NeuroImage 146 (2017) 1038–1049

Contents lists available at ScienceDirect

NeuroImage

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

BrainNetCNN: Convolutional neural networks for brain networks; towards predicting neurodevelopment

Jeremy Kawahara[a,1], Colin J. Brown[a,1], Steven P. Miller[b], Brian G. Booth[a], Vann Chau[b], Ruth E. Grunau[c], Jill G. Zwicker[c], Ghassan Hamarneh[a,*]

[a] Medical Image Analysis Lab, Simon Fraser University, Burnaby, BC, Canada
[b] Department of Paediatrics, The Hospital for Sick Children and the University of Toronto, Toronto, ON, Canada
[c] Child and Family Research Institute and the University of British Columbia, Vancouver, BC, Canada

ARTICLE INFO

Keywords:
Convolutional neural networks
Brain networks
Preterm infants
Diffusion MRI
Prediction
Connectome
Deep learning
Neurodevelopment

ABSTRACT

We propose BrainNetCNN, a convolutional neural network (CNN) framework to predict clinical neurodevelopmental outcomes from brain networks. In contrast to the *spatially* local convolutions done in traditional image-based CNNs, our BrainNetCNN is composed of novel edge-to-edge, edge-to-node and node-to-graph convolutional filters that leverage the *topological locality* of structural brain networks. We apply the BrainNetCNN framework to predict cognitive and motor developmental outcome scores from structural brain networks of infants born preterm. Diffusion tensor images (DTI) of preterm infants, acquired between 27 and 46 weeks gestational age, were used to construct a dataset of structural brain connectivity networks. We first demonstrate the predictive capabilities of BrainNetCNN on synthetic phantom networks with simulated injury patterns and added noise. BrainNetCNN outperforms a fully connected neural-network with the same number of model parameters on both phantoms with focal and diffuse injury patterns. We then apply our method to the task of joint prediction of Bayley-III cognitive and motor scores, assessed at 18 months of age, adjusted for prematurity. We show that our BrainNetCNN framework outperforms a variety of other methods on the same data. Furthermore, BrainNetCNN is able to identify an infant's postmenstrual age to within about 2 weeks. Finally, we explore the high-level features learned by BrainNetCNN by visualizing the importance of each connection in the brain with respect to predicting the outcome scores. These findings are then discussed in the context of the anatomy and function of the developing preterm infant brain.

1. Introduction

Preterm birth places infants at a higher risk for a variety of cognitive and neuromotor challenges. Despite decreasing mortality rates for preterm infants due to improving care, the rate of preterm birth is increasing in nearly every country, world-wide (where birth statistics are available) (World Health Organization, 2014). With information about specific brain injuries or abnormalities shortly after birth (i.e., via brain imaging), it may be possible to predict neurodevelopmental outcomes and potentially even improve those outcomes through targeted early interventions (Back and Miller, 2014; Bear, 2004). However, prediction of cognitive and neuromotor outcomes remains a challenging problem due to the complexity of the developing infant brain and the large number of confounding factors which may influence development (Brown et al., 2014). Some recent studies have used topological features from structural brain networks, derived from diffusion tensor images (DTI), to classify normal from abnormally low scores of general neurological and neuromotor function (Brown et al., 2015; Ziv et al., 2013). Other studies have confirmed that DTI-based features, such as fractional anisotropy (FA) in certain regions of the brain are correlated with neurodevelopmental outcomes of preterm infants (Ball et al., 2015; Chau et al., 2013).

Here, we use DTI-derived structural brain connectivity networks (i.e., connectomes) of preterm infants to predict Bayley-III cognitive and motor scores, assessed at 18 months of age, adjusted for prematurity. While direct prediction of the scores (i.e., regression) is perhaps a harder problem than prediction of abnormality (i.e., 2-class classification), having an actual predicted score may be more informa-

* Corresponding author at: TASC 9417, School of Computing Science, Simon Fraser University, 8888 University Drive, Burnaby, BC, Canada V5A 1S6.
E-mail address: hamarneh@sfu.ca (G. Hamarneh).
[1] Joint co-authors.

http://dx.doi.org/10.1016/j.neuroimage.2016.09.046
Received 6 July 2016; Accepted 19 September 2016
Available online 28 September 2016
1053-8119/

tive of the infant's development. To perform this prediction task, we employ a deep learning approach.

Artificial Neural Networks (ANNs),[2] specifically Convolutional Neural Networks (CNNs), have had much success lately in performing prediction tasks on medical image data (Cireşan et al., 2012, 2013; Roth et al., 2015). CNNs are especially useful when important features are too complex to be designed or even easily recognized by human observers (LeCun et al., 2015). In this paper, we propose BrainNetCNN, a novel type of CNN with specially designed *edge-to-edge*, *edge-to-node* and *node-to-graph* convolutional layer types for brain network data. These novel layer types are actually specific cases of more general convolutional filters that have meaningful interpretations in terms of network topology. BrainNetCNN is the first deep learning framework with architecture *designed specifically* for brain network data.

We validate our BrainNetCNN on both synthetic graph data and DTI-derived structural brain networks of preterm infants. Our infant dataset consists of 168 DTI images from a cohort of infants born very preterm and scanned between 27 and 45 weeks postmenstrual age (PMA). Due to the relatively few number of training instances available, a problem common to many neuroimaging applications, CNNs are advantageous as they share weights within layers which can reduce the number of free parameters to learn when compared to fully connected neural networks. We first demonstrate this in controlled experiments on synthetic graph data by showing that BrainNetCNN outperforms a fully connected neural-network with the same number of model parameters.

On the preterm infant connectome data, we first test BrainNetCNN with the task of predicting infant PMA at the time of scan. BrainNetCNN is able to predict an infant's age with an average error of about 2 weeks, demonstrating that it can learn relevant topological features from the connectome data. Finally, we apply BrainNetCNN to the much more challenging task of predicting neurodevelopmental scores. We were able to achieve statistically significant correlations between predicted scores and true scores, with an average prediction error of around 11%. Furthermore, we show that BrainNetCNN achieves significantly higher correlation values than other competing prediction methods on this task.

Finally, we explore the high-level features learned by the CNN by visualizing which connections in the brain are most predictive of age, cognitive outcomes and motor outcomes. We find that edges important for predicting age are well distributed across the brain network. Also, we find that edges important for motor score prediction are connected to regions known to be responsible for motor function, and that other unique connections are important to predict cognitive scores.

1.1. Related works

The usefulness of representing the brain as a structural brain network for inference or prediction of injury and disease in adults has been widely recognized (Cuingnet et al., 2011; Ghanbari et al., 2014; Munsell et al., 2015; Zhu et al., 2014). However, only a very limited number of studies have applied these techniques to scans of infants. Ziv et al. (2013) examined if it were possible to predict general neurological health of infants at 6 months after birth using brain networks derived from DTIs. They employed a support vector machine (SVM) trained on high-level topological features. In our recent previous work, we used similar features to predict neuromotor development outcomes at 18 months from scans of preterm infants acquired shortly after birth (Brown et al., 2015).

While the application of ANNs to medical image analysis is well established for some clinical applications, its use for neurological applications has only lately become more popular (Yoo et al., 2014; Yang et al., 2014; Liu et al., 2014; Li et al., 2014; Brosch and Tam, 2013; Suk et al., 2014, 2015; Dvorak and Menze, 2015). For instance, ANNs have recently been used to segment brain lesions in multiple sclerosis patients (Yoo et al., 2014), segment brain tumors in multimodal MRI volumes (Dvorak and Menze, 2015), and classify different types of cerebellar ataxia (Yang et al., 2014). Various deep architectures have also recently been used to predict stages of Alzheimer's disease progression (Liu et al., 2014; Li et al., 2014; Suk et al., 2014, 2015). Similarly, Brosch and Tam (2013) employed deep belief networks to learn a manifold describing variation in a population of Alzheimer's patients. The networks in these studies, however, were all trained over standard grid-like MR images of brain structure as opposed to graph or network representations of brain structure.

Very few papers have applied ANNs to brain connectivity data. Munsell et al. (2015) used a fully connected deep auto-encoder to extract features from connectomes, but did not explicitly consider the structure of the brain network in the fully connected layers. Plis et al. (2014) explored the use of deep belief networks for a variety of classification tasks over functional MR (fMRI) and standard MR brain data, but collapsed the spatial dimensions of each input image to a single vector of voxels .

Recently, Bruna et al. (2013) and Henaff et al. (2015) showed that CNNs could be applied to data over a graphical domain (as opposed to grid-like data such as images). Their work followed work by Shuman et al. (2012) who showed how to generalize convolutions to graph structured domains . In those works the input signal was given over the nodes of the graph with a single set of edge weights fixed for all samples. In contrast, for the case of structural brain networks, the input signal is given as weights over the edges (reflecting, e.g., connectivity strength), implying a different set of edge weights for each sample. Thus, the techniques described by those works are not immediately applicable to brain network data and so, here, we introduce specialized filters for the task. There is, however, a relationship between convolutions over graphs as defined by Shuman et al. and the edge-to-edge filters we propose in this paper (detailed in Section 2.1.1).

Finally, some recent works have leveraged graph kernels to facilitate kernel based learning on connectome data (Jie et al., 2014; Dodero et al., 2015). In contrast to graph convolutions, graph kernels do not explicitly extract graph features but instead define an inner product between graphs. As far as we are aware, however, none of these works have applied graph kernels to infant structural brain networks nor incorporated them into a deep learning framework. We know of no other work, to date, that has adapted CNNs for edge-weighted networks and applied them to the human connectome.

2. Method

Here, we present our novel CNN layer types, designed specifically for network data input (Sections 2.1.1–2.1.3), the dataset used in this study (Section 2.2), the overall architecture of BrainNetCNN (Section 2.3), how we implemented BrainNetCNN (Section 2.4) and finally our evaluation metrics (Section 2.5).

2.1. CNN layers for network data

A DTI-derived brain network, $G = (A, \Omega)$, is a compact representation of the white matter connections in a patient's brain, where Ω is a set of nodes representing regions in the brain and A is a weighted adjacency matrix of edges, representing the connection strength between each pair of brain regions (typically defined as the number of white-matter tracts connecting the regions).

One way to apply ANNs to brain network data is to ignore the structure of the brain network and treat the input edge weights as a

[2] We refer to two types of networks in the text: The artificial neural networks (e.g., CNN) and the human brain network (connectomes). To avoid possible confusion between the two, we have endeavoured to make the distinction clear from the context and use of qualifiers such as 'brain' or 'convolutional'.

vector of features (Munsell et al., 2015). This approach, however, discards the topological relationships between edges that are intrinsic to the data. An alternative approach is to treat the adjacency matrix as an image and use established convolutional filters designed to capture the spatial 2D grid locality of images (e.g., a 5×5 filter). However, spatial locality between entries of the adjacency matrix does not directly correspond to topological locality in the brain network. For an entry located at $A_{i,j}$, only those elements within the i-th row and j-th column are topologically local and so the typical grid convolutional filters used for images are not appropriate here.

We consider these topological differences between images and brain networks as we adapt the CNN paradigm to brain network data. To leverage the structure found within the adjacency matrix, we introduce three new layer types: edge-to-edge layers, edge-to-node layers, and node-to-graph layers. Each layer type consists of one or more simple convolutional filters of a particular shape and performs a specific operation on the brain network. A BrainNetCNN layer contains one or more filters (of the same type). Each filter takes all feature maps from the previous layer as input and then outputs a distinct feature map for the next layer. Note that for all equations of the filter types below, we omit the activation function and the standard bias term for simplicity.

2.1.1. Edge-to-edge Layers

An *edge-to-edge* (E2E) layer is similar to a standard convolutional layer in a CNN over grid-like data in that it filters data locally. Whereas in grid-like data, filters may be defined in terms of spatial locality, the E2E filter is defined in terms of topological locality, combining the weights of edges that share nodes together.

Formally, let $G^{\ell,m} = (A^{\ell,m}; \Omega)$ represent the m-th feature map of a weighted brain network at the ℓ-th layer of the CNN, where Ω is the set of nodes corresponding to brain regions and $A^{\ell,m} \in \mathbb{R}^{|\Omega|\times|\Omega|}$ is an adjacency matrix containing the network edge weights. Each layer takes M^{ℓ} feature maps as input, and for this study we assume that $M^1 = 1$ (i.e., the input feature map to the whole CNN is just a single adjacency matrix describing one connectome). Since the number of nodes do not change between input and output, Ω stays constant and the output of an E2E layer is a filtered adjacency matrix defined as,

$$A^{\ell+1,n}_{i,j} = \sum_{m=1}^{M^{\ell}} \sum_{k=1}^{|\Omega|} r^{\ell,m,n}_{k} A^{\ell,m}_{i,k} + c^{\ell,m,n}_{k} A^{\ell,m}_{k,j} \tag{1}$$

where $[\mathbf{c}^{\ell,m,n}, \mathbf{r}^{\ell,m,n}] = \mathbf{w}^{\ell,m,n} \in \mathbb{R}^{2|\Omega|}$ such that $[\mathbf{w}^{\ell,1,n}, \ldots, \mathbf{w}^{\ell,M^{\ell},n}] \in \mathbb{R}^{2|\Omega|\times M^{\ell}}$ are the learned weights of the nth filter at layer ℓ. Thus, for each pair of input and output feature maps, (m, n), at layer ℓ, the E2E layer learns a single vector of weights, $\mathbf{w}^{\ell,m,n} = [w^{\ell,m,n}_1, \ldots, w^{\ell,m,n}_{2|\Omega|}]$. The set of all weights, $\{\mathbf{w}^{\ell,m,n} | m \in \{1, 2, \ldots, M^{\ell}\}\}$, that contribute to one output feature map, n, in one layer, ℓ, defines a single filter. The E2E filter is illustrated, for a single input feature map, in Fig. 2 and in entirety as a block diagram on the left side of Fig. 1.

Intuitively, for some edge (i,j) in an adjacency matrix encoded in some feature map, m, an E2E filter computes a weighted sum of edge weights over all edges connected either to node i or j, like a convolution. This implies that a single weight, $w^{\ell,m,n}_k$, is applied to all edges of a given node. This, however, does not imply that edges from a given node are all treated with equal importance. A single edge, (i,j), may be highly weighted if both $r^{\ell,m,n}_j$ and $c^{\ell,m,n}_i$ are large. Multiple distinct edges may then be weighted in this way via different network feature maps.

While this study focuses on the application of BrainNetCNN to undirected graph data, the E2E filter can, more generally, operate on directed graphs. For symmetric input, $A^{\ell,m}$, the output of an E2E filter $A^{\ell+1,n}$ may be asymmetric since, in general, it is not necessarily true that $r^{\ell,m,n}_i + c^{\ell,m,n}_j = r^{\ell,m,n}_j + c^{\ell,m,n}_i$. The filter may weight the input asymmetrically. For undirected graphs, however, this is simply the same as having two output feature maps (one upper triangular, one lower triangular) and so it is not necessary to enforce symmetric output. While it might be possible to design a filter similar to the E2E filter that operates only over the upper (or lower) triangular elements, it would very likely preclude the use of standard convolutional filters (i.e., the $\mathbf{r}$ and $\mathbf{c}$ components of the E2E filter). The proposed formulation of the E2E filter allows us to leverage these efficient convolutional filters and implement this filter easily in established CNN software packages (see below).

The E2E filter is similar to a $3 \times 3 \times M^{\ell}$ convolution filter over a stack of 2D grid data, in that, for each feature map, it combines the signal at some point with the signal from the direct neighbours, but does so with a cross shape filter instead of a box-shaped filter. Note that unlike a 2D image, the brain network has no topological boundaries and so the output of the layer can be the same size as the input without requiring any padding. Another difference, as noted above, is that whereas a convolution typically acts on a signal defined over the nodes of the grid (or over a general network as in the case of Shuman et al. (2012)), here our filter acts on a signal defined over the edges (i.e., edge weights).

The connection between convolutions over the edges of a graph versus convolutions over the nodes of a graph can be understood in terms of the concept of a line graph (Godsil and Royle, 2013): Let $\mathcal{L}(G)$ represent the line graph of G. Briefly, $\mathcal{L}(G)$ is a graph with one node corresponding to each edge in G and one edge corresponding to each pair of edges in G that are joined by a node. The nodes of $\mathcal{L}(G)$ adopt the signal over the edges of G (i.e., edge weights) and because there is no signal over the nodes of G, the topology of $\mathcal{L}(G)$ is consistent over the entire dataset. Thus, by constructing $\mathcal{L}(G)$, the definition of convolution over a graph by Shuman et al. (2012) becomes applicable to brain network data. It turns out that an E2E filter over G is equivalent to a filter over $\mathcal{L}(G)$ with a K-hop of 1, which, as demonstrated by Shuman et al. (2013), can be written as a generalized convolution. Note, however, that for typical sizes of Ω, in the order of dozens to hundreds (e.g., 90, as is the case here), $\mathcal{L}(G)$ contains $\frac{1}{2}|\Omega|^3 - \frac{1}{2}|\Omega|(|\Omega| - 1) = 360,\ 495$ edges versus only $\frac{1}{2}|\Omega|(|\Omega| - 1) = 4,\ 005$ for G, making operations over $\mathcal{L}(G)$ much more memory intensive. Thus, for efficiency and ease of interpretation, we chose to define the E2E filter in terms of G rather than $\mathcal{L}(G)$.

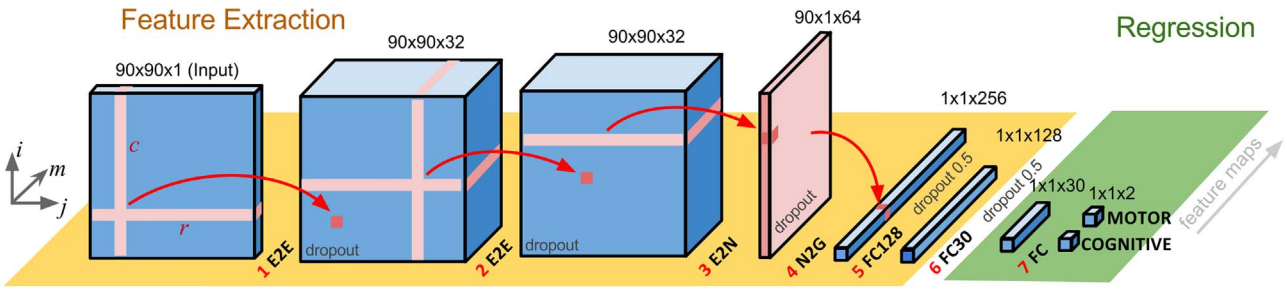

Fig. 1. Schematic representation of the BrainNetCNN architecture. Each block represents the input and/or output of the numbered filter layers. The 3rd dimension of each block (i.e., along vector m) represents the number of feature maps, M, at that stage. The brain network adjacency matrix (leftmost block) is first convolved with one or more (two in this case) E2E filters which weight edges of adjacent brain regions. The response is convolved with an E2N filter which assigns each brain region a weighted sum of its edges. The N2G assigns a single response based on all the weighted nodes. Finally, fully connected (FC) layers reduce the number of features down to two output score predictions.

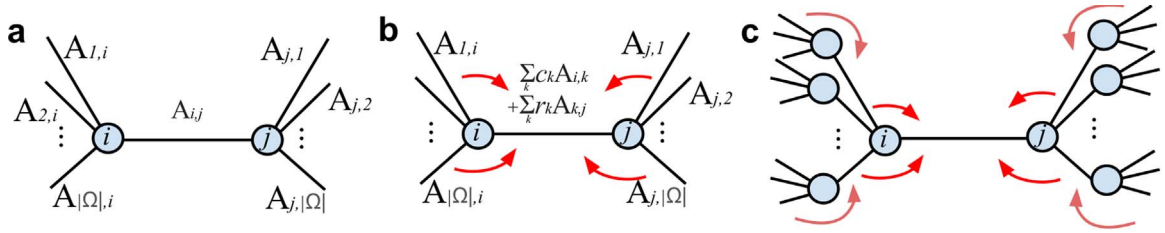

Fig. 2. An E2E filter at edge (i,j) shown, (a) before filtering, (b) after being applied once, and (c) after being applied twice. For simplicity, these examples assume only one input feature map and one output feature map. Accordingly, the feature map indices and layer indices are omitted.

2.1.2. Edge-to-node layer

An *edge-to-node* (E2N) filter takes an adjacency matrix, $A^{\ell,m}$, (representing a, possibly filtered, brain network) from each feature map as input and outputs a vector of size $|\Omega|$. Thus, the output of an E2N layer is defined as,

$$a_i^{\ell+1,n} = \sum_{m=1}^{M^\ell} \sum_{k=1}^{|\Omega|} r_k^{\ell,m,n} A_{i,k}^{\ell,m} + c_k^{\ell,m,n} A_{k,i}^{\ell,m}, \tag{2}$$

where, similar to an E2E layer, $[\mathbf{c}^{\ell,m,n}, \mathbf{r}^{\ell,m,n}] = \mathbf{w}^{\ell,m,n} \in \mathbb{R}^{2|\Omega|}$ such that $[\mathbf{w}^{\ell,1,n}, \ldots, \mathbf{w}^{\ell,M^\ell,n}] \in \mathbb{R}^{2|\Omega| \times M^\ell}$ are the learned weights of the nth filter at layer ℓ. However, the n-th output feature map, $\mathbf{a}^{\ell+1,n}$, of an E2N layer is a vector in $R^{|\Omega|\times 1}$, in contrast to an E2E layer whose output feature map is in $R^{|\Omega|\times|\Omega|}$.

An E2N filter is equivalent to convolving the adjacency matrix with a spatial 1D convolutional row filter and adding the result to the transpose of the output from a 1D convolutional column filter. This operation can be interpreted as computing a single output value for each node, i, by taking a weighted combination of the incoming and outgoing weights of each edge connected to i. Note that if we assume the input to the E2N filter is a symmetric matrix, we can drop either the term containing the row weights, $\mathbf{r}^{\ell,m,n}$, or the term containing the column weights, $\mathbf{c}^{\ell,m,n}$, since the incoming and outgoing weights on each edge will be equal. In all experiments in this paper, we used E2N filters with only the $|\Omega|$ row weights in $\mathbf{r}$ because we did not empirically find any clear advantage in learning separate weights for both incoming and outgoing edges when training over symmetric connectome data.

Similar to the E2E layer, the E2N layer does not necessarily discard information about distinct edges with particular importance: If weights $r_i^{\ell,m,n}$, $c_i^{\ell,m,n}$, $r_j^{\ell,m,n}$ and $c_j^{\ell,m,n}$ are all relatively large, then edge (i,j) will be weighted especially strongly and through multiple feature maps, many edges may be highly weighted in this way.

2.1.3. Node-to-graph layer

Finally, similar to the E2N layer, a node-to-graph (N2G) layer reduces the dimensionality of the input, in this case by taking a weighted combination of nodes to output a single scalar,

$$a^{\ell+1,n} = \sum_{m=1}^{M^\ell} \sum_{i=1}^{|\Omega|} w_i^{\ell,m,n} a_i^{\ell,m}, \tag{3}$$

per output feature map, n. The N2G filter, also a 1D spatial convolution, is applied after an E2N filter and reduces the spatial dimensions of the original input to single scalar per feature map. In the context of being applied after an E2N filter, which summarizes the responses of neighbouring edges into a set of node responses, the N2G filter can be interpreted as getting a single response from all the nodes in the graph.

2.2. Preterm data

The data for this study is from a cohort of infants born very preterm, between 24 and 32 weeks PMA, and imaged at BC Children's Hospital in Vancouver, Canada. The use of this data for this study was approved by the University of British Columbia Clinical Research Ethics Board. As detailed in Booth et al. (2016), after excluding images for poor scan quality (in short, first by visual inspection of the DTIs and then by examining tractography results for serious artefacts and directional bias), scans of 115 infants were used. Roughly half of the infants were scanned twice (shortly after birth and then again at about 40 weeks PMA), for a total of 168 scans. Full-brain streamline tractography was performed on each DTI to recover the neuronal connections in each brain. Using a neonatal atlas of $|\Omega| = 90$ anatomical regions from the University of North Carolina (UNC) School of Medicine at Chapel Hill (Shi et al., 2011), a weighted, undirected network was constructed from each scan by counting the number of tracts connecting each pair of anatomical regions. Each network is represented as a 90×90 symmetric adjacency matrix with zeros along the diagonal and is scaled to $[0, 1]$. At 18 months of age, adjusted for prematurity, the cognitive and neuromotor function of each subject was assessed using the Bayley Scales of Infant and Toddler Development (Bayley-III) (Bayley, 2006). Cognitive and motor scores from this test are normalized to a population mean of 100 with standard deviation of 15. See Brown et al. (2015) for further details about assessment protocol, scanning protocol and connectome construction.

Given the small data set (DTI of preterm infants is not standard procedure in clinical practice) and the imbalance (low numbers of preterm infants with high and low neurodevelopmental outcomes) we adopted the synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002) to balance and augment each training set. Training samples were binned by score (5 bins) and then SMOTE was run, repeatedly, to generate a synthetic sample from the bin with the fewest total number of real and synthetic samples, until the training set was augmented by a factor of 256. Note that in our previous work, we showed that the proposed LSI method outperformed SMOTE for improving prediction accuracy (Brown et al., 2015). While LSI worked well in that context, we were performing 2-class classification rather than regression. LSI is not applicable here because it augments data in individual classes, and in this paper we are performing regression over a single training set.

2.3. BrainNetCNN architecture

We base the architecture of our BrainNetCNN (for connectomes) on a common CNN (for images) where the first section of the network is composed of convolutional layers and the last section is composed of fully connected (FC) layers (e.g., Simonyan and Zisserman, 2015). Fig. 1 is a block diagram of a representative BrainNetCNN architecture with at least one layer of each of the proposed filter types.

The input to a BrainNetCNN model is a brain network, G^0, represented as a 90×90 adjacency matrix. The output layer of the network has two nodes where each node predicts a different neurodevelopmental outcome score (motor and cognitive). The second to last layer in the network of size $1 \times 1 \times 30$ can be interpreted as a set of high-level features learned by the previous layers. We selected a size of 30 features in order to directly compare the features learned by BrainNetCNN to the 30 network measure features used by Brown et al. (2015).

Since E2E layers operate on a whole adjacency matrix (per feature map), they can only be applied before E2N and N2G, which reduce the

input dimensionality (to a vector or a scalar per feature map). However, since E2E layers do not alter the input dimensionality, many E2E layers can be stacked (with the trade-off of an increased number of parameters to learn). An E2N layer reduces the 90 × 90 matrix to a single matrix of 90 × 1 elements and thus must be applied before an N2G layer. The N2G layer reduces the input dimensionality down to a single feature (per feature map) and thus cannot be applied before the E2E or E2N layers.

In the experiments below (Section 3) we test a variety of configurations of BrainNetCNN. Each configuration of BrainNetCNN can be understood as a CNN with a subset of the layers shown in Fig. 1. The basic configuration (E2Enet) contains one of each type of proposed layers along with 3 fully connected layers (i.e., layers 1, 3, 4, 5, 6 and 7 in Fig. 1). We also tested configurations with fewer layers: One model with the E2E layer removed (E2Nnet), and two more models similar to E2Enet and E2Nnet but with two of the fully connected layers removed (E2Enet-sml and E2Nnet-sml, respectively). Finding good results with these FC layers removed, we tested a model with the same layers as E2Enet-sml but with an additional E2E layer (2E2Enet-sml).

We compare our results from these BrainNetCNN configurations to one and two layer fully connected neural networks (FC30net and FC90net, respectively), which don't contain any of the proposed convolutional layers. The input to the FC networks is a 1×4005 vector consisting of the upper triangular values of the symmetric connectome matrix. FC90net is similar to layers 5, 6 and 7 in Fig. 1 but with only 90 responses between layers 5 and 6 to make the number of learnable parameters approximately equal to that in E2Nnet-sml and E2Enet-sml.

Generally, the number of output feature maps from each layer, M^{ℓ}, is independent of other network parameters and can be set freely. In the BrainNetCNN architecture, we increased the number of feature maps with each layer to compensate for the reductions along the other dimensions (i.e., dimensions i and j in Fig. 1); a common strategy for CNNs (e.g., Simonyan and Zisserman, 2015). Precisely, E2Nnet-sml has an E2N layer with 130 1×90 filters (layer 3 is increased from 64 to 130 to match the number of parameters with the other models) producing feature maps of size $1 \times 90 \times 130$. This is followed by an N2G layer with feature maps of size $1 \times 1 \times 30$ (layer 4) and a fully connected layer with an output of size 2 (layer 7). E2Enet-sml is constructed from layers 1, 3, 4, 7 (Fig. 1), with an E2E layer composed of 321×90 and 32 90×1 filters (layer 1) producing feature maps of size $90 \times 90 \times 32$. This is followed by an E2N layer with 64 $1 \times 90 \times 32$ filters (layer 3) producing feature maps of size $1 \times 90 \times 64$, an N2G layer with feature maps of size $1 \times 1 \times 30$ (layer 4), and a fully connected layer with an output of size 2 (layer 7) .

Every layer in our network uses very leaky rectified linear units as an activation function, where a leaky value of $x/3$ is assigned if $f(x) < 0$, as done by Graham (2014). For training, we employed dropout using a rate of 0.5 after the N2G layer and the FC layer of 128 units as shown in Fig. 1 (dropout was found to slightly improve correlation by ≈0.01 for the fully connected model). We used momentum of 0.9, a mini-batch of size 14, a weight decay of 0.0005, and a learning rate of 0.01. Mini-batch sizes, weight decay and learning rates were set to values that performed well over the fully connected model (see Section 3.2). All models minimized the training loss, which is defined as the Euclidean distance between the predicted and real outcomes plus a weighted $L2$ regularization term over the network parameters.

The ideal number of training iterations for a given model depends on the model architecture and on the training parameters. Thus, to minimize overfitting to the training data, and to ensure a fair comparison across all model types (both proposed and competing), we trained each model for a variable number of iterations, from 10 K to 100 K (in 10 K increments) and selected the model corresponding to the number of iterations that yielded the least overfitting (i.e., best performance on the test data).

2.4. Implementation

We implemented our BrainNetCNN using the popular deep learning framework, Caffe Jia et al., 2014. While the E2N and N2G filters were straightforward to implement using 1D filters, the E2E filter required a convolution of two 1D filters, $\mathbf{c} \in \mathbb{R}^{|\Omega| \times 1}$ and $\mathbf{r} \in \mathbb{R}^{1 \times |\Omega|}$, with the adjacency matrix, producing responses of dimensions $\mathbb{R}^{1 \times |\Omega|}$ and $\mathbb{R}^{|\Omega| \times 1}$, respectively. These response vectors are each replicated $|\Omega|$ times to produce two $\mathbb{R}^{|\Omega| \times |\Omega|}$ matrices, which are summed element-wise yielding a single matrix equivalent to Eq. (1).

2.5. Evaluation metrics

In addition to reporting mean absolute error (MAE) and the standard deviation of absolute error (SDAE) between the predicted and the true scores, we report the Pearson correlation coefficients between the predicted and the true scores, and the corresponding p-values. As our dataset contains many scores close to the mean value, MAE may be disproportionately low for regressors that frequently predict nearer to the mean score of the training data, even if they underfit the data. The Pearson correlation coefficient, however, measures the linear dependence between predicted and true scores and so is less affected by the distribution of the inputs. MAE is still important to report, however, since Pearson's correlation does not expose if a regressor is biased towards frequently predicting too high or too low. Thus, the measures are complementary.

3. Experiments

3.1. Simulating injury connectomes for phantom experiments

Before testing BrainNetCNN on real brain networks, we assessed its ability to learn and discriminate between differing network topologies using sets of synthetically generated networks. We first examined the performance of BrainNetCNN on data with increasing levels of noise and then compared BrainNetCNN to a fully connected neural network with the same number of model parameters. To simulate realistic synthetic examples, each example is based on the mean connectome, X_{μ} (Fig. 3-left), of our preterm infant data, perturbed by a simulated focal brain injury using a local signature pattern S. The symmetric matrix $S \in \mathbb{R}^{|\Omega| \times |\Omega|}$ has non-zero elements uniformly selected between [0, 0.1] (i.e., up to 10% of the values of X_{μ}) along the same row and column index. Thus, the simulated injury is to all connections (with varying intensity) emanating from a single brain region. We created two focal injury signatures, S^1 and S^2, with two corresponding injury regions. These two regions were chosen as the two rows in X_{μ} with the highest median responses in order to simulate injury to important regions (i.e., hubs) of the brain (Fig. 3-right). Mathematically, the i-th synthetic connectome, X_i, is formed as,

$$X_i = \frac{X_{\mu}}{(\mathbf{1} + \alpha_i S^1)(\mathbf{1} + \beta_i S^2)} + \gamma N_i \tag{4}$$

where $\mathbf{1}$ is a matrix composed of all ones; $N_i \in \mathbb{R}^{|\Omega| \times |\Omega|}$ is composed of random values simulating noise weighted by γ; and, α_i and β_i are scalar injury parameters that weight their respective signature matrices. α_i and β_i range between 50 and 140 as these are typical neurodevelopmental outcome ranges in our dataset. All operations are done element-wise and the resulting synthetic connectome X_i (Fig. 3-center) forms our observed example.

3.1.1. Predicting injury parameters over varying noise

We first tested our model's ability to predict the injury parameters (i.e., α_i and β_i) given the corresponding X_i under different level of noise, γ. The model was trained using 1000 synthetic examples and test over another 1000 examples. We chose 1000 training samples as it

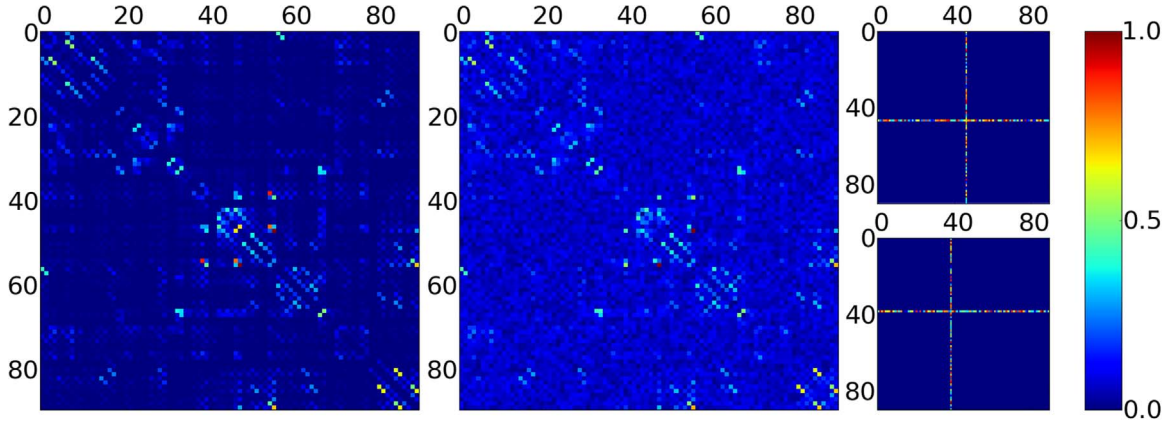

Fig. 3. (Left) The averaged connectome. An example synthetic connectome (*center*) used in our focal injury phantom experiments after introducing noise and the two signatures at the 47th and 39th regions (right).

Table 1
Synthetic experiments using E2Enet-sml to predict injury parameters α and β under different levels of noise measured by the peak-signal-to-peak-noise-ratio (PSPNR = $1/\gamma$). As expected, as the noise levels decrease, the Pearson correlation r increases (r_α indicates correlation with the α parameter), and the mean absolute error (MAE) and the standard deviation of the absolute error (SDAE) decrease.

PSPNR	r_α	MAE_α	SDAE_α	r_β	MAE_β	SDAE_β
4 (12 dB)	0.554	19.949	14.497	0.588	18.356	13.967
8 (18 dB)	0.873	9.732	7.870	0.873	9.980	8.259
16 (24 dB)	0.965	6.458	5.026	0.969	5.008	4.195
∞	1.000	1.071	0.682	0.999	1.088	0.879

represents a realistic best-case scenario for a large dataset of DTI scans. As shown in Table 1, under moderate noise, our BrainNetCNN model (E2Enet-sml) accurately predicts α and β, indicating an ability to recognize multiple subtle, synthetically induced connectome perturbations.

3.1.2. Predicting focal injury parameters with different models

We also used the phantom data to assess the difference in predictive ability on a small training set, between a fully connected model (FC90net) and two models based on our proposed BrainNetCNN layers (E2Nnet-sml, E2Enet-sml), each with a similar number of model parameters.

To more closely approximate our real dataset, we used 112 synthetic samples to train, 56 synthetic samples to test and used relatively high, fixed PSNR of 8 (or 18 dB, where PSPNR = $1/\gamma$). The results are reported in Table 2.

The E2Enet-sml outperformed the FC90net model achieving an average increase in mean correlation of 15.54% and an average decrease in MAE of 29.17% over both parameters, and slightly outperformed E2Nnet-sml across all measures. The E2Nnet-sml also outperformed FC90net across all measures. As these models all have nearly the same number of parameters to learn, and E2Nnet-sml has the same number of non-linear layers as the FC90net model, these tests indicate that the BrainNetCNN convolutional filters contribute greatly to the improvements in prediction accuracy on this realistic phantom.

Table 2
Comparison of a fully connected model (*top row*) with two proposed BrainNetCNN models (*bottom rows*), all with similar numbers of parameters on phantom data.

Model	r_α	MAE_α	SDAE_α	r_β	MAE_β	SDAE_β
FC90net	0.648	20.583	11.609	0.688	20.080	11.513
E2Nnet-sml	0.736	16.380	10.977	0.752	16.492	9.834
E2Enet-sml	**0.812**	**13.760**	**9.494**	**0.772**	**15.021**	**9.761**

3.1.3. Predicting diffuse injury parameters with different models

It is thought that poorer neurodevelopmental outcomes in many preterm infants, especially low cognitive scores, may be caused by diffuse white matter injuries rather than focal lesions (Back and Miller, 2014). Thus, in addition to simulating focal injuries, we also test our method on a phantom dataset with simulated diffuse injuries, spread across the whole brain. The diffuse injury synthetic connectomes are created using the same method described above, in Section 3.1, except that the focal injury pattern matrices, S^1 and S^2 are replaced with diffuse injury pattern matrices, D^1 and D^2. Diffuse injury patterns are simulated by selecting a random injury weight (again in [0, 0.1]) for each region. Given a 90×1 vector, d^k, of injury weights, a symmetric diffuse injury pattern is computed as $D^k_{i,j} = \frac{1}{2}(d_i^k + d_j^k)$. Examples of diffuse injury patterns and a diffuse injury synthetic connectome are shown in Fig. 4. While the same level of noise (PSPNR of 8) was applied to this dataset as for the focal injury phantoms, the broader injury pattern produces a weaker overall connectivity signal, causing the noise to appear more pronounced.

As with the experiment on focal injury phantoms, here we test the ability of FC90net, E2Nnet-sml and E2Enet-sml models to predict two independent injury patterns. On this more challenging phantom data, the BrainNetCNN models again outperform the FC model in terms of both MAE and correlation (Table 3). Here, however, the E2Nnet-sml model slightly outperforms the E2Enet-sml.

3.2. Infant age and neurodevelopmental outcome prediction

To test the BrainNetCNN on the preterm infant data, we performed 3-fold cross-validation. The data was split randomly into three folds of 56 scans with the constraint that scans of the same subject were in the same fold. We chose three folds because, despite giving a larger training set size, more folds would require an increased number of (deep) models to be trained. In each round, two folds were selected as a training set, augmented (as described in Section 2.2) and then used to train a model. As ANNs can find different local minima and thus produce different solutions, for each test involving an ANN, we trained each model with five different random initializations and averaged the predicted scores (Cireşan et al., 2012, 2013; Simonyan and Zisserman, 2015).

3.2.1. Model sensitivity to initialization and number of iterations

As was mentioned above, for a fair comparison, the reported correlation values (i.e., capturing the prediction accuracy) for each architecture were the best achieved for that architecture across different numbers of training iterations. Fig. 5 compares the correlation values across increasing numbers of training iterations (from 10 K to 100 K) for both FC90net and E2Enet-sml architectures. For each

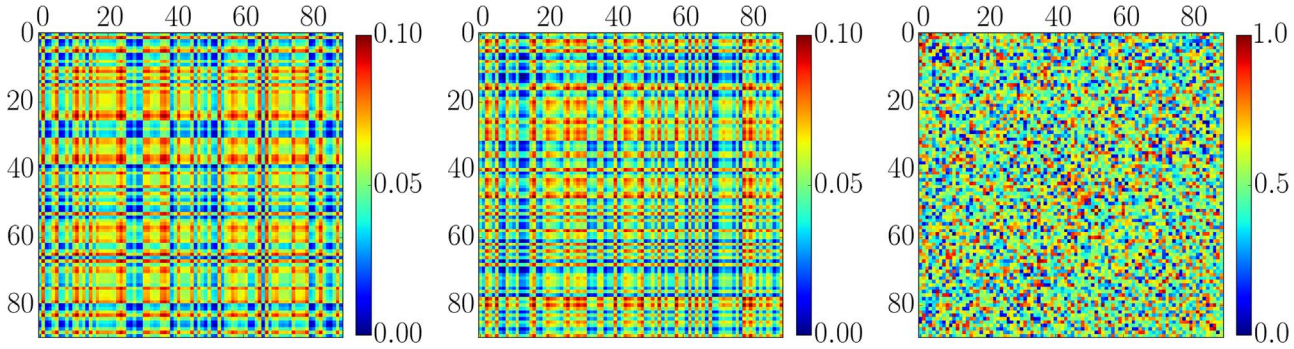

Fig. 4. (Left and center) Sample diffuse whole brain injury patterns. (Right) Sample diffuse injury synthetic connectome with two diffuse injury patterns and noise applied.

Table 3
Comparison of a fully connected model (*top row*) with two proposed BrainNetCNN models (*bottom rows*), all with similar numbers of parameters on diffuse injury phantom data.

Model	r_α	MAE_α	$SDAE_\alpha$	r_β	MAE_β	$SDAE_\beta$
FC90net	0.129	22.614	**11.946**	0.217	20.796	13.838
E2Nnet-sml	**0.398**	**19.570**	12.476	**0.326**	**19.724**	**13.223**
E2Enet-sml	0.386	19.712	12.483	0.315	19.938	13.531

type of architecture, predicting each neurodevelopmental outcome type, the correlation values increase rapidly and then roughly plateau after about 30 K training iterations. So, while we chose the best number of iterations for each method to be fair to each type of architecture, we observe that the correlation value is fairly insensitive to this parameter. Furthermore, Fig. 5 validates that 100 K is a good upper limit for number of training iterations, as no model appears like it would greatly improve given more training. In the case of cognitive score prediction using the E2Enet-sml model, the correlation values appear to slightly decrease after 80 K iterations, potentially indicating that the model is beginning to over-fit to the training data past this point. Results for both E2Enet-sml and FC90net models are reported in Table 4 at the 60 K mark since it is the peak of each of their combined correlations. Note that the mean correlation value slightly differs from what is reported in Table 4 since Table 4 averages the predictions together over the five models before taking the correlation (instead of computing the mean of the correlations for each model as is displayed in Fig. 5).

Fig. 5 shows the mean and the standard deviation of the correlation values across the predictions of the five different randomly initialized models. Furthermore, the standard deviation decreases with the number of iterations, meaning that the different independently initialized models converge to similar performance after training.

3.2.2. Age prediction

Before applying BrainNetCNN to the very difficult task of predicting neurodevelopmental outcomes, we first trained it to predict infant PMA at the time of scan. We performed this test to establish an upper-bound on the predictive performance of BrainNetCNN, as there are perhaps fewer complicating factors in predicting age compared to predicting neurodevelopmental outcomes (which we discuss in Section 4). Using E2Enet-sml, we were able to accurately predict PMA, with an MAE of 2.17 weeks (or 11.1% of the total age range) and an SDAE of 1.59 weeks. The correlation between predicted and ground-truth age was 0.864. While the purpose of this test is only to show the ability of our model to learn some clinical parameters given the connectome data, for completeness, we also tested the FC90net model and the E2Nnet-sml model. On this baseline task, the FC90net model performed slightly worse than E2Enet-sml, achieving an MAE of 2.29 weeks, SDAE of 1.65 weeks and a correlation of 0.858. Similarly, the E2Nnet-sml model slightly underperformed E2Enet-sml, achieving an MAE of 2.377, SDAE of 1.72 and a correlation of 0.843.

We found that absolute error of age prediction (using the E2Enet-sml model) was correlated with PMA (r=0.224), implying that age predictions were more accurate for younger infants. In Section 3.2.4, we visualize and discuss which edges and regions of the infant connectomes BrainNetCNN determined to be most important for predicting age.

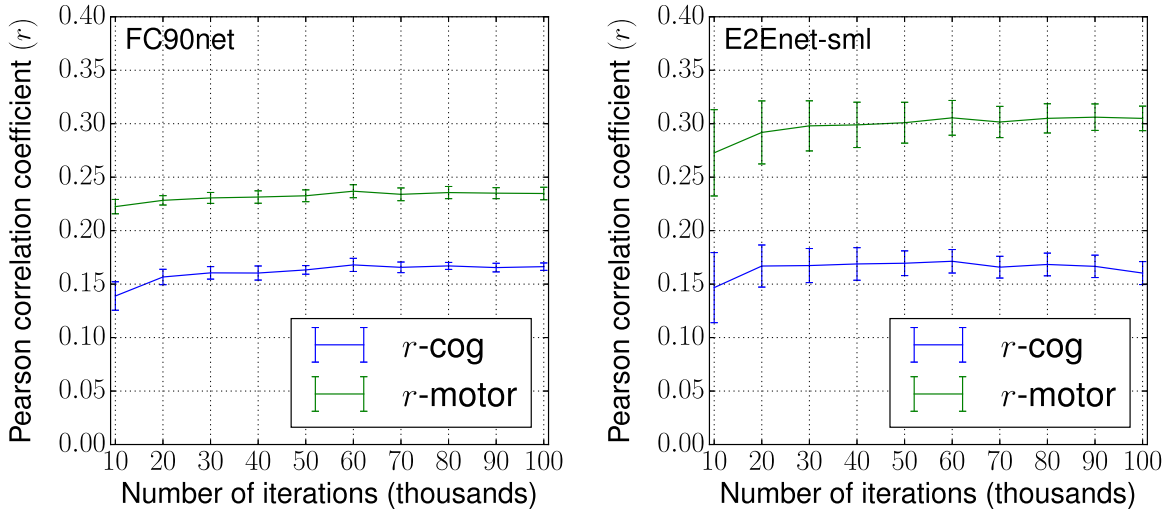

Fig. 5. The effect of the number of training iterations on correlations between predicted and ground truth outcome scores (on test data) for FC90net (left) and E2Enet-sml (right) architectures. The standard deviation of each of the five randomly initialized models is shown at each 10K iterations as vertical error bars.

Table 4
Correlation, r, corresponding p-values, MAE and standard deviation of absolute error (SDAE) between true and predicted Bayley-III motor and cognitive scores. Results for different configurations of BrainNetCNN (with different subsets of the layers shown in Fig. 1) and for competing models trained on different features. Our proposed, full BrainNetCNN model with one E2E layer for motor score and two E2E layers for cognitive outperform all other methods in terms of correlation.

	Model	Layers	Motor				Cognitive			
			r	p	MAE	SDAE	r	p	MAE	SDAE
competing	Clinical	-	0.106	0.170	16.139	13.737	0.086	0.271	15.339	12.053
	Network	–	0.227	0.003	13.345	9.761	0.143	0.064	13.564	9.722
	PCA30	–	0.181	0.019	12.186	8.259	0.069	0.374	11.682	8.809
	Raw Edges	7	0.176	0.023	27.399	27.273	0.063	0.420	27.502	26.529
	FC30net	6,7	0.231	0.003	10.915	8.075	0.158	0.041	10.583	8.572
	FC90net	5,6,7	0.237	0.002	11.142	7.986	0.169	0.029	10.545	8.631
proposed	E2Nnet	3,4,5,6,7	0.271	0.0004	11.095	7.797	0.154	0.046	10.845	8.902
	E2Enet	1,3,4,5,6,7	0.281	0.0002	11.506	7.833	0.182	0.018	11.132	8.964
	E2Nnet-sml	3,4,7	0.263	0.0006	**10.640**	8.075	0.162	0.0355	**10.493**	8.459
	E2Enet-sml	1,3,4,7	**0.310**	**<0.0001**	10.761	7.734	0.174	0.0239	11.231	**8.424**
	2E2Enet-sml	1,2,3,4,7	0.290	0.0001	11.153	**7.686**	**0.188**	**0.0148**	11.077	8.574

3.2.3. Neurodevelopmental outcome prediction

We explored the more challenging outcome prediction task using different configurations of BrainNetCNN and competing methods (see Table 4). We compared the ANN models (i.e., FC and BrainNetCNN models) to linear regressors trained on features from (**i**) the raw edge weights (Raw Edges), (**ii**) 30 principal components of the edge features using PCA (PCA30), (**iii**) high-level network features (Network), as used by Brown et al. (2015), and (**iv**) 6 clinically relevant metadata features (Clinical) including age at birth, age at scan, gender and ratings of brain white matter injury (Miller et al., 2005), ventriculomegaly (Cardoza et al., 1988) and intraventricular hemorrhaging (Papile et al., 1978) that are used by clinicians to assess risks to preterm infants neurodevelopmental outcomes. As with the size of the last layer of BrainNetCNN, we chose 30 PCA features in order to provide the most direct comparison to Brown et al. (2015).

Table 4 reports MAE, SDAE, correlations and correlation p-values between ground-truth and predicted scores. The statistical significance ($p < 0.05$), reports the very small likelihood that the positive correlation obtained is coincidental.

In terms of MAE, many models performed similarly well over motor and cognitive outcomes. PCA30 performed nearly as well as the neural network based models which all achieved average absolute errors of <11% (based on a range of scores between 50 and 155). This result, alone, appears to suggest that the simplest models can perform with similar accuracy to more complex models. However, the correlation results contradict this and suggest that the PCA model has actually underfit the data, predicting a similar output for every input, resulting in comparatively low r values.

Different configurations of our BrainNetCNN produce the highest prediction correlation values for both motor and cognitive scores. Despite having the same number of trainable model parameters as FC90net (and significantly less parameters than E2Nnet and E2Enet) the E2Enet-sml model results in the highest motor correlation. Similarly for cognitive scores, a model with an additional E2E layer, 2E2net-sml, attains the highest prediction correlation. The E2Nnet-sml yields the lowest MAE for both motor and cognitive scores.

Paired t-tests were used to check the significance of the improvement of the BrainNetCNN models over FC90net, the next best model. To do this, 1000 random subsets of 56 instances (i.e., the size of each fold) were selected. For each model, the correlation between scores predicted by that model and the ground truth scores were computed within each subset. (Note that for all models, the distributions of correlation values across the 1000 subsets were found to be normal using Kolmogorov–Smirnov tests.) Each paired t-test was performed between a pair of models with the null hypothesis that the mean of the distribution of correlation values were equal. The paired t-tests showed that all models with an E2E layer performed significantly better, on average, than the FC90net model on both prediction tasks with $p < 0.05$ except for the E2Enet-sml model which did not perform significantly better at predicting cognitive scores. For the 2E2Enet-sml model, correlations improved over FC90net an average of 8.44% for motor scores and 10.4% for cognitive scores.

To ensure that BrainNetCNN was not consistently predicting too high or too low (i.e., prediction bias), a t-test on the prediction errors of E2Enet with respect to each score type was performed. The mean difference between predicted and ground truth values for cognitive and motor scores were not found to be statistically significantly different from zero (p-values of 0.6817 and 0.9731 respectively), meaning that our model was unbiased.

3.2.4. Maps of predictive edges

In order to uncover which connections were learned by BrainNetCNN to be predictive of age, cognitive outcome and motor outcome, we used the method of Simonyan et al. (2013), which computes the partial derivatives of the outputs of an ANN with respect to the input features. For each outcome y_s (i.e., either age or motor or cognitive scores), Simonyan et al.'s method computes $\frac{\partial y_s}{\partial A_{i,j}^{1,m}}$ for every input edge (i,j). In Fig. 6, the partial derivatives of motor and cognitive scores, averaged over the entire dataset, are plotted for all connectome edges, both spatially on line segments connecting centroids of the UNC atlas regions and schematically in Circos plots (Krzywinski et al., 2009). A complete list of region names corresponding to the region codes used in the Circos plots can be found in the Appendix of the recent paper by Brown et al. (2014).

While many of the partial derivatives are positive (red) indicating connections that, when strong (i.e., high tract count), contribute to better outcomes there are also many negative partial derivatives (blue). We see that many brain connections (edges) from the right middle frontal gyrus (MFG) are selected as being predictive of positive outcomes for both motor and cognitive scores. The left precuneus (PCUN), fusiform gyrus (FFG), superior frontal gyrus (SFGdor) and right lingual gyrus (LING) also appear to be prominent hubs of important connections for both scores. For motor scores, the connection between the two superior frontal gyri appears to be of particular importance. In contrast, the connection between the left FFG and right LING is highlighted as being relatively more important for cognitive scores than for motor scores. However, there is considerable overlap between the two sets of edges.

Compared to the sets of edges found to be important for predicting neurodevelopmental outcomes, those found to be important for predicting age are much more widely distributed across the brain network (Fig. 6). Only the connection between the right LING and the right FFG

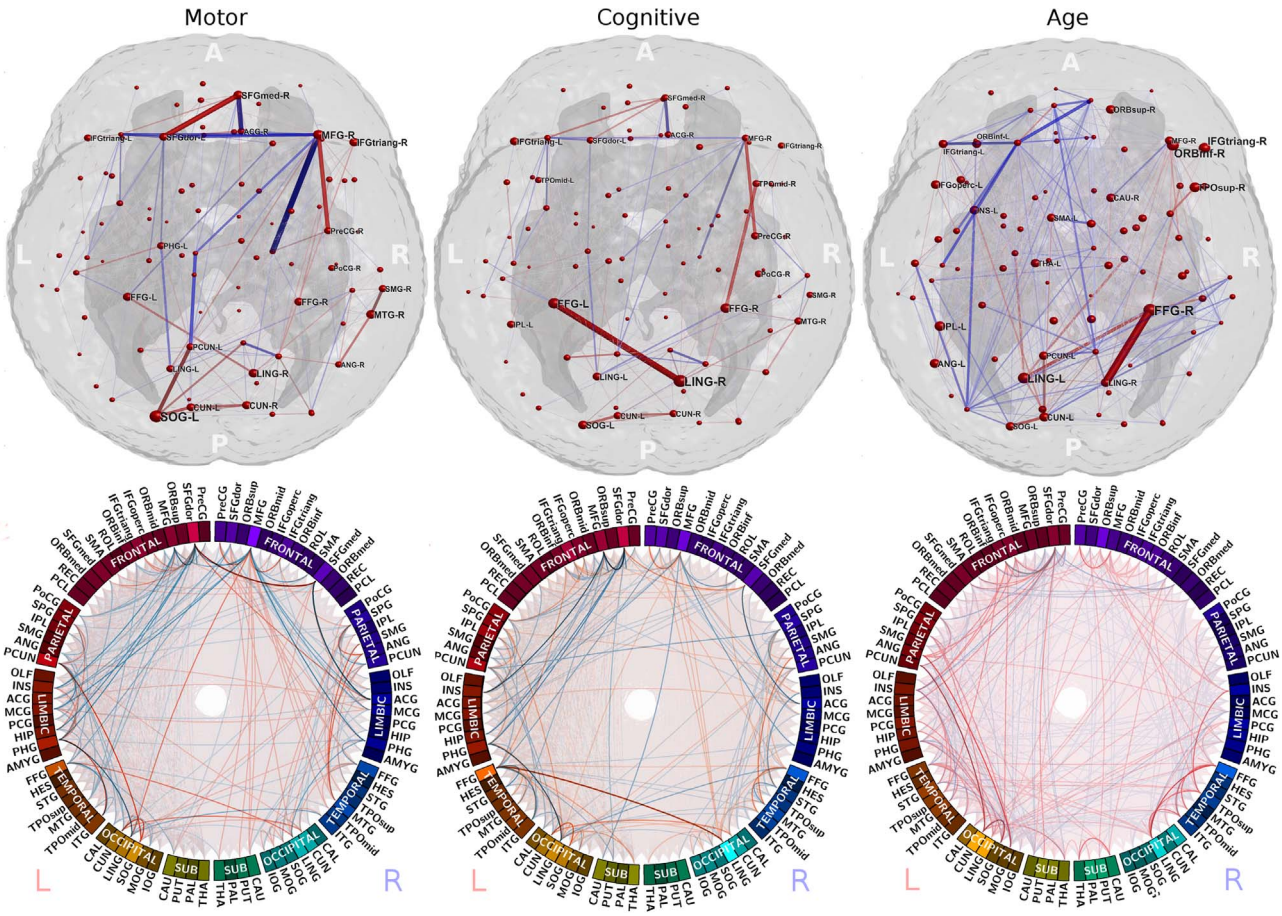

Fig. 6. Connections learned by BrainNetCNN to be most predictive of outcomes and ages. *Top row*: Edges with positive (red) and negative (blue) partial derivatives with respect to motor outcomes (left), cognitive outcomes (middle) and ages (right). Edge thickness and opacity represent the magnitude of each partial derivative. Very small magnitudes (< 0.001) were omitted for clarity. Node sizes represent the sum of partial derivative magnitudes of all neighbouring edges with positive derivatives. *Bottom row*: The same partial derivatives mapped on to Circos ideograms. Brightness of the color of the regions in each ring denotes the sum of positive partial derivative magnitudes.

appears to stand out as being a particularly strong predictor. We discuss possible anatomical reasons for these observations below.

4. Discussion

Broadly, the proposed BrainNetCNN performed well, predicting motor and cognitive scores with the highest correlations to the ground truth scores. Furthermore, it was found that, with respect to most accuracy measures, our convolution based models (e.g., E2Enet-sml, 2E2Enet-sml) were able to outperform other models without relying on the large fully connected layers. This increased accuracy was found for both real connectome data and carefully controlled phantom data. These results validate that our novel E2E, E2N and N2G filters, are able to learn important structures for prediction with a relatively small number of parameters. As well, it suggests that an alternative to learning larger models with more layers is to employ convolutional layers that leverage the topology of the input data.

It was also found that for prediction of cognitive scores, it was helpful to stack E2E layers as seen by the comparatively high correlation value for 2E2Enet-sml. This stacking of E2E layers enables learning of complex structural patterns while requiring the optimization of relatively few additional parameters.

When BrainNetCNN was used to predict age, it was found that prediction was more accurate for the younger infants. One factor that likely contributed to this result is that there are more scans of younger infants in our data set (60% of scans are below the age range mid point), which provided more training data for these cases. If true, it suggests that larger training sets could further improve prediction results.

Despite the discrepancy in prediction error between younger and older infants, our E2Enet-sml model was able to predict PMA with high accuracy. However, when predicted with the same model, the correlation values for neurodevelopmental outcome scores were relatively low (e.g., 0.310 for motor scores versus 0.866 for age). While statistically significant, these values for prediction of outcome scores entail only weak to moderate correlations. Nevertheless, note that the correlation values and relative improvement of BrainNetCNN over FC models were only slightly lower for this real data than for the simulated diffuse injury phantoms. Fig. 4 (right), especially as compared to Fig. 3 (center), gives a sense of the level of difficulty of the prediction problem to result in correlation values in this range.

A number of factors contribute to the increased difficulty of predicting outcome scores compared to predicting age. Probably the most significant factor is the ≈18 months of intervening time between scan and Bayley-III assessment. This task of predicting neurodevelopmental outcomes of infants 18 months into future is made more difficult by the fact that, shortly after birth, infants are developing very rapidly and environmental and genetic factors will affect the course of this development. The infant brain may also be impacted by preterm birth and postnatal illness through mechanisms that do not alter DTI metrics of diffusivity. Furthermore, the amount of available data for training remains relatively small compared to the dimensionality of the input networks, especially for the minority of cases with very high and low outcome scores. While data augmentation can be use to expand and balance a dataset, it is not a substitute for real data.

In all of the experiments on real connectome data, we trained each

ANN model on motor and cognitive outcomes jointly. This was done because the scores are strongly correlated (r=0.68 in our dataset) and we expected that prediction of two outcomes would help regularize our underdetermined models. We did explore training motor and cognitive outcomes separately but found little difference in our metrics compared to joint training. Given that joint training requires only a single model trained for both tasks, significantly reducing the computational burden and training times, it was our adopted choice for all experiments. While its possible that low cognitive outcomes and low motor outcomes do not share a common aetiology, the 30 high-level features of the last layer of the proposed models provides these models with high flexibility to identify injury patterns of different types.

In comparison to our BrainNetCNN learned features, the network measure features used in our earlier work (Brown et al., 2015) performed poorly. This was somewhat surprising, as they were shown to perform well on the similar task of motor classification (Brown et al., 2015). However, in that work, these network measure features were combined with several meta-data features, including information about age, gender and brain white matter injury grade, then dimensionality reduced using PCA before performing prediction. It is possible then that the network measures are much more powerful only in combination with meta-data.

Generally, prediction results were more accurate for motor scores than for cognitive scores. It is likely that this is mainly due to motor scores having a higher accuracy of assessment at 18 months of age; our ability to accurately assess cognitive function improves over time, as more functions can be assessed with age. The disparity in prediction accuracy could also be partly due to motor scores having a simpler dependence on the input features compared to the cognitive scores. This is plausible since a few particular regions (e. g., primary motor and premotor cortices) are well known to be responsible for many motor functions (Meier et al., 2008) whereas cognitive function likely relies on a complex network of many regions (which may be unique per cognitive task) (Bressler and Menon, 2010). Furthermore, compared to motor outcomes, cognitive outcomes may be more sensitive to environmental factors not captured by imaging such as maternal education and socioeconomic status (Grunau et al., 2009).

We regard our work as only a proof of concept, showing that filters designed to leverage the structure of the input brain networks can outperform other models on this prediction task. Consequently, as with other non-medical applications of deep learning, given the large number of parameters to be learned, the full potential of CNNs like BrainNetCNN would be realized when applied to applications with much larger neuroimaging datasets, which in turn will require further time and effort to explore a wide array of architectures and parameter settings. To accelerate this exploration, we make our BrainNetCNN publicly available, downloadable at http://www.BrainNetCNN.cs.sfu.ca. Additionally, here we identify three important avenues for future investigation.

First, while it was found that our connectome based models performed better than the models learned from clinical features, it is likely that these features may contribute complementary information to that derived from the connectomes. If the features from both sources could be combined intelligently, the prediction accuracies would likely increase.

Second, as noted, a lack of training data is a major challenge for complex models like the ones proposed. However, other works have shown that transfer learning can occur after pre-training a deep convolutional neural network over larger, similar datasets (Donahue et al., 2014). Since diffusion tensor images of preterm infants are difficult to acquire, perhaps pre-training BrainNetCNN with connectomes from infants born at term or other similar data could improve its predictive ability.

Finally, how to generate the most realistic synthetic training data is still an open research question. We were motivated to attempt to perform data augmentation here because it was clear that even with convolutional filters, the number of parameters to learn in a deep network is high. It is possible that a more advanced data augmentation strategy than SMOTE could perform better. We plan to extend our recently proposed LSI method from the context of binary classification to regression in the hope that it would perform better than SMOTE for this sparsely sampled, high-dimensional data (Brown et al., 2015). We expect that by improving our approach in these ways, we will move towards achieving clinically useful predictive power.

When visualizing which edges BrainNetCNN selected as most predictive of positive cognitive and motor outcomes, it was found that many edges were common to both tasks. This is not surprising since, as mentioned above the two scores are well correlated and since BrainNetCNN was trained to predict both scores simultaneously. However, it may also indicate that some of these common connections in the brain are ones which are at higher risk for damage from the external factors that can lead to poor neurodevelopmental outcomes (e.g., white matter injury and infection) and thus are good common predictors of healthy outcomes. The right middle-frontal gyrus (MFG), in particular, was connected to many strong predictors of both outcomes (Fig. 6). This region is known to be associated with spatial memory (Leung et al., 2002), recognition and recall (Rugg et al., 1996), among other functions, and so may be of particular importance for high Bayley-III scores. However, we note that 18 month outcomes have limited sensitivity to distinguish specific motor and cognitive skills. A longer term follow-up of this cohort is underway and will be helpful to examine specificity of these connections.

Fig. 6 also showed that the most predictive connections of both outcomes had clear laterality. Ratnarajah et al. (2013) found asymmetric functional specializations in the structure of the neonatal connectome. Our finding of laterality may then be due to connections between specific asymmetrical functional regions of the brain that are important for the Bayley-III cognitive and motor tests.

In terms of motor outcomes specifically, we found that the right precentral gyri (PreCG) was highly predictive of motor function, which is expected since the PreCG houses the primary motor cortex. Similarly, the premotor cortex is located, at least in part, within the superior frontal gyri (SFGdor), which were found to be connected to many strongly predictive edges, especially in the left hemisphere. One connection very predictive of cognitive outcomes and not motor outcomes was that between the left FFG and right LING. Both regions have been found to be associated with reasoning about sequences of events (Brunet et al., 2000), however exactly why this particular link is important for prediction of the cognitive outcomes, is unclear. Again, a longer term follow-up may help answer these kinds of questions.

Edges found to be important for prediction of PMA were much more widespread across the brain network compared to those for predicting neurodevelopmental outcomes. This is expected since the whole brain is developing during this early period of development (i.e., many connections changing with age) whereas motor or cognitive functions depend predominantly on specific subnetworks (Betzel et al., 2013). One connection that stood out as being especially positively predictive of age was between the right LING and FFG. This result is consistent with our analysis of the development of healthy preterm infants (Brown et al., 2014).

Generally, extracting the important features from the trained network provides candidate regions and connections for further investigation. This is especially important given the complexity of the brain and what remains to be fully understood about its function and development.

5. Conclusions

In this work we presented BrainNetCNN, the first CNN regressor for connectome data. We introduced three specialized convolutional layer types, designed to leverage the structure inherent in weighted brain networks. We first demonstrated the ability of our framework to

learn multiple independent injury patterns to brain networks by predicting the input parameters of each instance in a realistic phantom dataset. We then tested BrainNetCNN on a set of 168 preterm infant brain networks and showed that our method was able to predict Bayely cognitive and motor scores 18 months into the future. Cognitive and motor scores predicted by BrainNetCNN had significantly higher correlations to the ground truth scores than those predicted by competing methods. Finally, those edges that were learned by BrainNetCNN to be important for each neurodevelopmental outcome were visualized. We found that, as expected, connections from the premotor and primary motor cortices were found to be predictive of better motor outcomes. We also found a general asymmetry in the important connections consistent with other reports in the literature.

Acknowledgments

The authors thank the families for their participation, as well as Anne Synnes for her assistance with gathering and interpreting the infant data and also the staff in the Neonatal Follow-Up Program of BC Children's & Women's Hospitals for their valuable contribution in assessing these children. This work is supported by Canadian Institutes for Health Research (CIHR) operating grants MOP-79262 (S.P.M.) and MOP-86489 (R.E.G.). S.P.M. is currently the Bloorview Children's Hospital Chair in Pediatric Neuroscience. R.E.G. is supported by a Senior Scientist Award from the Child & Family Research Institute. J.G.Z. is supported by a Michael Smith Foundation for Health Research Scholar Award and a Canadian Child Health Clinician Scientist Program Career Enhancement Award. We also thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for partial funding. Finally, we thank the reviewers for their valuable feedback that resulted in a much improved paper.

References

Back, S.A., Miller, S.P., 2014. Brain injury in premature neonates: a primary cerebral dysmaturation disorder? Ann. Neurol. 75 (4), 469–486.
Ball, G., Pazderova, L., Chew, A., Tusor, N., Merchant, N., Arichi, T., Allsop, J.M., Cowan, F.M., Edwards, A.D., Counsell, S.J., 2015. Thalamocortical connectivity predicts cognition in children born preterm. Cereb. Cortex.
Bayley, N., 2006. Manual for the Bayley Scales of Infant Development 3rd edition. Harcourt, San Antonio.
Bear, M. Laurel, 2004. Early identification of infants at risk for developmental disabilities. Pediatr. Clin. N. Am. 51 (3), 685–701.
Betzel, Richard F., Griffa, Alessandra, Avena-Koenigsberger, Andrea, Goñi, Joaquín, Thiran, Jean-Philippe, Hagmann, Patric, Sporns, Olaf, 2013. Multi-scale community organization of the human structural connectome and its relationship with resting-state functional connectivity. Netw. Sci. 1 (03), 353–373.
Booth, B.G., Miller, S.P., Brown, C.J., Poskitt, K.J., Chau, V., Grunau, R.E., Synnes, A.R., Hamarneh, G., 2016. Steamstatistical template estimation for abnormality mapping: a personalized DTI analysis technique with applications to the screening of preterm infants. NeuroImage 125, 705–723.
Bressler, Steven L., Menon, Vinod, 2010. Large-scale brain networks in cognition: emerging methods and principles. Trends Cognit. Sci. 14 (6), 277–290.
Brosch, T., Tam, R., 2013. Manifold learning of brain MRIs by deep learning. In: MICCAI, Lecture Notes in Computer Science, vol. 8150. Springer International Publishing, Cham, pp. 633–640.
Brown, C.J., Miller, S.P., Booth, B.G., Andrews, S., Chau, V., Poskitt, K.J., Hamarneh, G., 2014. Structural network analysis of brain development in young preterm neonates. NeuroImage 101, 667–680.
Brown, C.J., Miller, S.P., Booth, B.G., Poskitt, K.J., Chau, V., Synnes, A.R., Zwicker, J.G., Grunau, R.E., Hamarneh, G., 2015. Prediction of motor function in very preterm infants using connectome features and local synthetic instances. In: MICCAI, Lecture Notes in Computer Science, vol. 9349. Springer International Publishing, Cham, pp. 69–76.
Bruna, J., Zaremba, W., Szlam, A., LeCun, Y., 2013. Spectral Networks and Locally Connected Networks on Graphs. arXiv preprint, p. 14. http://arXiv:1312.6203.
Brunet, Eric, Sarfati, Yves, Hardy-Baylé, Marie-Christine, Decety, Jean, 2000. A pet investigation of the attribution of intentions with a nonverbal task. Neuroimage 11 (2), 157–166.
Cardoza, Jimmy D., Goldstein, Ruth B., Filly, Roy A., 1988. Exclusion of fetal ventriculomegaly with a single measurement: the width of the lateral ventricular atrium. Radiology 169 (3), 711–714.
Chau, V., Synnes, A., Grunau, R.E., Poskitt, K.J., Brant, R., Miller, S.P., 2013. Abnormal brain maturation in preterm neonates associated with adverse developmental outcomes. Neurology 81 (24), 2082–2089.
Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16 (1), 321–357.
Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In: NIPS, pp. 2843–2851.
Cireşan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: MICCAI, Lecture Notes in Computer Science, vol. 8150. Springer International Publishing, Cham, pp. 411–418.
Cuingnet, R., Rosso, C., Chupin, M., Lehéricy, S., Dormont, D., Benali, H., Samson, Y., Colliot, O., 2011. Spatial regularization of SVM for the detection of diffusion alterations associated with stroke outcome. Med. Image Anal. 15 (5), 729–737.
Dodero, L., Minh, H.Q., Biagio, M.S., Murino, V., Sona, D., 2015. Kernel-based classification for brain connectivity graphs on the Riemannian manifold of positive definite matrices. In: 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI), pp. 42–45.
Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T., 2014. DeCAF: a deep convolutional activation feature for generic visual recognition. ICML 32, 647–655.
Dvorak, Pavel, Menze, Bjoern, 2015. Structured prediction with convolutional neural networks for multimodal brain tumor segmentation. In: Proceeding of the Multimodal Brain Tumor Image Segmentation Challenge, pp. 13–24.
Ghanbari, Y., Smith, A.R., Schultz, R.T., Verma, R., 2014. Identifying group discriminative and age regressive sub-networks from DTI-based connectivity via a unified framework of non-negative matrix factorization and graph embedding. Med. Image Anal. 18 (8), 1337–1348.
Godsil, C., Royle, G.F., 2013. Algebraic graph theory, In: Graduate Texts in Mathematics, vol. 207. Springer Science & Business Media, Berlin.
Graham, B., 2014. Spatially Sparse Convolutional Neural Networks. arXiv Preprint. pp. 1–13. http://arXiv:1409.6070.
Grunau, R.E., Whitfield, M.F., Petrie-Thomas, J., Synnes, A.R., Cepeda, I.L., Keidar, A., Rogers, M., MacKay, M., Hubber-Richard, P., Johannesen, D., 2009. Neonatal pain, parenting stress and interaction, in relation to cognitive and motor development at 8 and 18 months in preterm infants. Pain 143 (1), 138–146.
Henaff, M., Bruna, J., Lecun, Y., 2015. Deep Convolutional Networks on Graph-Structured Data. arXiv preprint, pp. 1–10.
Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe : Convolutional architecture for fast feature embedding. In: ACM Conference on Multimedia, pp. 675–678.
Jie, B., Zhang, D., Wee, C.-Y., Shen, D., 2014. Topological graph kernel on multiple thresholded functional connectivity networks for mild cognitive impairment classification. Hum. Brain Map. 35 (7), 2876–2897.
Krzywinski, M.I., Schein, J.E., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19 (9), 1639–1645.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.
Leung, H.-C., Gore, J.C., Goldman-Rakic, P.S., 2002. Sustained mnemonic response in the human middle frontal gyrus during on-line storage of spatial memoranda. J. Cogn. Neurosci. 14 (4), 659–671.
Li, F., Tran, L., Thung, K-H, Ji, S., Shen, D., Li, J., 2014. Robust Deep Learning for Improved Classification of AD/MCI Patients. In: Proceedings of MICCAI 2014 Machine Learning in Medical Imaging (MLMI) Workshop, pp. 240–247.
Liu, S., Liu, S., Cai, W., Pujol, S., Kikinis, R., Feng, D., 2014. Early diagnosis of Alzheimer's disease with deep learning. In: IEEE ISBI, Beijing, IEEE, pp. 677–680.
Meier, Jeffrey D., Aflalo, Tyson N., Kastner, Sabine, Graziano, Michael S.A., 2008. Complex organization of human primary motor cortex: a high-resolution fmri study. J. Neurophysiol. 100 (4), 1800–1812.
Miller, Steven P., Ferriero, Donna M., Leonard, Carol., Piecuch, Robert., Glidden, David V., Partridge, J Colin., Perez, Marta., Mukherjee, Pratik., Vigneron, Daniel B., Barkovich, A. James, 2005. Early brain injury in premature newborns detected with magnetic resonance imaging is associated with adverse early neurodevelopmental outcome. J. Pediatr. 147 (5), 609–616.
Munsell, B.C., Wee, C.-Y., Keller, S.S., Weber, B., Elger, C., da Silva, L.A.T., Nesland, T., Styner, M., Shen, D., Bonilha, L., 2015. Evaluation of machine learning algorithms for treatment outcome prediction in patients with epilepsy based on structural connectome data. NeuroImage 118, 219–230.
Papile, Lu-Ann, Burstein, Jerome, Burstein, Rochelle, Koffler, Herbert, 1978. Incidence and evolution of subependymal and intraventricular hemorrhage: a study of infants with birth weights less than 1,500 gm. J. Pediatr. 92 (4), 529–534.
Plis, S.M., Hjelm, D.R., Slakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H., Paulsen, J., Turner, J., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. 8 ((8) July), 1–11.
Ratnarajah, N., Rifkin-Graboi, A., Fortier, M.V., Chong, Y.S., Kwek, K., Saw, S.M., Godfrey, K.M., Gluckman, P.D., Meaney, M.J., Qiu, A., 2013. Structural connectivity asymmetry in the neonatal brain. NeuroImage 75, 195–202.
Roth, H.R., Lu, L., Farag, A., Shin, H-C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. DeepOrgan: multi-level deep convolutional networks for automated pancreas segmentation. In: MICCAI, vol. 9349, pp. 556–564.
Rugg, M.D., Fletcher, P.C., Frith, C.D., Frackowiak, R.S.J., Dolan, R.J., 1996. Differential activation of the prefrontal cortex in successful and unsuccessful memory retrieval. Brain 119 (6), 2073–2083.
Shi, F., Yap, P.-T., Wu, G., Jia, H., Gilmore, J.H., Lin, W., Shen, D., 2011. Infant brain atlases from neonates to 1-and 2-year-olds. PLoS One 6 (4), e18746.
Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P., 2013. The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Process. Mag. 30 (3),

83–98.
Shuman, D.I., Ricaud, B., Vandergheynst, P., 2012. A windowed graph Fourier transform. In: Statistical Signal Processing Workshop (SSP), IEEE, pp. 133–136.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: ICLR.
Simonyan, Karen, Vedaldi, Andrea, Zisserman, Andrew, 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency smaps. arXiv preprint http://arXiv:1312.6034.
Suk, H.-I., Lee, S.W., Shen, D., 2015. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220 (2), 841–859.
Suk, H.-I., Lee, S.-W., Shen, D., Alzheimer's Disease Neuroimaging Initiative, et al., 2014. Hierarchical feature representation and multimodal fusion with deep learning for ad/mci diagnosis. NeuroImage 101, 569–582.
World Health Organization, November 2014. Preterm Birth Fact Sheet No. 363. ⟨http://www.who.int/mediacentre/factsheets/fs363/en/⟩. Accessed 2015-19-08.
Yang, Z., Zhong, S., Carass, A., Ying, S.H., Prince, J.L., 2014. Deep learning for cerebellar ataxia classification and functional score regression. In: Proceedings of MICCAI 2014 Machine Learning in Medical Imaging (MLMI) Workshop. Springer International Publishing, Cham, pp. 68–76.
Yoo, Y., Brosch, T., Traboulsee, A., Li, D.K., Tam, R., 2014. Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation. In: Machine Learning in Medical Imaging. Springer International Publishing, Cham, pp. 117–124.
Zhu, D., Shen, D., Jiang, X., Liu, T., 2014. Connectomics signature for characterizaton of mild cognitive impairment and schizophrenia. In: IEEE ISBI, pp. 325–328.
Ziv, E., Tymofiyeva, O., Ferriero, D.M., Barkovich, A.J., Hess, C.P., Xu, D., 2013. A machine learning approach to automated structural network analysis: application to neonatal encephalopathy. PloS One 8 (11), e78824.