score is observed when the image presentation rate is reduced from 10 to 4 images/second, validating our hypothesis.
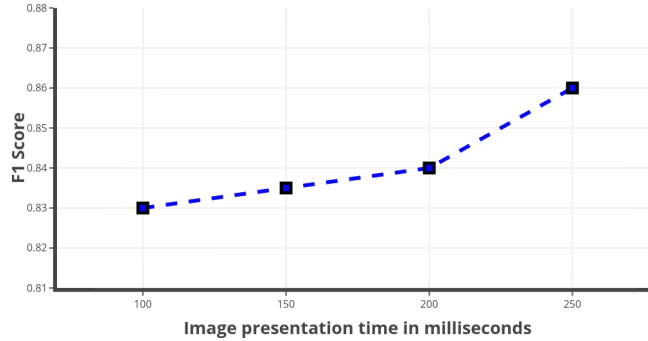
Fig. 5. Presentation rate vs annotation performance: Variation in F1-score with image display rate.

Conversely, since our annotation system is solely based on P300 signatures which are task specific but target class agnostic. Therefore, it is not mandatory to train the EEGNet with object class-specific EEG responses. To validate this aspect, we trained and tested the EEGNet with EEG responses corresponding to different object categories. Table 2 presents the F1 scores achieved for the five viewers with class-agnostic train and test EEG data. Note that only a marginal difference in annotation performance is noticeable with class-specific and class-agnostic EEG data across viewers. Since we are using the pre-trained VGG-19

model exclusively to extract feature descriptors, it can be used without further fine tuning for any new target class categories.

Table 2. Annotation performance with class-specific vs class-agnostic EEG data for five viewers.

F1 Score	P1	P2	P3	P4	P5
Class-specific train and test	0.88	0.86	0.89	0.87	0.88
Class-agnostic train and test	0.85	0.85	0.84	0.86	0.86

6   Conclusion

In order to facilitate large-scale image annotation efforts for computer vision and scene understanding applications, we propose an EEG-based fast image annotation system. Our annotation system exclusively relies on the P300 ERP signature, which is elicited upon the viewer detecting a pre-specified object class in the displayed image. A further outlier removal procedure based on binary feature-based clustering significantly improves annotation performance.

Overall, our system achieves a peak F1-score of 0.88 with a 10 Hz annotation throughput. Another advantage of our method is that the P300 signature is specific to the target detection task, but not the underlying object class.Therefore, any novel image category can be annotated with existing models upon compiling the viewer EEG responses. Future work will focus on discovering and exploiting object-specific EEG signatures, and combining multiple human responses (*e.g.*, EEG plus eye movements) for fine-grained object annotation and classification.

References

1. S. Zhang, J. Huang, Y. Huang, Y. Yu, H. Li, and D. N. M., "Automatic image annotation using group sparsity," in *CVPR, 2010*.
2. Y. Verma and C. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *ECCV*, 2012.
3. V. Yashaswi and C. Jawahar, "Exploring svm for image annotation in presence of confusing labels," in *BMVC,2013*.
4. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.
5. H. Fu, Q. Zhang, and G. Qiu, "Random forest for image annotation," in *-ECCV 2012*.
6. M. Wang and X.-S. Hua, "Active learning in multimedia annotation and retrieval: A survey," *TIST,2011*.
7. G. Sychay, E. Chang, and K. Goh, "Effective image annotation via active learning," in *Proceedings. IEEE International Conference on Multimedia and Expo*, vol. 1, 2002, pp. 209–212 vol.1.

8. P. Bakliwal and C. Jawahar, "Active learning based image annotation," in *NCVPRIPG*. IEEE, 2015.

9. H. Katti, R. Subramanian, M. Kankanhalli, N. Sebe, T.-S. Chua, and K. R. Ramakrishnan, "Making computers look the way we look: Exploiting visual attention for image understanding," in *ACM International Conference on Multimedia*, 2010, pp. 667–670.

10. R. Subramanian, D. Shankar, N. Sebe, and D. Melcher, "Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes." *Journal of Vision*, vol. 14, no. 3, pp. 1–18, 2014.

11. A. Oliva, "Gist of the scene," *Neurobiology of attention*, vol. 696, pp. 251–256, 2005.

12. C. Keysers, D. Xiao, P. Foldiak, and D. Perrett, "The speed of sight," *J. Cognitive Neurosci.*, pp. 90–101, 2001.

13. D. E. Linden, "The P300: where in the brain is it produced and what does it tell us?" *The Neuroscientist*, vol. 11, no. 6, pp. 563–576, 2005.

14. E. Mohedano, G. Healy, K. McGuinness, X. Giró-i Nieto, N. E. OConnor, and A. F. Smeaton, "Improving object segmentation by using eeg signals and rapid serial visual presentation," *Multimedia tools and applications*, vol. 74, no. 22, pp. 10 137–10 159, 2015.

15. E. A. Pohlmeyer, J. Wang, D. C. Jangraw, B. Lou, S.-F. Chang, and P. Sajda, "Closing the loop in cortically-coupled computer vision: a brain–computer interface for searching image databases," *Journal of neural engineering*, vol. 8, no. 3, p. 036025, 2011.

16. S. Koelstra, C. Mühl, and I. Patras, "Eeg analysis for implicit tagging of video data," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE, 2009, pp. 1–6.

17. R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe, "ASCERTAIN: Emotion and personality recognition using commercial sensors," *IEEE Transactions on Affective Computing*, 2016.

18. A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian, "Affect recognition in ads with application to computational advertising," in *ACM International Conference on Multimedia*, 2017.

19. A. Kapoor, P. Shenoy, and D. Tan, "Combining brain computer interfaces with vision for object categorization," in *CVPR, 2008*, 2008.

20. C. Spampinato, S. Palazzo, I. Kavasidis, D. Giordano, M. Shah, and N. Souly, "Deep learning human mind for automated visual classification," 2017.

21. M. Bilalpur, S. M. Kia, M. Chawla, T. Chua, and R. Subramanian, "Gender and emotion recognition with implicit user signals," in *International Conference on Multimodal Interaction*, 2017.

22. T. W. Picton *et al.*, "The p300 wave of the human event-related potential," *Journal of clinical neurophysiology*, vol. 9, pp. 456–456, 1992.

23. V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "Eegnet: A compact convolutional network for eeg-based brain-computer interfaces," *arXiv preprint arXiv:1611.08024*, 2016.

24. D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.

25. D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

26. A. Paszke, S. Chintala, R. Collobert, K. Kavukcuoglu, C. Farabet, S. Bengio, I. Melvin, J. Weston, and J. Mariethoz, "Pytorch: Tensors and dynamic neural

networks in python with strong gpu acceleration, may 2017," https://github.com/pytorch/pytorch.

27. R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for eeg decoding and visualization," *Human Brain Mapping*, aug 2017. [Online]. Available: http://dx.doi.org/10.1002/hbm.23730

28. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Advances in neural information processing systems*, 2014, pp. 3320–3328.

29. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

30. L. V. D. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

31. L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59–70, 2007.

32. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results."

33. J. Polich, "Updating P300: An integrative theory of p3a and p3b," *Clinical neurophysiology*, vol. 118, no. 10, pp. 2128–2148, 2007.