HALLUCINATING BRAINS WITH ARTIFICIAL BRAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human brain function as measured by functional magnetic resonance imaging (fMRI), exhibits a rich diversity. In response, understanding the individual variability of brain function and its association with behavior has become one of the major concerns in modern cognitive neuroscience. Our work is motivated by the view that generative models provide a useful tool for understanding this variability. To this end, this manuscript presents two novel generative models trained on real neuroimaging data which synthesize task-dependent functional brain images. Brain images are high dimensional tensors which exhibit structured spatial correlations. Thus, both models are 3D conditional Generative Adversarial networks (GANs) which apply Convolutional Neural Networks (CNNs) to learn an abstraction of brain image representations. Our results show that the generated brain images are diverse, yet task dependent. In addition to qualitative evaluation, we utilize the generated synthetic brain volumes as additional training data to improve downstream fMRI classifiers (also known as decoding, or brain reading). Our approach achieves significant improvements for a variety of datasets, classification tasks and evaluation scores. Our classification results provide a quantitative evaluation of the quality of the generated images, and also serve as an additional contribution of this manuscript.

1 INTRODUCTION

Functional Magnetic Resonance Imaging (fMRI) is a common tool used by cognitive neuroscientists to investigate the properties of brain function in response to stimuli. Classic analysis approaches (Poldrack et al., 2011) focused on analyzing group-averaged brain function images. However, it was discovered that brain activation patterns vary significantly between individuals. Thus, modern analysis now prioritizes understanding the inter-subject variability of brain function (Dubois & Adolphs, 2016; Geerligs et al., 2017). Our work is motivated by the view that generative models provide a useful tool for understanding this variability – as they enable the synthesis of a variety of plausible brain images representing different hypothesized individuals, and high-quality generative models can be analyzed to posit potential mechanisms that explain this variability (Horn et al., 2008). The results presented in this paper provide – to our knowledge for the first time, positive results suggesting that it is indeed possible to generate high quality diverse and task dependent brain images.

While we can qualitatively evaluate generative brain images, quantitative evaluation allows us to objectively compare between various results. To this end, we utilize the generated synthetic brain volumes as additional training data to improve downstream fMRI classifiers. The use of classifiers to predict behavior associated with brain images is also known as decoding or brain reading (Varoquaux & Thirion, 2014; Pereira et al., 2009). Classifiers such as support vector machines and deep networks have been applied for decoding brain images. For example, Cox & Savoy (2003) attempted to classify which of 10 categories of object a subject was looking at (including similar categories, such as horses and cows) based on limited number of brain images. Besides visual tasks, Mitchell et al. (2008) distinguished active region of brains when subjects listened to linguistic words, where the stimuli included five items from each of 12 semantic categories (animals, body parts etc.).

Beyond providing a model for individual variability, high quality brain image synthesis addresses pressing data issues in cognitive neuroscience. Progress in the neurosciences is stifled by the difficulty of obtaining brain data either because of limited culture of sharing, or due to medical privacy regulations (Poldrack & Gorgolewski, 2014). For the computational neuroscientist, generated images deliver *unlimited* quantities of high quality brain imaging data. Generative models provide

data that can be used to develop state of the art tools before application to real data (Varoquaux & Thirion, 2014). The simulated data approach has already proven invaluable to many scientific fields including particle physics and astronomy (Castelvecchi et al., 2017).

One of the promising generative models are Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). They are able to capture complex distributions using a non-cooperative two-player game formulation: a generator produces synthetic data by transforming samples drawn from a simple distribution; and a discriminator focuses on distinguishing synthetic and real data. Despite (or possibly due to) the compelling formulation, GAN training is known to be unstable. To address this difficulty various variants have been proposed. Wasserstein GANs (WGANs) (Arjovsky et al., 2017) formulate the objective using the Wasserstein distance rather than the classical Jenson-Shannon divergence. Improved training of WGANs (Gulrajani et al., 2017) applies an additional gradient penalty, which avoids the critical weight clipping in WGANs which might lead to pathological behavior. Dualing GANs restrict the discriminator and formulate the dual objective (Li et al., 2017).

Beyond improving the stability, conditional GANs (Mirza & Osindero, 2014) make it possible to control the data generation process by conditioning the model on additional information. Auxiliary Classifier GANs (AC-GANs) (Odena et al., 2016) unify a GAN and classifier to a single architecture, employing labels to generate ImageNet samples. 3D GANs (Wu et al., 2016) reconstruct 3D objects and Edward & David (2017) propose to use improved WGAN to enhance the stability of 3D GANs.

We make the following contributions in this paper:

1. We develop Improved Conditional Wasserstein GANs (ICW-GAN) and Auxiliary Classifier and Discriminator GANs (ACD-GAN), two types of 3D conditional GANs to synthesize fMRI brain data, both of which we found to be stable to train and to generate high quality and diverse 3D brain images.
2. We evaluate of our models on three different datasets using a series of image classification tasks with support vector machines and deep net classifiers at two levels of resolution. Results show that synthetic data generated by our models can greatly improve test set classification accuracy of brain volumes.
3. Use metrics and decoding tools, we assess the quality of generated brain volumes and their diversity.

2 Related Work

2.1 Neural Nets for fMRI Brain Data

Neural networks have been used by several authors for classifying brain imaging data. Firat et al. (2014) and Koyamada et al. (2015), used 2D deep nets to extract features of fMRI brain images to classify brain states. D. Nathawani et al. (2016) applied both 2D and 3D neural networks to classify fMRI brain data. Svanera et al. (2017) decoded fMRI data of video stimuli and classify data into visual categories. Similarly, D. Nathawani et al. (2016) extracted features from 4-D fMRI data and used deep learning methods for discrimination of cognitive processes. For manuscripts that apply both neural networks and other classifiers such as support vector machines, the classification results with CNN are generally much better than using SVM in the single subject prediction task, which inspires us to continue to use neural networks in multiple subject classification and to explore ways to improve it.

2.2 Generative Adversarial Nets

To learn a distribution over data $\boldsymbol{x}$, a GAN (Goodfellow et al., 2014) formulates a 2-player non-cooperative game between two deep nets. The *generator* G uses a random noise vector $\boldsymbol{z}$ sampled from a prior distribution $P_{\boldsymbol{z}}(\boldsymbol{z})$ as input and produces an image $G(z)$. The generator is trained to fool the discriminator, which receives either synthetic data or real data and differentiates between them. Formally, G and D play the following two-player minimax game with value function $V(G, D)$:

$$\min_G \max_D V(G, D) := \mathbb{E}_{\boldsymbol{x} \sim P_{data}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z} \sim P_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \tag{1}$$

For the conditional GAN (Mirza & Osindero, 2014), both G and D are conditioned on some extra information $\boldsymbol{y}$, for instance, class labels or other features. The conditioning can be presented by feeding $\boldsymbol{y}$ into both the generator and the discriminator. This work is successful in one-to-many mappings such as image labeling with many tags.

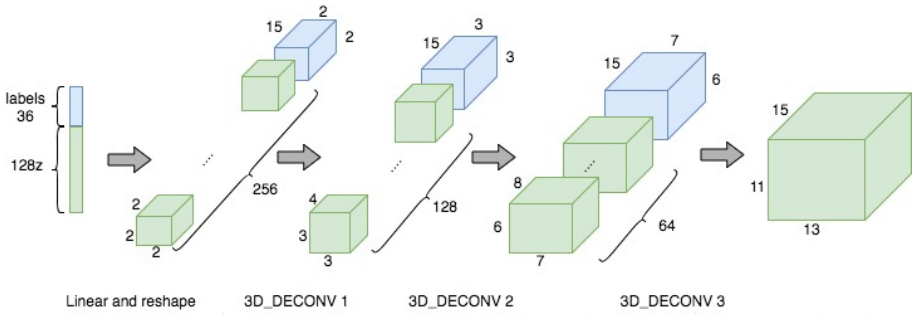


Figure 1: The generator of our ICW-GAN. The encoding z is drawn from a multi-variate normal distribution. The label vector is a one-hot encoding, i.e., the one entry that equals one represents the class that the generated volume belongs to. Our stride in the de-convolutional layers is [1,2,2,2,1] in the batch, height, width, length, feature map dimension.

The Wasserstein GAN (WGAN) (Arjovsky et al., 2017) uses the *Earth-Mover* distance, also known as *Wasserstein-1* distance W(q,p) as the objective function. It computes the minimum cost of transporting mass to transform the distribution q into the distribution p. Under this circumstance, W(q,p) is continuous everywhere and its gradient with respect to its input was found to be more stable than its classical GAN counterpart. Improved Wasserstein GAN (IWGAN) (Gulrajani et al., 2017) argues that weight clipping of the critic, i.e., the discriminator, in WGANs inevitably causes the gradient to either vanish or to explode. To address this issue Gulrajani et al. (2017) propose an alternative penalty term in the critic loss based on the gradient norm.

In an auxiliary classifier GAN (AC-GAN) (Odena et al., 2016), every generated sample has a corresponding label which can be used as a condition in the *generator* and the *classifier*. The *discriminator* is modified to contain an auxiliary decoder network to reconstruct the training data class. Formally, the objective function differs from other versions of GANs, which also contain the log-likelihood of the classifier.

Further, 3D-GANs (Wu et al., 2016) extend GANs to 3D object generation. Different from classical GANs, 3D-GANs apply the three dimensional convolution in both the generator and the discriminator. By learning deep object representations, 3D GANs can generate visually appealing yet variable 3D object volumes.

3 Approach

In the following we introduce two models for fMRI data generation: ICW-GAN and ACD-GAN. They differ in model structure and in the way label information is taken into account. We first show the model structure of 3D ICW-GAN and its downstream classifier in Section 3.1, and present ACD-GAN in Section 3.2.

3.1 3D Improved Conditional Wasserstein GAN (ICW-GAN)

Similar to classical generative adversarial nets (GANs), ICW-GANs are formulated as a non-cooperative two-player game between two adversaries: (1) a generator $\hat{x} = G_\theta(z)$, which generates artificial samples $\hat{x}$ from randomly drawn latent encodings z via a transformation using a deep net parameterized by θ; and (2) a discriminator $D_w(x) = \sigma(f_w(x))$ which computes the probability that a given data point x is real rather than artificial, by using a sigmoid function σ of the logit f_w obtained from a deep net parameterized by w. Following Arjovsky et al. (2017), we build our models based on minimizing the Wasserstein distance via

$$\min_\theta \max_{w \in W} \mathbb{E}_{\boldsymbol{x} \sim P_{data}(\boldsymbol{x})}[f_w(\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{z} \sim P_{\boldsymbol{z}}(\boldsymbol{z})}[f_w((G_\theta(\boldsymbol{z})))], \tag{2}$$

where $\{f_w\}_{w \in W}$ denotes a set of functions that are K-Lipschitz for some K. IW-GANs (Gulrajani et al., 2017) provide a more stable method for enforcing the Lipschitz constraint by employing a gradient penalty rather than weight clipping, when optimizing:

$$\mathbb{E}_{\boldsymbol{z} \sim P_{\boldsymbol{z}}(\boldsymbol{z})}[(1 - D_w(G_\theta(\boldsymbol{z})))] - \mathbb{E}_{\boldsymbol{x} \sim P_{data}(\boldsymbol{x})}[D_w(\boldsymbol{x})] + \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{\boldsymbol{x}}}}[(\| \nabla_{\hat{\boldsymbol{x}}} D_w(\hat{\boldsymbol{x}}) \|_2 - 1)^2]. \tag{3}$$

Hereby $\hat{\boldsymbol{x}}$ is a convex combination of real data and artifical samples, i.e., $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon) G(\boldsymbol{z})$ with $\epsilon \sim U[0, 1]$.

In our models, we extend IW-GAN in two directions. First, because fMRI data is three dimensional, 3D convolution and deconvolution are used instead of 2D ones in order to maintain the spatial

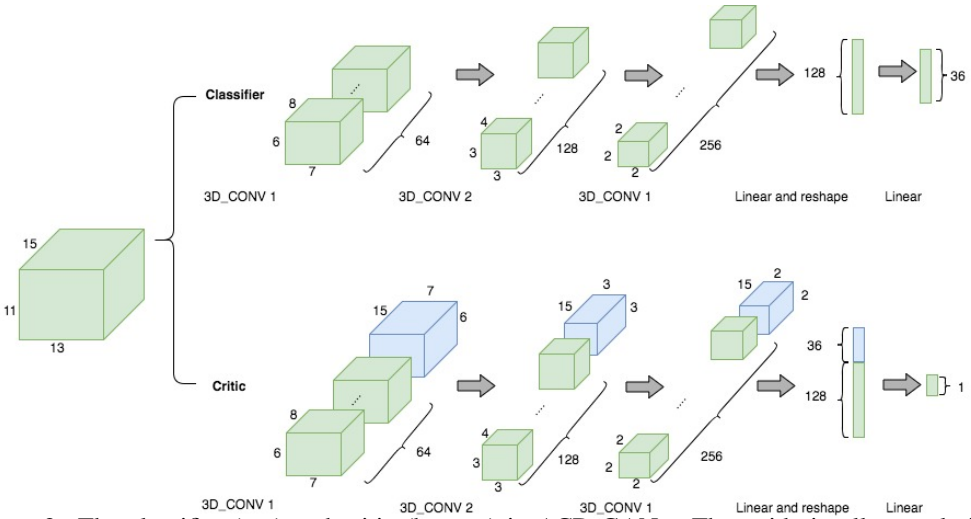


Figure 2: The classifier (top) and critic (bottom) in ACD-GANs. The stride in all convolutional layers is [1,2,2,2,1] in the batch, height, width, length, feature map dimension.

structure and voxel information of the fMRI data. Second, we condition both the discriminator D_w and the generator G_θ on available labeled data.

As shown in Figure 1, our generator consists of three fully convolutional layers. In convolutional layers, we use kernels with a size $4 \times 4 \times 4$ and strides 2, with batch normalization and Leaky ReLU layers added between and a Tanh layer at the end. The D looks like a mirror of the G but no activity function is used at the end of the D.

This ICW-GAN model follows models proposed in (Gulrajani et al., 2017; Mirza & Osindero, 2014; Wu et al., 2016), however, we found the following modifications to stabilize training and supervise the generation process for classification data: compared to existing GANs, our CIW-GAN applies a more stable upper bound for Wasserstein distance as an objective function, leverages conditioning labels to supervise the generation process, and utilizes 3D convolution methods for 3D object generation.

To include label information, we concatenate labels to the input and to the hidden layers. At the input of the generator, one-hot labels are combined with the brain vector. Then, for each of the intermediate layers, we use a fully connected layer followed by a $\tanh$ activation to transform the one-hot vector to a volume of appropriate size, i.e., $15 \times 2 \times 2$ for the first hidden layer, and $15 \times 3 \times 3$ for the next. We empirically found dimensions other than 15, e.g., 3 to work well too. We concatenate the label volume to intermediate volumes on feature dimension and pass the joint one to the next deconvolution layer. We follow the same procedure in the architecture of the discriminator.

Referring to the volume labels via $\boldsymbol{y}$, the objective function L of our ICW-GAN model is as follows:

$$L = \mathbb{E}_{\boldsymbol{z} \sim P_{\boldsymbol{z}}(\boldsymbol{z})}[(1 - D(G(\boldsymbol{z}|\boldsymbol{y})))] - \mathbb{E}_{\boldsymbol{x} \sim P_{data}(\boldsymbol{x})}[D(\boldsymbol{x}|\boldsymbol{y})] + \lambda \mathbb{E}_{\hat{x} \sim P_{\hat{\boldsymbol{x}}}}[(\parallel \nabla_{\hat{\boldsymbol{x}}} D(\hat{\boldsymbol{x}}|\boldsymbol{y}) \parallel_2 - 1)^2], \quad (4)$$

where $\hat{\boldsymbol{x}} \leftarrow \epsilon \boldsymbol{x} + (1 - \epsilon) G(\boldsymbol{z})$, $\epsilon \sim U[0, 1]$. λ is a gradient penalty coefficient. We optimize both the discriminator and generator loss using Adam optimizer(Kingma & Ba (2014)).

Downstream Classifiers for ICW-GAN

Our work is motivated by the difficulty to acquiring fMRI data, and our goal is to enlarge the dataset for downstream tasks such as classification, using synthetically generated samples. To assess the suitability of the developed generation procedure, we employ a variety of downstream classification processes to illustrate the benefits of training with synthetic brain images. More specifically, we train a classifier with real ('Real') or real plus generated ('Real+G') data. Common classifiers in neuroscience are Support Vector Machines (SVMs) and a 3D deep nets. We think usage of both provides valuable insights as their number of trainable parameters differs significantly, i.e., 14 parameters for SVMs and several hundred thousand for our deep nets. To measure the results obtained from different training datasets we use accuracy, macro F1, precision and recall metrics.

SVM: We use a simple linear SVM to classify test data and don't extract any intermediate features. Instead, we use raw brain data, vectorized to a 1-dimensional vector for classification to avoid any form of collusion.

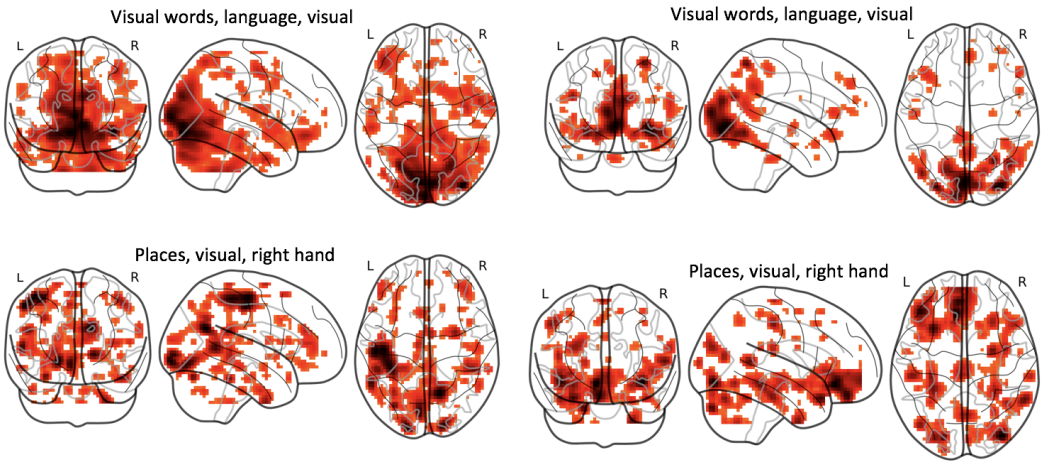


Figure 3: 2D projections of synthetic brain volumes generated using the proposed CIW-GAN. The training dataset is Collection 1952. The tag above each image sequence shows the class the picture belongs to.

Scaling	Input	Classifier	Accuracy	Macro F1	Precision	Recall
4.0	Real	SVM	0.797	0.797	0.813	0.797
	Real	NN	0.802	0.802	0.817	0.802
	Real+G	SVM	0.806	0.803	0.823	0.807
	Real+G	NN	**0.819**	**0.817**	**0.830**	**0.819**
2.0	Real	SVM	0.855	0.857	0.867	0.857
	Real	NN	0.863	0.863	0.872	0.863
	Real+G	SVM	0.860	0.863	0.860	0.857
	Real+G	NN	**0.891**	**0.894**	**0.906**	**0.891**

Table 1: Results on collection 1952. Scaling means times for brain volumes to be scaled down. Input represents the input of a classifier: Real means only use real training data, while Real + G means real training data plus generated data which is produced by CIW-GAN. Two classifiers are utilized: SVM and neural nets (NN).

Deep Net: The deep net structure is similar to the discriminator, having a 3 dimensional structure and identical number of convolution layers with a Leaky ReLU activation. Obviously the classifier doesn't concatenated intermediate and input data with any label information. Moreover, instead of only discriminating between real or synthetic brain volumes, generally, the classifier has a higher dimensional output depending on the number of classes of the dataset. We'll provide the respective information for each dataset in the experimental section.

3.2 3D Auxiliary Classifier and Discriminator GAN (ACD-GAN)

Instead of assessing the classifier performance in a separate downstream classification task, we also develop a conditional architecture of GANs which jointly trains the generator, the discriminator and the classifier. This form of training is inspired by the AC-GAN (Odena et al., 2016) formulation. Our generator architecture is illustrated in Figure 1 and identical to the one employed for the ICW-GAN. The combination of classifier and discriminator is provided in Figure 2. As before, the discriminator assesses the probability that a provided sample is real or artificially generated. In addition, the classifier provides a probability distribution over the class labels. In this model, training passes labels through the generator and the discriminator in order to condition the adversarial process. However, no labels are employed in the classification stream. The classifier loss, L_C, is the log-likelihood of the correct class. Compared to L in Eq. (4), the discriminator loss, L_D includes the additional classifier component

$$L_C = \mathbb{E}_{\boldsymbol{x}\sim P_{data}(\boldsymbol{x})}[\log P(C = c|\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}(\boldsymbol{z})}[\log P(C = c|G(\boldsymbol{z})]. \tag{5}$$

The generator loss function, L_G, consists of the likelihood of fake images and one part of L_C, i.e.,

$$L_G = \mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}(\boldsymbol{z})}[P(D = \text{fake}\,|G(\boldsymbol{z})) + \mathbb{E}_{\boldsymbol{z}\sim P_{\boldsymbol{z}}(\boldsymbol{z})}[P(C = c|G(\boldsymbol{z})]. \tag{6}$$

Similarly, the discriminator loss function is given by

$$L_D = L + L_C \tag{7}$$

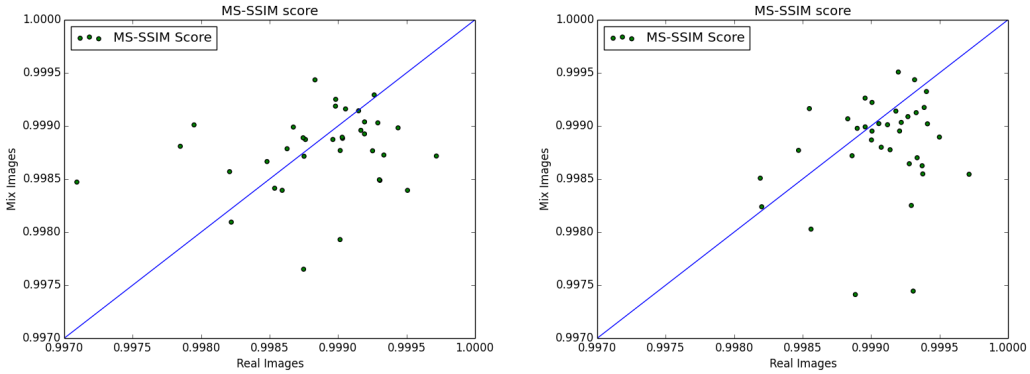


Figure 4: Mean MS-SSIM scores between pairs of images within a given class for brain volumes of high resolution (left) and low resolution (right). Each point represents an individual class. Values in horizontal axis are MS-SSIM scores computed on real brain images, while values on the vertical axis are calculated using mixed images.

Scaling	Input	Classifier	Accuracy	Macro F1	Precision	Recall
8.0	Real	SVM	0.523	0.480	0.497	0.523
	Real	NN	0.530	0.517	0.545	0.530
	Real+G	SVM	0.531	0.493	0.510	0.533
	Real+G	NN	**0.562**	**0.539**	**0.568**	**0.563**
4.0	Real	SVM	0.555	0.507	0.517	0.533
	Real	NN	0.723	0.712	**0.737**	**0.723**
	Real+G	SVM	0.562	0.517	0.527	0.563
	Real+G	NN	**0.737**	**0.715**	0.727	**0.723**

Table 2: Results on Collection 2138.

We optimize the generator, classifier and discriminator loss functions using Adam optimizer(Kingma & Ba (2014)).

4 EXPERIMENTS

Our work is motivated by the fact that fMRI data is rarely available. To assess the suitability of generative training methods for data generation, in this section, we examine the performance of the two aforementioned models on three different datasets: collection 1952, 2138 and 503[1]. To illustrate the benefits of generated fMRI data for classifier training, we focus on two variants of the model architecture for generating images at $13 \times 15 \times 11$ and $26 \times 31 \times 23$ spatial resolutions for collection 1952, and $13 \times 15 \times 11$ and $26 \times 31 \times 22$ resolutions for the other two collections.

In Section 3, we present the architectures of our models for generating low resolution data, which only have 3 convolutional layers. For high resolution data we add one more convolutional and deconvolutional layer in the discriminator and generator respectively. At the same time, for high resolution data the classifier also has 4 convolutional layers in the ACD-GAN. We first show qualitative results of generated 3D volumes. Then we provide quantitative results for 3D volume classification.

To carefully assess the classifier performance we use cross validation in all reported results. See the supplementary material for details.

Dataset 1: Collection 1952: Collection 1952 has 6573 brain images, 45 classes with a total number of 19 sub-classes. The labels include task description that subjects were going through during the imaging process, such as visual, language and calculate. A combination of 2 to 4 sub-classes are subsumed for a class. There are a total of 45 classes because not every potential combination is reasonable. In our experiments, we on a subset of 36 classes because 9 of the 45 classes have less than 30 examples. If a class has more than 100 images, we split images in that class 7:1:2 into training, validation and test datasets. Below 100 images but more than 30 images we use the proportion of 3:1:2 for splitting images. We ensure that the test data always consists of real images only. We ensure that the test data always consists of real images only. We train the ICW-GAN and ACD-GAN for 2500 epochs with a learning rate of 1e-4 and 50% exponential decay for each 3000 iterations and a batch size of 50. 2-D projections of several brain volumes generated by ICW-GAN

[1] publicly available from https://neurovault.org/collections/503

Scaling	Input	Classifier	Accuracy	Macro F1	Precision	Recall
6.0	Real	SVM	0.212	0.210	0.220	0.210
	Real	NN	0.293	0.287	0.302	0.293
	Real+G	SVM	0.225	0.227	0.24	0.223
	Real+G	NN	**0.299**	**0.298**	**0.321**	**0.298**
3.0	Real	SVM	0.233	0.230	0.233	0.233
	Real	NN	0.358	**0.386**	0.390	0.358
	Real+G	SVM	0.231	0.227	0.230	0.230
	Real+G	NN	**0.373**	0.380	**0.410**	**0.372**

Table 3: Results on collection 503.

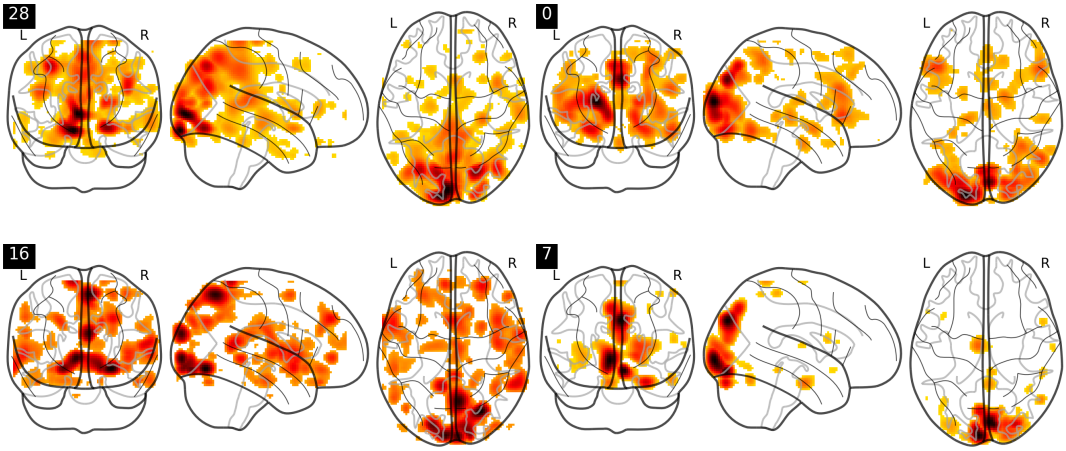


Figure 5: 2-D projections of synthetic brain volumes by ACD-GAN using training data in collection 503. Still, a left top tag is the class generated brains belong to. From the top to the bottom, and left to right, the class for each projections is 28, 0, 16 and 7. Each class represents a picture being shown to the subjects. Linear normalization is used and the lower threshold is 0.3.

are illustrated in Figure 3. The projections in the top row of Figure 3 are brains of class 'visual words, language, visual', and those at the bottom are from class 'places, visual, right hand'. By examining the generated brain images, we find diversity in both intra-class and between-class. We refer the interested reader to the supplementary material for a comparison of generated brains and real brain recordings.

To assess the suitability of the generated data we conduct a classification task. Note that the test data is always composed of real images. The classification results are shown in Table 1. Scaling 4.0 means that the training images are rescaled by a factor of 4.0 times to $13 \times 15 \times 11$ and by a factor of 2.0 to a shape of $26 \times 31 \times 23$. The second column indicates the type of training data we use for training a classifier: only using real data ('Real'), or using the mixed data of real and generated volumes ('Real+G'). The third column denotes the classifier type, i.e., an SVM or a deep net ('NN'). We use the validation dataset to choose the best training models and use these models to classify the test data. We observe the deep net classifier to generally outperform SVMs and we observe generated data to be beneficial for classifier training.

Inspired by Odena et al. (2016), we also compute the multi-scale structural similarity (MS-SSIM) (Wang et al., 2003) to examine the intra-class diversity of the generated data. MS-SSIM score is a well-designed similarity metric which utilizes image details at different resolutions. Its values are within $[0, 1]$. Higher MS-SSIM values correspond to more similar images. We measure the mean MS-SSIM score with 100 randomly chosen volumes of mixed data with a given class, and ones of real training data. In $13 \times 15 \times 11$ resolution, we find that 22 classes with mixed data have a lower MS-SSIM score than only with real training data. In other words, 61.1% classes with mixed data have sample variability that exceeds those only with real training data. In $26 \times 31 \times 23$ resolution, 69.4% classes with mixed data have a lower MS-SSIM score. See Figure 4 for details.

Dataset 2: Collection 2138: For data in collection 2138, there are 1847 brain images, 61 classes and 50 labels. Because of the small size of the dataset, we randomly choose 70% of the brain images as training data and leave 30% as test data. In this case, we do not have development data to supervise the training process; thus, we train our models for 1000 epochs in several runs and record the best classification results, which are summarized in Table 2. In this collection, we downsample brains by

Dataset	Model	Accuracy	Macro F1	Precision	Recall
1952	ICW-GAN	0.819	**0.817**	**0.830**	**0.819**
	ACD-GAN	**0.835**	0.815	0.828	0.817
2138	ICW-GAN	0.562	0.539	0.568	**0.563**
	ACD-GAN	**0.569**	**0.547**	**0.579**	0.560
503	ICW-GAN	0.299	0.298	0.321	0.298
	ACD-GAN	**0.326**	**0.299**	**0.329**	**0.301**

Table 4: Results of ACD-GAN

a factor of 8 and 4 to volume sizes of $13 \times 15 \times 11$ and $26 \times 31 \times 22$ respectively. Similar to the earlier reported results we observe deep nets to outperform SVMs, but more importantly, generated data during training was again suitable to improve the classifier performance on real data.

Dataset 3: Collection 503: 5067 brain images are in collection 503. Subjects in this dataset are required to respond to 30 images from the International Affective Picture Set (Chang et al., 2015). These 30 images were used to train the Picture Induced Negative Emotion Signature also described in this paper. We consider the 30 pictures classes of brain data. In other words, there are overall 30 classes and each class represents a picture shown to subjects. This collection needs extra preprocessing because brain volumes have two shapes: $79 \times 95 \times 68$ and $91 \times 109 \times 91$. We conform large volumes to the small ones. Similarly, we do the experiment in the two levels of resolutions. Classification results are summarized in Table 3 and again follow the trend reported earlier.

Results of Using ACD-GAN: We apply the ACD-GAN to brain data of low resolution which is efficient and time-saving way to examine the model. Since our ICW-GAN gains the best performance in the three databases if we use mixed training data, we compare these best results to results of ACD-GAN. 2-D projections of brain volumes generated by ACD-GAN are shown in Figure 5. In Table 4, we present classification results of our models and show that ACD-GAN obtains improved accuracy scores compared to a ICW-GAN on most of the investigated metrics.

Comparison and Analysis: To the best of our knowledge, we are the first to apply generative models to artificially enlarge brain datasets and improve downstream classification tasks. As a result, we consider SVM classification results as a baseline, no matter which sources of training data is used. Though the distribution of each evaluation score for the three datasets is different, they hold two similar patterns. First, according to all the results presented above, our ICW-GAN has overall better performance in these three datasets. Further, the ICW-GAN hold the same status at both levels of resolution for brain images. We only test the ACD-GAN in small resolution, results show that it can better improve the accuracy scores compared to ICW-GAN.

We also want to point out that our strategy, to mix real and generated data, although competitive, does not always yield the best results. For example, using real data only to train the NN classifier achieves the largest macro F1 score at high resolution in Table 3. On the one hand, accuracy metrics are the most popular and outstanding score to estimate multi-class classification tasks. On another hand, to explain why our models cannot always achieve the highest score in other scores, we think there exists a reasonable range of variation for those scores because of a few unavoidable factors such as parameter initialization.

5 Conclusions

Generative models provide a useful tool for understanding the individual variability of brain images. The results of this manuscript show – to our knowledge for the first time, that 3-D conditional GANs, in particular our proposed CIW-GAN and ACD-GAN, can generate high quality diverse and task dependent brain images. We hope our results inspire additional research on generative models for brain imaging data. Beyond qualitative evaluation, we evaluate quantitative performance by using the generated images as additional training data in a predictive model – mixing synthetic and real data to train classifiers. The results show that our synthetic data augmentation can significantly improve classification accuracy under various circumstances – a result which may be of independent interest. Future work will be focused additional qualitative evaluation of the generated images by neuroscience experts and exploration of various applications. We also plan to more throughly investigate the trained models to understand what it my teach neuroscientists about individual variability. Finally, we plan to expand to model to combine data across multiple studies by exploring techniques for merging labels across studies.

REFERENCES

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GANs. *arXiv preprint arXiv:1701.07875*, 2017.

Davide Castelvecchi et al. Astronomers explore uses for AI-generated images. *Nature*, 2017.

L.-J. Chang, P.-J. Gianaros, S.-B.Manuck, A. Krishnan, and T.-D. Wager. A sensitive and specific neural signature for picture-induced negative affect. *PLoS biology*, 2015.

D.-D Cox and R.-L Savoy. Functional magnetic resonance imaging (fMRI)brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 2003.

Dhruv D. Nathawani, T. Sharma, and Y. Yang. Neuroscience meets deep learning, 2016.

J. Dubois and R. Adolphs. Building a science of individual differences from fMRI. *Trends in cognitive sciences*, 2016.

S. Edward and M. David. Improved Adversarial Systems for 3D Object Generation and Reconstruction. *CoRR*, 2017.

O. Firat, L. Oztekin, and F. Vural. Deep learning for brain decoding. In *Image Processing (ICIP), 2014 IEEE International Conference on*, 2014.

L. Geerligs, K. A. Tsvetanov, and R. N. Henson. Challenges in measuring individual differences in functional connectivity using fMRI: The case of healthy aging. *Human Brain Mapping*, 2017.

Ian J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Nets. In *Advances in neural information processing systems*, 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville. Improved training of Wasserstein GANs. *arXiv preprint arXiv:1704.00028*, 2017.

J.-D. Van Horn, S. T. Grafton, and M. B. Miller. Individual variability in brain activity: a nuisance or an opportunity? *Brain imaging and behavior*, 2008.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

S. Koyamada, Y. Shikauchi, K. Nakae, M. Koyama, and S. Ishii. Deep learning of fmri big data: a novel approach to subject-transfer decoding. *arXiv preprint arXiv:1502.00093*, 2015.

Y. Li, A. G. Schwing, K.-C. Wang, and R. Zemel. Dualing GANs. In *Proc. NIPS*, 2017.

M. Mirza and S. Osindero. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*, 2014.

T. M. Mitchell, S. V. Shinkareva, A. Carlson, K. Chang, V.-L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. *science*, 2008.

A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier GANs. *arXiv preprint arXiv:1610.09585*, 2016.

F. Pereira, T. Mitchell, and M. Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 2009.

R. A. Poldrack, J. A. Mumford, and T. E. Nichols. *Handbook of functional MRI data analysis*. 2011.

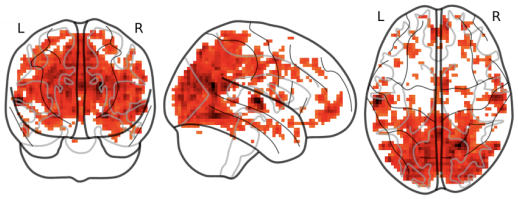
Russell A Poldrack and Krzysztof J Gorgolewski. Making big data open: data sharing in neuroimaging. 2014.

M. Svanera, S. Benini, G. Raz, T. Hendler, R. Goebel, and G. Valente. Deep driven fMRI decoding of visual categories. *arXiv preprint arXiv:1701.02133*, 2017.

G. Varoquaux and B. Thirion. How machine learning is shaping cognitive neuroimaging. *GigaScience*, 2014.

Z. Wang, E.-P. Simoncelli, and A.-C. Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2004. Conference Record of the Thirty-Seventh Asilomar Conference on*, 2003.

J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. 2016.



Analysis	Correlation
auditory cortex	0.207
auditory	0.198
heschi	0.195
heschl gyrus	0.194
sounds	0.190
fractional anisotropy	0.189
fa	0.188
pitch	0.187
anisotropy fa	0.186
anisotropy	0.185

Figure 7: The projections of a generated brain volume(left) and its corresponding analysis report(right). The volume belongs to the class 'Non-human sound, auditory'

6 Supplementary Material

6.1 Cross-Validation

In Figure 6, the blue bar represents training data and the green bar is test data. In our case, we construct a 3-round cross-validation task to measure performance of our models. First, we partition test data to three parts and each fold still maintains similar proportion of data for each class as original test data. For each round of cross-validation, we use one of the three parts as test data and left two and training data to train a model. The evaluation scores below are the average score of the three rounds.

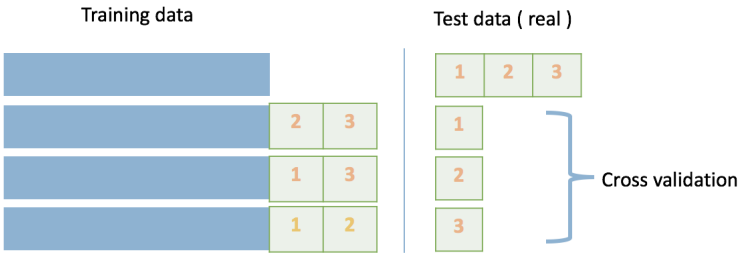


Figure 6: 3 fold cross validation.

6.2 Data Analysis

We use NeuroSynth, a brain decoding tool to analyze generated volumes qualitatively. Its principle is to compare activity regions of the given brain volume to a data pool and an analysis report about correlation to specific regions or tasks the subject was receiving. Figure 7 shows a generated brain image and an analysis report for it with top 10 rank of correlation. The supervised class of this generated brain is "Non-human sound, auditory", and several correlations with top rank are related to auditory and sound.

6.3 New Multi-Class Results

We also examine our models in a more tough multi-class task. As described before, one class in Collection 1952 consists of several labels and the collection has totally 19 labels. We encode classes in 19 dimensions, each of which dimension represents a label. The value in a dimension is 1 only if the class possesses that label, otherwise, value is 0. For example, if class 1 has label 0 and label 1, then it will be encoded as a 19 dimensional vector with the value of the fist two dimensions 1 and

Input	Accuracy	Macro F1	Precision	Recall
Real	0.4838	0.57	0.76	0.48
Real+G	**0.4992**	0.57	0.75	0.49

Table 5: New multi-class results in Collection 1952

0 in other dimensions(we can formally present it as $([1, 1, 0, ..., 0])$). This task is highly demanding since that there are larger probability space, specifically, 2^{19} possibilities, for a classifier to learn and discriminate. We use SVM to train real and mixed training data, both of which are in the resolution of $13 \times 15 \times 11$, and utilize 3 fold cross-validation as illustrated in Section 6.1 to this multi-class task. Several evaluation scores are shown in Table 2. It is shown that even in this highly demanding task, accuracy score with mixed data(the second row) used is still visually higher than the baseline(the first row).

6.4 Real brain images of Collection 1952 with labels on top

We present several real images of Collection 1952 with their labels above the 2D projections of brain images in Figure 8.

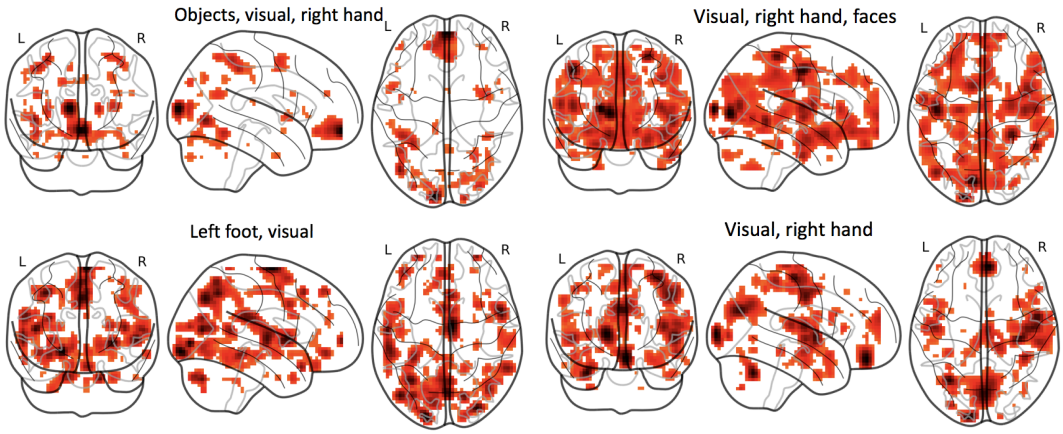


Figure 8: Real images in Collection 1952 with labels on top. Linear normalization is used and the lower threshold for plotting these images is 0.48.

6.5 More brain generations

We want to show more synthetic brain images using ICW-GAN in this section. Figure 9∼11 show generated ones from Collection 1952, 2138 and 503 respectively. For Collection 1952, classes are on the top of images.

As for Collection 2138, the corresponding categories of the classes are listed as follow:

- **14:** visual form recognition, feature comparison, response selection, response execution, relational comparison, visual pattern recognition
- **22:** response execution, working memory, body maintenance, visual body recognition
- **4:** response selection, response execution, punishment processing
- **43:** motion detection
- **2:** response selection, response execution, animacy perception, animacy decision, motion detection
- **3:** response selection, response execution, motion detection

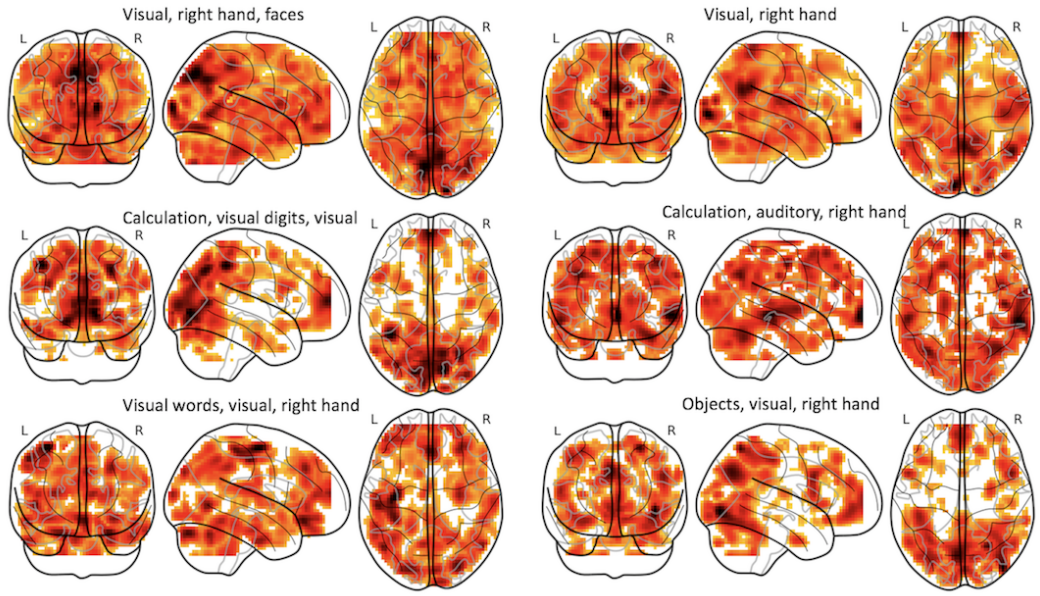


Figure 9: Brain images in Collection 1952 with labels on top. The lower threshold for plotting these images is 0.3.

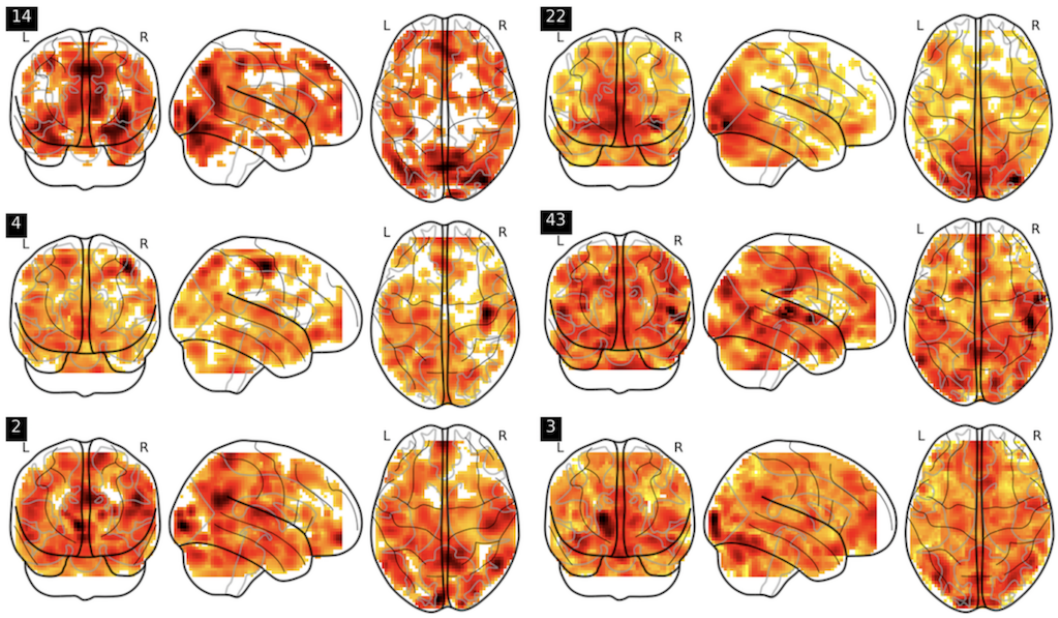


Figure 10: Brain images in Collection 2138. A left top tag is the class generated brains belong to and we list their corresponding categories below.

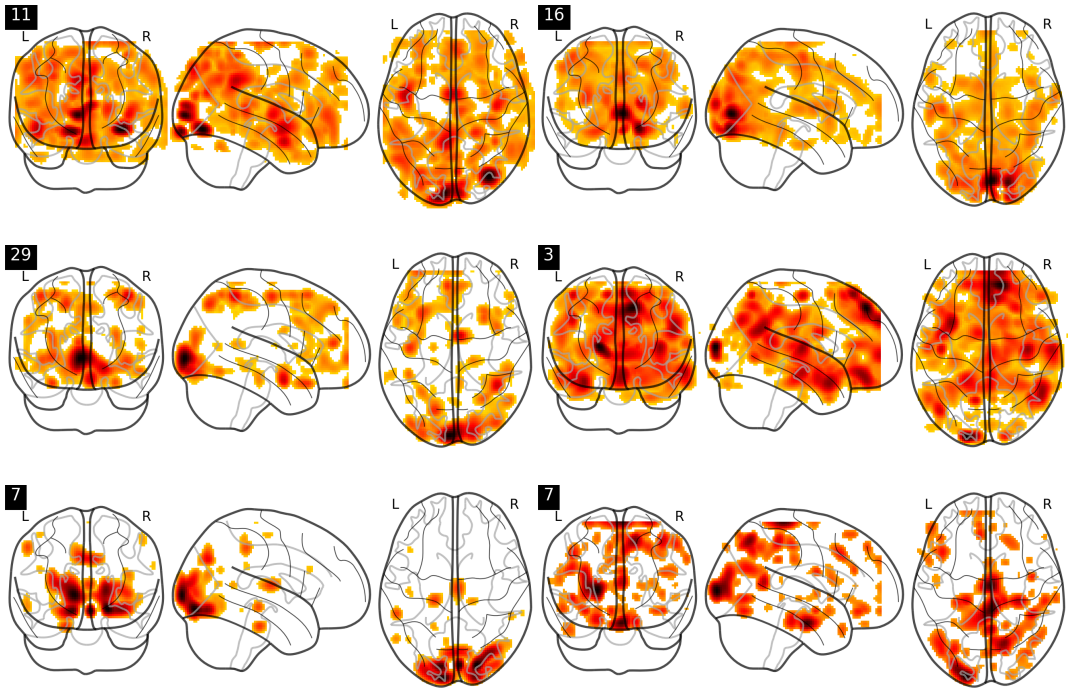


Figure 11: Brain images in Collection 503 and their left top tag represents the stimulating picture for subjects.