NeuroImage 61 (2012) 1113–1119

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs

Marc N. Coutanche *, Sharon L. Thompson-Schill

Department of Psychology, University of Pennsylvania, 3720 Walnut Street, Philadelphia, PA 19104, USA

ARTICLE INFO

Article history:
Accepted 25 March 2012
Available online 3 April 2012

Keywords:
MVPA
fMRI
Multivariate
Runs
Design
Classification

ABSTRACT

Functional magnetic resonance imaging (fMRI) studies are broken up into runs (or 'sessions'), frequently selected to be long to minimize across-run signal variations. For investigations that use multi-voxel pattern analysis (MVPA), however, employing many short runs might improve a classifier's ability to generalize across irrelevant pattern variations and detect condition-related activity patterns. We directly tested this hypothesis by scanning participants with both long and short runs and comparing MVPA performance using data from each set of runs. Every run included presentations of faces, places, man-made objects and fruit in a blocked 1-back design. MVPA performance significantly improved from using a large number of short runs, compared to several long runs, in across-run classifications with identical amounts of data. Superior classification was found across variations in the classifier employed, feature selection procedure and region of interest. Performance improvements also extended to an information brain mapping 'searchlight' procedure. These results suggest that investigators looking to maximize the detection of subtle multi-voxel patterns across runs might consider employing short fMRI runs.

Introduction

Investigators planning an fMRI study are faced with an abundance of potentially important design considerations, ranging from acquisition parameters like pulse repetition time (TR) and flip angle, to experimental paradigm decisions, such as whether to arrange trial types randomly or in blocks. One decision on which experimenters may place little import is the number and length of runs (or in SPM parlance, "sessions"). Typical fMRI studies divide the total scanning period into a small number of runs, each 5–10 min long. For many investigators, this choice of run duration is guided by conventions that have their roots in memory limitations that accompanied early fMRI applications. As these limitations have been reduced, some investigators have opted to make runs as long as possible, to reduce variability in the data across runs (a point to which we will return later). Any breaks between runs are largely thought to benefit the subject (i.e., a brief chance to relax) and not the investigator.

The breaking-up of an fMRI scan into runs has more than an organizational impact: the start of each run re-equalizes image intensity. This produces so much variability that the first several TRs must be routinely discarded early in data pre-processing. The steady-state eventually reached by each run is not identical to the last, leading investigators to typically include examples of each condition in every run, so that real activation differences between conditions are not confounded with inter-run variability. During the pre-processing stages of data analysis, data from all runs are then concatenated after some correction (e.g., mean centering) to allow the data from different runs to be analyzed together. Other than this correction, the number and length of runs are not generally given much attention in fMRI studies that use the General Linear Model to describe task-related changes in individual voxels or regions of interest (ROIs).

The focus of the current study is the effect of run length on a different type of fMRI approach: a family of analytic tools known as multi-voxel pattern analysis (MVPA). Unlike univariate analyses that examine whether voxels or regions differ in their overall response to conditions of interest, MVPA evaluates the information distributed across groups of voxels, where the response levels of each voxel alone may be uninformative. Since Haxby et al. (2001) showed that MVPA can detect information that is inaccessible to traditional univariate approaches, MVPA techniques are being applied to address an increasingly broad array of questions, including detection of stimulus orientation in visual areas (Kamitani and Tong, 2005), identification of the content of recalled items (Polyn et al., 2005), decoding of reward predictions with learning (Kahnt et al., 2011), and prediction of symptom severity in a clinical sample (Coutanche et al., 2011; for MVPA reviews see: Haynes and Rees, 2006; Mur et al., 2009; Norman et al., 2006; O'Toole et al., 2007).

For researchers employing MVPA, runs often serve an important purpose: they can act as units of independence for training and testing. In order to evaluate the information possibly present within activity patterns, a machine learning classifier is frequently trained to

* Corresponding author. Fax: +1 215 898 1982.
E-mail address: coumarc@psych.upenn.edu (M.N. Coutanche).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2012.03.076

distinguish trials of different conditions in one part of the dataset and then tested (by predicting condition labels of trials) in the remaining data. The runs of a scan are commonly used to divide the dataset into training and testing groups to help ensure these data remain independent (e.g., Haynes and Rees, 2005; Lee et al., 2011, 2012; Reddy et al., 2010; Wolbers et al., 2011; an approach suggested in Misaki et al., 2010; Mur et al., 2009, p. 106). The most common approach to allocating different runs to training versus testing sets is 'cross-validation', where each 'fold' of the data (typically one run) takes a turn as the testing set, while the other runs are used for training ('leave-one-run-out'). This run-based cross-validation method is a natural way to divide a dataset, but has additional consequences for classification, which motivate the current work.

To successfully classify activity patterns in a testing set, a classifier must generalize across data variability found within each class. One source of variability within a class arises from differences in the patterns evoked by different exemplars (or trials) of a category: A classifier trains on exemplars that each have an idiosyncratic pattern in order to learn the consistent features of that class of patterns. Indeed, the requirement of classifiers to generalize across exemplars allows one to make inferences about the representation of categories, such as "faces" or "verbs that involve hand actions". But, another source of variability within a class is less theoretically interesting and arises from the changes in image intensity across different runs discussed above. Just as classifiers must detect a consistent class signal from different exemplars and apply this to a new exemplar, they also often generalize from run-related noise in training to detect a signal in unseen testing runs. For MVPA, where distinguishing conditions can depend on subtle differences between activity patterns, training/testing inter-run variation can be an obstacle. The challenge of coping with between-run signal differences was noted recently by Misaki et al. (2010) in their comparison of MVPA approaches. The task of generalizing across runs was difficult enough that their results were "consistent with the notion that run-related changes reflecting scanner state and head motion are substantially larger than activity-pattern effects" (p. 117).

MVPA studies have largely continued the tradition of presenting trials within four to eight fMRI runs that are each 5–10 min long. However, if we were designing an MVPA study from scratch without influence from this tradition, we might have good reason to break up the scanning session into as many runs as possible. This idea follows the same reasoning regularly applied to selecting stimuli. If we wish to maximize the classification of 'chair' activity patterns from the patterns for other objects, we would ensure that a classifier is trained on trials from a variety of chair exemplars, perhaps from multiple viewpoints (as employed in Haxby et al., 2001; O'Toole et al., 2005; Spiridon and Kanwisher, 2002). Presenting a classifier with this variation would help the model find central 'chair' features and therefore generalize to unseen chair exemplars and viewpoints. Analogously, training a classifier on trials from a variety of runs may help a classifier model the signal of interest, which will remain largely consistent across runs. Structuring a scanning session to contain many short runs may therefore increase a classifier's ability to generalize to unseen exemplars in an unseen run.

A secondary benefit of increasing the run number is to increase the percentage of data that can be used to train a classifier. Classifier performance is improved from including more data in the training set (see Pereira et al., 2009 for a discussion). In a typical leave-one-run-out cross-validation approach over n runs, $(n-1)/n$ of the data is used for training. With four large runs, this would translate into 75% of the data being employed to train each iteration. With sixteen small runs, 93.75% (15/16) of the data can be used for training.

Naturally, any new run structure should continue to incorporate trials from all conditions in each run, to avoid confounding condition-relevant signals with between-run differences. As long as this point is heeded, however, there may be advantages to structuring a scanning session into many short runs.

In this study we directly compared the detection of multi-voxel information from data acquired across a small number of long runs, with data acquired across many short runs. We used a within-subject design by presenting participants with four long runs intermixed with sixteen short runs. The short runs cumulatively contained the same number of trials as all the long runs, allowing us to directly evaluate the effects of run-structure within individuals. All runs included presentations of faces, scenes, man-made objects and fruits in short blocks while participants performed a rapid 1-back task. We directly compared MVPA performance in the two sets of runs (short and long) for different classifiers, voxel-selection techniques, regions, and in a spherical searchlight analysis ('information-based brain mapping'; Kriegeskorte et al., 2006). These analyses seek to answer the question of whether the typical fMRI design of a small number of long runs could be switched to a large number of short runs for MVPA studies looking to maximize the detection of information across runs.

Materials and methods

Participants

Ten participants (9 females, mean age = 23 years, range = 19–27 years) were recruited for this study. All participants were right-handed with normal or corrected-to-normal vision and reported no history of neurological problems. All participants provided written informed consent and received monetary compensation for their participation. The human subjects review board at the University of Pennsylvania approved all experimental procedures.

Stimuli and experimental design

Participants were presented with a series of images, while undergoing fMRI. Following an anatomical scan, participants underwent functional imaging as they viewed blocks of faces, scenes, man-made objects and fruit.

Participants received a total of twenty functional runs: four runs of 120 TRs ('long runs') and sixteen runs of 36 TRs ('short runs'). Every short run contained one block of each category, giving a combined total of sixteen blocks of each stimulus type in the short runs. The four long runs each contained four blocks of each image category, also giving a combined total of sixteen blocks for each stimulus type.[1] The blocks of stimuli were randomly ordered within each run. Each block contained ten images, including one randomly selected repeat. Participants were instructed to respond on a button-box when they saw an image repeat (1-back task). Images were presented for 400 ms followed by a 500 ms interstimulus interval, producing blocks that were 9 s long. Blocks were separated by 12 s of rest. Participants saw 144 unique color photographic images of each stimulus category. Each image was used once in the short runs and once in the long runs, in a random order in both cases. As all of the analyses use data from *either* short or long runs, the MVPA classifiers are only exposed to one presentation of each image (excepting immediate repeats from the 1-back task).

The two kinds of run (short and long) were intermixed in each participant (illustrated in Fig. 1). In order to ensure that the runs of each type were equally close in time, the order of the twenty runs was determined as follows. First, 10,000 possible run orders were randomly generated. For each theoretical sequence, the mean number of TRs separating all of the sixteen short runs and the mean

[1] The long runs were slightly shorter than four short runs combined as each run contained four pseudo-TRs (removed in pre-processing) and two TRs of fixation at the beginning and end.

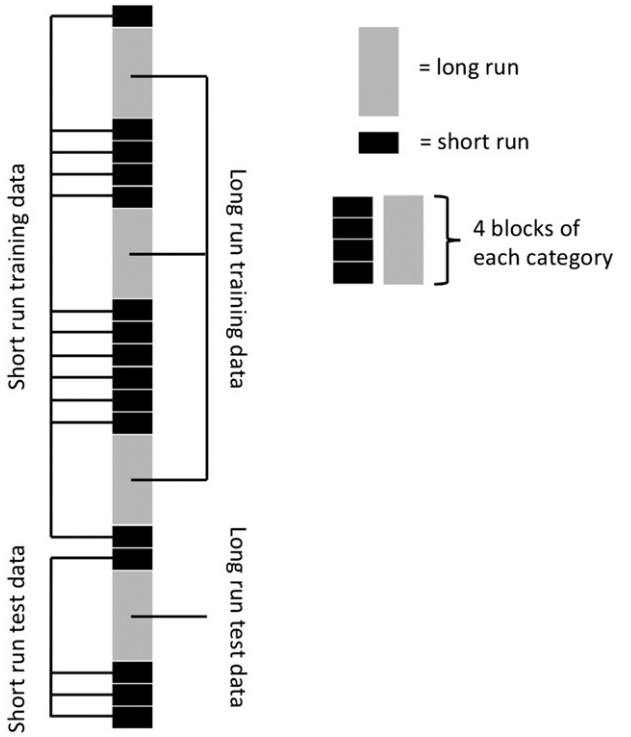

Fig. 1. An example of the design. Each participant experienced both long and short runs in a pseudo-random sequence. Long runs contained the same amount of data as four short runs. In four-fold cross-validation, training data consisted of either three long runs (testing on the fourth) or twelve short runs (testing on the remaining four).

number separating all of the four long runs were calculated. Five run sequences with identical mean TR distances for short and long runs were randomly selected for use in the experiment. Participants were randomly allocated to one of these matched sequences. This procedure allowed us to have confidence that any classification differences between short and long runs were not due to runs of one length being closer in time to each other.

Image acquisition

Imaging data were acquired with a 3T Siemens Trio system with eight-channel head coil and foam padding for stabilizing the head. T1-weighted anatomical images were acquired at the beginning of each session (TR = 1620 ms, TE = 3 ms, TI = 950 ms, voxel size = 0.977 mm × 0.977 mm × 1.000 mm). This was followed by BOLD imaging using interleaved gradient-echo EPI (TR = 3000 ms, TE = 30 ms, field of view = 19.2 cm × 19.2 cm, voxel size = 3.0 mm × 3.0 mm × 3.0 mm) for twenty runs. Sixteen runs contained 36 TRs and four contained 120 TRs. The run order was randomized according to the procedure described above.

Imaging preprocessing

Imaging data were preprocessed using the Analysis of Functional NeuroImages (AFNI) software package (Cox, 1996). The first four volumes of each functional run were first removed to allow the signal to reach steady-stage magnetization. All functional images were slice time corrected and a motion correction algorithm registered all volumes to a mean functional volume (Cox and Jesmanowicz, 1999). Low frequency trends were removed from all runs using a high-pass filter threshold of 0.0159 Hz (1/three times the stimulus onset asynchrony). Voxel activation was scaled to have a mean of 100, with a maximum limit of 200. Two of the ten participants were removed from further analyses because of excessive motion during the functional scan, leaving eight participants for the target analyses.

Multi-voxel pattern analysis

The preprocessed functional data, reflecting voxel response amplitudes at each TR, were analyzed using MATLAB. Each voxel's values were z-scored within each run. The condition labels were convolved with a time-shifted model of the hemodynamic response (compensating for the hemodynamic delay). The convolved condition function was thresholded at 0.5 for each TR to give three time points (TRs) in each block. Multi-voxel analyses were then conducted using the pattern vector (n voxels long) associated with each TR, labeled according to the shifted condition timing (as employed in, for example, Haynes and Rees, 2005; Rissman et al., 2010; Wolbers et al., 2011).

A cross-validation procedure was employed for classification. This approach separates the data into folds and takes turns to test on one fold while training on the rest. The primary aim of this study was to investigate the impact of run length on MVPA performance, so for many of the subsequent analyses we controlled the amount of training data by applying a 4-fold cross-validation structure to both long and short runs, producing folds containing one long run or four short runs respectively. In other analyses, we employed leave-one-run-out cross-validation, giving 16 folds for the short runs and 4 folds for the long runs.

MVPA was performed using all voxels in the ROIs and also on visually responsive voxels from each ROI. Visually responsive voxels were selected based on an orthogonal ANOVA of whether each voxel in the ROI had significantly different levels of activation for all image trials (collapsed across condition) and fixation. To protect against a possible selection bias (Kriegeskorte et al., 2009), we based the voxel selection on data separate from those used in classification: we used an ANOVA across short runs to select voxels for use in the long run analyses, and across long runs to select voxels for use in the short run analyses.

Several types of classifiers were used to examine the consistency of results across different classification approaches. We employed a Gaussian Naïve Bayes (GNB) classifier, frequently used in MVPA studies, to perform four-way classifications through the MATLAB Statistics Toolbox. We also conducted a linear discriminant analysis (LDA) classification approach, through the MATLAB Statistics Toolbox. To reduce the dimensionality of the data to lower than the number of trials, we applied a principle component analysis (PCA) immediately prior to the LDA (e.g., as used in Carlson et al., 2003). In each cross-validation fold, we selected the top principle components that accounted for 95% of the training data variance. These components were then applied to the testing data and used as features for the classifier. We also applied a frequently used correlation-based minimum-distance classifier. For each cross-validation fold, the activity patterns for testing trials were correlated with the mean pattern of each class in the training data. The class with the highest correlation value reflected the classifier's prediction for each testing trial. Pairwise class comparisons were examined using a linear support vector machine (SVM) classifier through the MATLAB Bioinformatics Toolbox. The SVM 'C' parameter was set at 1.0 for all classifications, although we also employed an embedded cross-validation procedure to select the parameter for each classification. This approach explored a range of C parameter values (0.001, 0.01, 0.1, 1, 10, 100) by training and testing within the 'training' data. This embedded cross-validation trained and tested across three folds for both long and short run analyses. The parameter giving the best performance in the training data of a fold was then used for classifying that fold's independent testing set.

We also assessed MVPA performance through a searchlight analysis (Kriegeskorte et al., 2006) to examine whether any classification

improvements extend to this approach. Three-dimensional searchlight clusters with a 3-voxel radius were mapped onto the VT ROI (described below), producing volumes of up to 123 voxels when not restricted by the region's boundaries. Each searchlight was subjected to cross-validation using a GNB classifier, as described above. The searchlight analyses were performed separately in the long and short runs, producing two classification values for every searchlight. Each searchlight's classification performance was allocated to its central voxel, creating a vector of accuracies for each run-type.

Regions of interest

Each subject's T1 anatomical image was subjected to automated cortical reconstruction and volumetric segmentation (Fischl et al., 2002) using the FreeSurfer image analysis package. We focused our analyses on two ROIs known to contain information about visually presented stimuli: gray matter of the ventral temporal (VT) lobes and gray matter of the occipital lobe. The VT region has an established role in high-level visual processing (Epstein and Kanwisher, 1998; Ishai et al., 1999; Kanwisher et al., 1997; Tanaka, 1996). Findings of distributed information relating to visual categories in this region (Haxby et al., 2001) prompted a number of MVPA studies examining VT cortex (Cox and Savoy, 2003; O'Toole et al., 2005; Spiridon and Kanwisher, 2002). The occipital lobe was selected as a second ROI for its central role in visual perception (Grill-Spector and Malach, 2004). This region has also been a successful target for multi-voxel pattern investigations of visual processing, making it of interest here (Haynes and Rees, 2005; Kamitani and Tong, 2005; Serences et al., 2009).

The VT ROI was constructed from segmented gray matter of the parahippocampal, inferior temporal, fusiform and lingual gyri. The fusiform and lingual gyri extend into the occipital lobe, so these structures were cut off at the approximate border of the occipital lobe. This gave a VT volume of 1594–2234 voxels (M = 1849, s.d. = 184). The occipital ROI was constructed from segmented gray matter of the lateral occipital gyrus, cuneus, pericalcarine, calcarine sulcus and occipital parts of the fusiform and lingual gyri, giving a volume of 1278–1753 voxels (M = 1495, s.d. = 161).

Results and discussion

Behavioral results

Before examining MVPA performance, we assessed whether task performance differed in the long and short runs, to give confidence that any differences in MVPA performance are not due to attentional effects. Behavioral results were not available for one participant due to a technical difficulty. In the remainder of the participants, however, performance in the fast 1-back task (using a d′ measure to consider hits and false positives) did not differ (paired $t_6 = 0.62$, $p = 0.56$) between the short runs (M = 3.68, s.d. = 0.66) and long runs (M = 3.58, s.d. = 0.51).

MVPA results for different run lengths

Our primary question was whether MVPA performance differed when using data from several long runs compared to many short runs, for equal amounts of training data. We performed 4-way classifications using 75% of the data for training and 25% for testing (4-fold cross-validation) using GNB, correlation-based and LDA classifiers. Classifying with data from short runs gave greater performance than with data from long runs across all classifiers in both the VT and occipital regions (see Table 1 for full results). The performance improvements ranged from 6.8% (GNB in occipital) to 8.8% (correlation-based in VT), where chance was at 25%. The VT results are displayed in Fig. 2.

These findings were not driven by any systematic motion differences between the runs: participants' movement levels were not significantly different for short and long runs (paired $t_7 = 1.29$, $p = 0.24$), with the mean values, if anything, toward larger motion in the set of short runs. We also verified that the short run advantage was not dependent on grouping the runs into folds sequentially (e.g., the first four short runs forming the first fold). Although the runs were deliberately grouped in this manner to match the long runs as much as possible, we also randomly assigned short runs to folds in 100 permutations. We then conducted a 4-way GNB classification in VT cortex for each permutation, averaging the results for each participant. The mean classification performance from the different 4-fold permutations in short runs (M = 0.64, s.d. = 0.11) continued to be greater than performance using long runs (paired $t_7 = 11.42$, $p < 0.001$). The same analysis in the occipital lobe also yielded short run classification performance (M = 0.57, s.d. = 0.08) that was greater than for long runs (paired $t_7 = 3.71$, $p = 0.008$). We examined two separate regions in this study to assess if short run benefits are consistent across independent regions of voxels. Combining the two regions into a larger occipitotemporal area did not disrupt the advantage: a 4-fold GNB analysis in the combined region using short runs (M = 0.63, s.d. = 0.10) gave greater classification performance (paired $t_7 = 3.75$, $p = 0.007$) than using long runs (M = 0.55, s.d. = 0.11).

MVPA performance is frequently improved by selecting relevant features, a process that aims to restrict analyses to informative voxels. We examined whether the short-run advantage would

Table 1
Short and long run classification performance for combinations of regions, feature selection techniques and 4-way classifiers. Mean classification accuracies are listed with standard deviations in parentheses. Chance performance is at 0.25. All classifiers were run in a 4-fold cross-validation structure with the same amount of training and testing data for each type of run. Mean voxel counts (and standard deviations) are listed for ROIs and sets of visually responsive features used in long (L) and short (S) run analyses. Each *p*-value reflects the outcome of a two-tailed paired *t*-test comparing classification performance using data from short runs versus using data from long runs (without correction for multiple comparisons).

	Voxel count	GNB classifier			Correlation-based classifier			LDA classifier		
		Long run performance	Short run performance	*p*	Long run performance	Short run performance	*p*	Long run performance	Short run performance	*p*
Anatomical ROI: VT lobe	1849 (184)	0.56 (0.10)	0.64 (0.12)	0.003*	0.59 (0.11)	0.68 (0.10)	0.009*	0.67 (0.09)	0.76 (0.04)	0.009*
Anatomical ROI: occipital lobe	1495 (161)	0.50 (0.12)	0.56 (0.09)	0.02*	0.58 (0.15)	0.66 (0.09)	0.02*	0.69 (0.11)	0.76 (0.06)	0.057
Visually responsive VT voxels	L: 953 (101) S: 912 (138)	0.61 (0.10)	0.68 (0.11)	0.002*	0.65 (0.09)	0.76 (0.06)	<0.001*	0.71 (0.06)	0.80 (0.03)	0.002*
Visually responsive occipital voxels	L: 1252 (181) S: 1227 (206)	0.50 (0.11)	0.57 (0.09)	0.01*	0.60 (0.13)	0.69 (0.09)	0.005*	0.71 (0.10)	0.77 (0.06)	0.03*

* Statistically significant at $p < 0.05$.

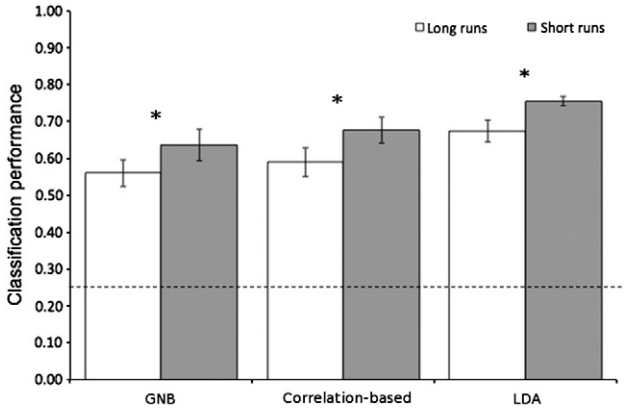

Fig. 2. Classification performance using data from long and short runs in the ventral temporal lobes. Accuracies were calculated from four-way classifications of presentations of faces, scenes, man-made objects and fruits during a 1-back task, using Gaussian Naïve Bayes, correlation-based and linear discriminant analysis classifiers. The dashed line indicates chance performance. Asterisks denote statistical significance in a two-tailed paired *t*-test comparing long and short run classification performance ($p<0.05$). Error bars reflect the standard error of the mean.

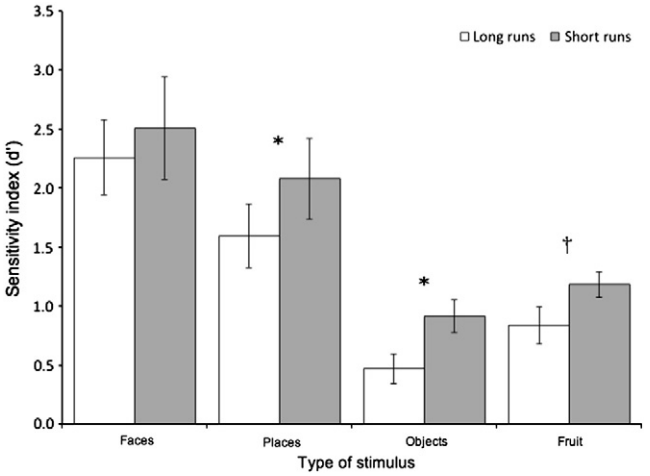

Fig. 3. Sensitivity for detecting each class of stimulus using data from long and short runs in the ventral temporal lobes. D-prime scores were calculated from four-way Gaussian Naïve Bayes classifications of presentations of faces, scenes, man-made objects and fruits during a 1-back task. Asterisks denote statistical significance in a two-tailed paired *t*-test comparing long and short run d′ scores ($p<0.05$). A cross symbol represents a statistical trend ($p<0.10$). Error bars reflect the standard error of the mean.

remain after a typical feature selection procedure. We selected visually responsive voxels in each region and then conducted MVPA on these restricted voxel sets for the short and long run data (using voxels selected in one run length for classifications in the other, to ensure independence). The numbers of selected voxels are listed in Table 1. There was substantial overlap between the long and short run visually responsive voxel sets: an average of 80% (s.d. = 2%, range = 76%–84%) of voxels in the smallest feature set were included in the larger set for VT cortex, and 93% (s.d. = 2%; range = 89%–95%) of voxels overlapped in the two occipital feature sets. The numbers of features selected were not significantly different for VT (paired $t_7 = 1.38$, $p = 0.21$) or occipital (paired $t_7 = 1.56$, $p = 0.16$) voxels. MVPA using these selected features followed the prior pattern of results: classification performance using short run data was significantly higher than using long run data for the GNB, correlation-based and LDA classifiers (Table 1).

The preceding analyses all compared short and long run designs when the same amount of data was used for each (i.e., train on 75%, test on 25%); however, another benefit of employing short runs is having the opportunity to train on a greater proportion of the data, in this case 15/16 instead of 3/4. We compared performance for the different fold structures using a GNB classifier. Classifications using the short runs in a 16-fold structure (M = 0.59, s.d. = 0.08; chance = 0.25) gave significantly greater classification performance than using short runs in a 4-fold structure (M = 0.56, s.d. = 0.09) in the occipital lobe (paired $t_7 = 3.40$, $p = 0.01$), although this did not reach significance in the VT area (16-fold: M = 0.66, s.d. = 0.10; paired $t_7 = 1.57$, $p = 0.16$).

We examined how different stimulus classes were affected by the run structures. We calculated category sensitivity indices (d′, which accounts for hits and false positives in the classifiers' predictions) from the results of the 4-fold GNB analyses conducted in each ROI. This yielded four d′ values from each set of classification predictions. These results are displayed in Figs. 3 and 4. In the VT area, places (paired $t_7 = 2.97$, $p = 0.02$), man-made objects (paired $t_7 = 2.86$, $p = 0.02$) and fruit (paired $t_7 = 2.24$, $p = 0.06$) had significantly or close-to-significantly greater sensitivity from classifications using the short runs than using the long runs, but this comparison did not reach significance for faces (paired $t_7 = 1.34$, $p = 0.22$). The occipital lobe showed greater sensitivity in the short runs for places (paired $t_7 = 2.51$, $p = 0.04$) and man-made objects (paired $t_7 = 2.61$, $p = 0.03$), although this only approached significance for faces (paired $t_7 = 1.90$, $p = 0.10$) and fruit (paired $t_7 = 1.68$, $p = 0.14$).

Two-way classifications

We were interested in whether short run improvements would generalize to pairwise comparisons, so we employed 2-way linear SVMs on VT voxels in a 4-fold cross-validation structure. The mean accuracies for the six possible 2-way comparisons in long runs (M = 0.86, s.d. = 0.04) and short runs (M = 0.90, s.d. = 0.05) were only significantly different at a trend level (paired $t_7 = -1.87$, $p = 0.10$). A similar result was found when the C parameter was optimized in the training data of each classification (paired $t_7 = -1.98$, $p = 0.09$). Five of these comparisons involved very localizable stimuli (faces and places), which have produced at- or near- ceiling classification performance in previous investigations, unlike the more difficult comparison of different object types (e.g., see Fig. 3 in Spiridon and Kanwisher, 2002). We therefore had an a priori hypothesis that the pairwise comparison of fruit and man-made objects would be most sensitive to improvements from using many short runs (if for no other reason than classification performance being off-ceiling). This was borne out: fruit and man-made object

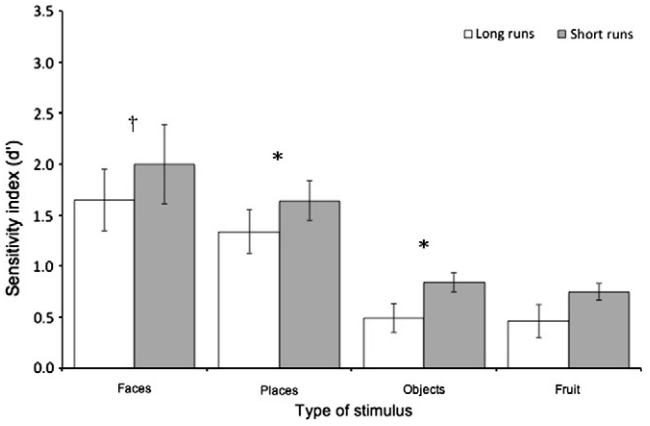

Fig. 4. Sensitivity for detecting each class of stimulus using data from long and short runs in the occipital lobes. D-prime scores were calculated from four-way Gaussian Naïve Bayes classifications of presentations of faces, scenes, man-made objects and fruits during a 1-back task. Asterisks denote statistical significance in a two-tailed paired *t*-test comparing long and short run d′ scores ($p<0.05$). A cross symbol represents a statistical trend ($p<0.10$). Error bars reflect the standard error of the mean.

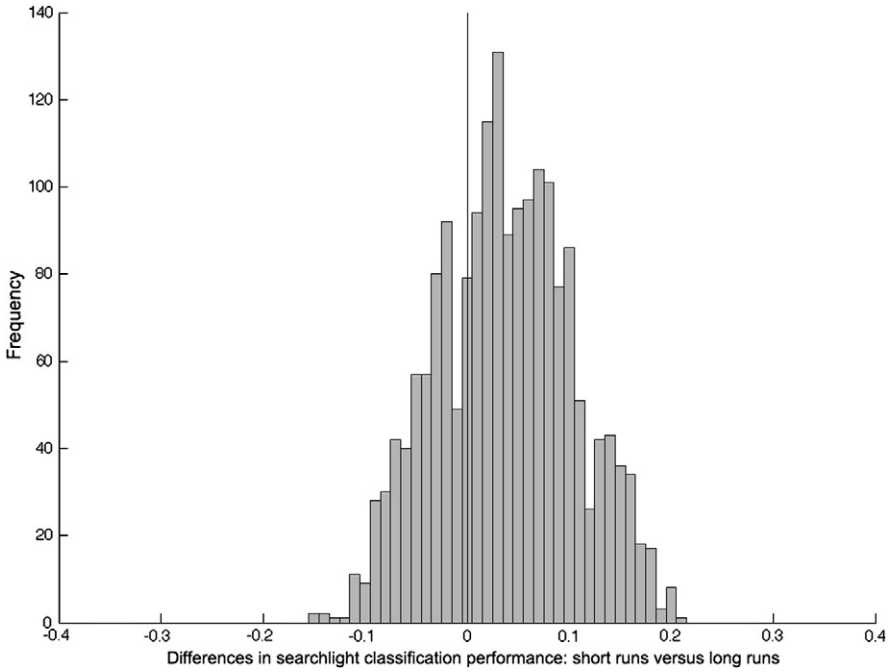

Fig. 5. A histogram of differences in classification performance using short run data versus long run data for ventral temporal searchlights of a representative participant. Two sets of searchlight analyses were conducted in each participant: based on data from short runs and data from long runs. For each searchlight, classification performance using long run data was subtracted from classification performance using short run data to give an improvement score for that searchlight. All improvement scores were then plotted in a histogram. The histogram's shift from zero reflects the higher frequency and larger magnitude of performance improvements from using short runs compared to long runs.

classifications with short runs (M = 0.70, s.d. = 0.10; chance = 0.50) showed significantly higher performance (paired $t_7 = 3.38$, $p = 0.01$) than with long runs (M = 0.58, s.d. = 0.07). This was also found when the C parameter was optimized in the training data of each classification (paired $t_7 = 3.57$, $p = 0.009$). This finding that short runs benefited this difficult comparison is important because many studies apply MVPA to classify categories that are not easily discriminable with univariate analyses. It is precisely these difficult classifications that may benefit most from short runs.

Searchlight analyses

A successful approach for examining the information content of regions has been to conduct information brain mapping, where a series of searchlights are systematically applied. To examine if using many short runs would also have an advantage with this method, we conducted searchlight analyses with a 4-way GNB classifier (with 4-folds) in the VT area using short runs and then long runs. Mean searchlight performance was significantly greater (paired $t_7 = 4.02$, $p = 0.005$) using short run data (M = 0.37, s.d. = 0.03; chance = 0.25) than long run data (M = 0.35, s.d. = 0.03). The general level of searchlight performance was likely lower than performance from using the entire ROI for several reasons, including the influence of uninformative searchlights on mean accuracy (bringing it closer to chance) and the ROI analysis having access to widely distributed information.

Classification using short runs also showed greater performance in direct comparisons of all searchlights within participants: in paired comparisons of all searchlights, short run data gave significantly higher performance than long run data for all but one participant (seven at $p < 0.001$; one at $p < 0.001$ for the reverse direction).[2] The distribution of searchlight improvements (displayed for a representative participant in Fig. 5) included both positive and negative values, indicating that not every searchlight had a higher accuracy value using short runs. Noisy fluctuations of classification scores around chance may account for these 'long > short' searchlights. Consistent with this explanation, 'short run–long run' performance differences (i.e., short run improvement scores) were significantly positively correlated with *overall* searchlight information (the mean of both long and short run accuracies together) in all but one participant ($p < 0.001$ in 7, $p = 0.097$ in 1). In other words, searchlights with greater benefits from long runs tended to be closer to chance in overall accuracy (collapsed across run-type) than searchlights benefiting from short runs. Notably, the one participant showing significant long run improvements in the previous within-participant comparison of all searchlights had the most positive correlation in this latter analysis, hinting that fluctuations around chance may have at least partially driven their result.

Discussion

Of the many design decisions fMRI experimenters must face, we suspect that run length is a factor that is not routinely given much thought; however, we have shown here that breaking up an fMRI session into a large number of runs can enhance MVPA classification performance in across-run cross-validation. This benefit was seen when the quantity of data was held constant, suggesting that this advantage comes from the number of runs. This result was apparent across different regions, for multiple stimuli classes, using several classifiers, and in a roaming searchlight analysis.

It is a common practice for fMRI investigators to present participants with variations in stimuli, such as different chair exemplars, to help identify a core signal from signal fluctuations that are due to other characteristics. Although we frequently incorporate such variance into our selections of stimuli, another source of signal variance, fluctuations across scanning runs, is viewed more as a necessary nuisance than an opportunity to hone in on a consistent signal. Investigations that rely on extracting distributed patterns of activity (MVPA) often depend on identifying a consistent (usually

[2] Short runs continued to give greater performance when non-overlapping independent searchlights were examined: 25 randomly sampled independent searchlights had higher mean short run performance in an average of 97.1 of 100 permutations (s.d. = 4.5) in the 7 participants showing the effect.

subtle) signal that will then generalize to new 'unseen' stimuli and runs. The findings presented here suggest that structuring a scanning session to contain a large number of short runs can help with identifying activity patterns that reflect conditions of interest in new unseen runs. Our findings were especially apparent for categories of stimuli that are not easily discriminable with univariate analyses, making the results particularly important for conditions represented by subtle activity patterns, i.e., those often examined with MVPA.

Our study suggests that exposing a classifier to a large number of short runs can improve its ability to generalize to new runs. Although this advantage is found when the amount of training data is held constant, we also confirmed a further benefit of increasing the proportion of data used for training. In this study, the combined benefits of having more runs and a larger proportion of training data gave a performance boost of 10 percentage points (with chance at 25%) in the VT area. As we were restricted to using half the time of a typical one-hour scanning session for the short runs (the rest of the session included longer runs), the magnitude of these benefits may be greater for investigations that employ short runs for a full session.

This investigation speaks to the benefits of shortening runs for detecting patterns in new scanner runs, but we acknowledge that other factors may sometimes preclude this design: It is still the case that examples from all conditions should be included in every run where possible, which could necessitate longer runs if the stimulus duration or number of conditions is high. Similarly, counterbalancing procedures may influence the run length (and therefore number) ultimately selected for an experiment.

The detection of activity patterns in new scanner runs is particularly relevant for investigations that employ cross-validation across runs (e.g., Haynes and Rees, 2005; Lee et al., 2011, 2012; Reddy et al., 2010; Wolbers et al., 2011). If a study is conducting within-run comparisons, however, where the investigators are confident that this will not impact the conclusions drawn, other design options may produce greater sensitivity to patterns. For example, having one very long run would avoid run-variability, although this in turn may require additional considerations (e.g., to protect against standard preprocessing steps producing dependencies between time points; Mur et al., 2009, p. 106). Similarly, it is also important to note that while our findings were evident for an experiment using a block design, we cannot conclude that this will extend to other design options from these data alone. A vast number of analysis choices are available to investigators using MVPA. We observed performance improvements across several classification methods, however, it is possible that this will not be seen for all MVPA decisions and approaches. Future investigations are required to understand the boundaries of the 'many versus few' design decision.

In summary, we have demonstrated that structuring an fMRI scanning session to have many short runs can give greater MVPA classification performance than employing fewer long runs in studies generalizing across runs. This benefit is found when the amount of data and cross-validation approach is held constant, suggesting that exposures to trials from a variety of runs strengthen subsequent generalization to new runs and activity pattern detection. MVPA investigators may wish to consider increasing the number of runs in a scanning session to take advantage of this effect.

Acknowledgments

We thank Avi Chanales and Carol Gianessi for assistance with stimulus preparation, and members of the Thompson-Schill lab for helpful discussions. We thank the anonymous reviewers for useful comments and suggestions. This work was funded by NIH grant R01MH070850 (to ST-S). MC is funded by a fellowship from the Howard Hughes Medical Institute.

References

Carlson, T.A., Schrater, P., He, S., 2003. Patterns of activity in the categorical representations of objects. J. Cogn. Neurosci. 15 (5), 704–717.
Coutanche, M.N., Thompson-Schill, S.L., Schultz, R.T., 2011. Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. Neuroimage 57 (1), 113–123.
Cox, R.W., 1996. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput. Biomed. Res. 29 (3), 162–173.
Cox, R.W., Jesmanowicz, A., 1999. Real-time 3D image registration for functional MRI. Magn. Reson. Med. 42 (6), 1014–1018.
Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19 (2 Pt 1), 261–270.
Epstein, R., Kanwisher, N., 1998. A cortical representation of the local visual environment. Nature 392 (6676), 598–601.
Fischl, B., Salat, D.H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., et al., 2002. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 33 (3), 341–355.
Grill-Spector, K., Malach, R., 2004. The human visual cortex. Annu. Rev. Neurosci. 27, 649–677.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293 (5539), 2425–2430.
Haynes, J.-D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8 (5), 686–691.
Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7 (7), 523–534.
Ishai, A., Ungerleider, L.G., Martin, A., Schouten, J.L., Haxby, J.V., 1999. Distributed representation of objects in the human ventral visual pathway. Proc. Natl. Acad. Sci. 96 (16), 9379–9384.
Kahnt, T., Heinzle, J., Park, S.Q., Haynes, J.-D., 2011. Decoding the formation of reward predictions across learning. J. Neurosci. 31 (41), 14624–14630.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8 (5), 679–685.
Kanwisher, N., McDermott, J., Chun, M.M., 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. J. Neurosci. 17 (11), 4302–4311.
Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103 (10), 3863–3868.
Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12 (5), 535–540.
Lee, Y.S., Janata, P., Frost, C., Hanke, M., Granger, R., 2011. Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. Neuroimage 57 (1), 293–300.
Lee, Y.S., Turkeltaub, P., Granger, R.H., Raizada, R.D.S., 2012. Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. J. Neurosci. 32 (11), 3942–3948.
Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53 (1), 103–118.
Mur, M., Bandettini, P.A., Kriegeskorte, N., 2009. Revealing representational content with pattern-information fMRI—an introductory guide. Soc. Cogn. Affect. Neurosci. 4 (1), 101–109.
Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10 (9), 424–430.
O'Toole, A.J., Jiang, F., Abdi, H., Haxby, J.V., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. J. Cogn. Neurosci. 17 (4), 580–590.
O'Toole, A.J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J.P., Parent, M.A., 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J. Cogn. Neurosci. 19 (11), 1735–1752.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45 (1 Suppl.), S199–S209.
Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A., 2005. Category-specific cortical activity precedes retrieval during memory search. Science 310 (5756), 1963–1966.
Reddy, L., Tsuchiya, N., Serre, T., 2010. Reading the mind's eye: decoding category information during mental imagery. Neuroimage 50 (2), 818–825.
Rissman, J., Greely, H.T., Wagner, A.D., 2010. Detecting individual memories through the neural decoding of memory states and past experience. Proc. Natl. Acad. Sci. 107 (21), 9849–9854.
Serences, J.T., Ester, E.F., Vogel, E.K., Awh, E., 2009. Stimulus-specific delay activity in human primary visual cortex. Psychol. Sci. 20 (2), 207–214.
Spiridon, M., Kanwisher, N., 2002. How distributed is visual category information in human occipito-temporal cortex? An fMRI study. Neuron 35 (6), 1157–1165.
Tanaka, K., 1996. Inferotemporal cortex and object vision. Annu. Rev. Neurosci. 19 (1), 109–139.
Wolbers, T., Zahorik, P., Giudice, N.A., 2011. Decoding the direction of auditory motion in blind humans. Neuroimage 56 (2), 681–687.