EVALUATION OF WEIGHT SPARSITY REGULARIZION SCHEMES OF DEEP NEURAL NETWORKS APPLIED TO FUNCTIONAL NEUROIMAGING DATA

*Hyun-Chul **Kim**, Jong-Hwan **Lee***

Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea
{hyunchul_kim, jonghwan_lee}@korea.ac.kr

ABSTRACT

The paper presented a systematic evaluation of the weight sparsity regularization schemes for the deep neural networks applied to the whole brain resting-state functional magnetic resonance imaging data. The weight sparsity regularization was deployed between the visible and hidden layers of the Gaussian-Bernoulli restricted Boltzmann machine (GB-RBM), in which the L0-norm based non-zero value ratio and L1-/L2-norm based Hoyer's sparseness were used to define the weight sparsity. Also, the weight sparsity regularization schemes between the two consecutive layers (i.e. layer-wise) and between the layer and the node in the subsequent layer (i.e. node-wise) were compared in terms of the convergence property. Finally, the reproducibility of 10 sets of weight features extracted from the GB-RBMs trained using 10 sets of random initial weights was evaluated.

Index Terms—Deep neural network, Gaussian-Bernoulli restricted Boltzmann machine, Hoyer's sparseness, Human Connectome Project, weight sparsity

1. INTRODUCTION

The deep neural network (DNN) with an explicit sparsity control of weights [1-3] has recently been shown its efficacy (1) to enhance the classification performance and (2) to extract the hierarchical feature representations from the functional magnetic resonance imaging (fMRI) data. Using this method, the sparsity level of each weight matrix between two consecutive layers (i.e. layer-wise weight sparsity regularization) was explicitly controlled by the percentage of non-zero weights (PNZ) [1, 2].

This layer-wise sparsity control via the PNZ is advantageous due to its computational simplicity. However, the convergence of the PNZ-based sparsity regularization is potentially sensitive to the learning rates during the DNN training [2] due to its dependence on the threshold to define a non-zero value since the PNZ is scale-variant. To address this issue, the Hoyer's sparseness (HSP) measure which is based on the ratio between the L1-/L2-norms and is scale-invariant was adopted [4, 5].

The weight sparsity level can also be defined from the weight vector between the layer and the node in the subsequent layer (i.e. node-wise weight sparsity regularization). Fine-grained sparsity regularization would be achieved from the node-wise sparsity regularization by minimizing the variability of the sparsity levels across weight feature vectors within a layer (see Fig. S2-S3 in [1]).

To investigate the aforementioned motivations, the Gaussian-Bernoulli restricted Boltzmann machine (GB-RBM) [6-8] was employed in this study compared to the auto-encoders (AE) based on minimization of mean-squares error between input and reconstructed input data by learning hidden representations of input samples [5]. Also, the whole brain resting-state fMRI (rfMRI) data in the grayordinates [9, 10] from the Human Connectome Project[1] (HCP) were used as input samples compared to the whole brain functional connectivity patterns of the HCP rfMRI obtained in the volumetric coordinates [5]. The convergence properties of the GB-RBM training from (1) the HSP-based compared to PNZ-based weight sparsity measurement and from (2) the node-wise compared to layer-wise weight sparsity regularization were explored across several learning rates to train the GB-RBM. In addition, the reproducibility of the weight features extracted from the GB-RBM was evaluated using ten GB-RBM results trained with 10 sets of random initial weights.

2. METHODS AND MATERIALS

2.1. rfMRI data from Human Connectome Project

In the HCP rfMRI data from 900 subjects release (S900), the rfMRI data from one subject ('100307') were used. A gradient-echo echo-planar-imaging (EPI) pulse sequence was applied to acquire the rfMRI data (time-of-repetition, or TR = 720 ms; time-of-echo, or TE = 33.1 ms; field-of-view = 208×180 mm^2; 2 mm isotropic voxel size; 72 axial slices; slice thickness = 2 mm; multi-band factor = 8; 1,200 EPI volumes). The grayoridnates FIX-Denoised (Extended) rfMRI data [10] from the subject were spatially smoothed on the surface (for the cortices) and on the volume (for the sub-cortices and cerebellum) spaces with 8 mm isotropic Gaussian kernel using the Connectome Workbench command[2] (i.e. 'wb_command –cifti-smoothing').

[1] www.humanconnectome.org
[2] http://www.humanconnectome.org/software/connectome-workbench.html

2.2. Weight sparsity regularized Gaussian-Bernoulli restricted Boltzmann machine (GB-RBM)

The RBM is a single layer computational model with the input layer (with input/visible nodes) and the hidden layer (with hidden nodes). There are undirected connections between the visible nodes and hidden nodes and there is no connection across the nodes in each layer. In the GB-RBM, the values of the visible node and hidden node are modeled as the Gaussian and Bernoulli distributions, respectively. The grayordinates rfMRI data in each time point (i.e. TR) were used as the values in the visible nodes of the GB-RBM.

The cost function of the GB-RBM was defined from the energy function that is inversely proportional to the log-likelihood of the joint probability between the visible and hidden nodes:

$$E(\mathbf{v},\mathbf{h}) = \sum_i \frac{(v_i - b_i)^2}{2\sigma_i^2} - \sum_j c_j h_j - \sum_{i,j} \frac{h_j w_{ji} v_i}{\sigma_i} \quad (1)$$

where $E(\mathbf{v},\mathbf{h})$ is the energy function, $\mathbf{v}$ and $\mathbf{h}$ are visible and hidden nodes, respectively, w_{ji} is a weight that connects the i^{th} visible node and the j^{th} hidden node, $\mathbf{b}$ and $\mathbf{c}$ are the biases of the visible and hidden nodes, respectively, σ_i is the standard deviation of the Gaussian visible node v_i.

By applying the stochastic gradient descent method to the energy function in Eq. (1) with the contrast divergence-1 approximation of the Gibbs sampling (i.e. $\Delta_{RBM}\mathbf{W}(t)$) [8] as well as the L1-/L2-norm regularizations (i.e. elastic net), the learning rule of the GB-RBM weights, $\mathbf{W}$ was defined as follows:

$$\Delta\mathbf{W}(t) = \alpha(t)\{(1-\beta(t))\Delta_{RBM}\mathbf{W}(t) + \beta(t)sign(\mathbf{W}(t)) + \gamma\mathbf{W}(t)\}, \quad (2)$$

where t is an epoch number, $\alpha(t)$ is an overall learning rate (i.e. an initial learning rate was gradually reduced to 10^{-6} after 200 epochs), $\beta(t)$ (≥ 0) and λ are the L1- and L2-norm regularization parameters, respectively. The L2-norm regularization parameter was fixed to 10^{-4} to prevent an over-fitting during the training (i.e. ridge regression) [11] .

2.3. Layer-wise weight sparsity regularization

The $\beta(t)$ in Eq. (2) controls the degree of the L1-norm regularization of weights by balancing the cost function of the GB-RBM and the L1-norm regularization. The weight matrix of the GB-RBM becomes sparser when $\beta(t)$ is higher (and subsequently the cost function of GB-RBM is less stringent), whereas there is no weight sparsity regularization when $\beta(t)$ equals to zero (and thus this is equivalent to the GB-RBM training with the ridge regression). The $\beta(t)$ is adaptively changed based on the difference between the target weight sparsity level and the sparsity level of the current weights using the PNZ:

$$\Delta\beta(t) = \mu \cdot sign\left(\frac{\|\mathbf{W}(t)\|_0}{N} - \rho_{PNZ}\right), \quad (3)$$

where μ is a learning rate (i.e. 10^{-2}), N is a total number of elements of the weight matrix, ρ_{PNZ} is a PNZ-based target weight sparsity level (i.e. 0 to 1; low ρ_{PNZ} indicates the high sparsity level and vice versa), $\|\cdot\|_0$ is the L0-norm that counts the number of non-zero elements in the weight matrix, i.e. any value outside an interval ε (i.e. 10^{-3}) from 0 (i.e. [-ε, ε]) was defined as non-zero value in our study. The $\beta(t)$ value was limited to 0.5 to prevent the potential instability caused by the substantial L1-norm regularization of weights.

The PNZ is potentially sensitive to the overall learning rate α in Eq. (2) due to the scale variance of the PNZ caused by the threshold ε. Alternatively, the HSP that is based on the ratio between the L1- and L2-norms of the weights is scale-invariant and thus the HSP can be a viable option to substitute the PNZ. Also, the HSP is less computationally demanding compared to the Gini index, in which these are the two most reliable measures of sparsity [12].

The update term of $\beta(t)$ using the HSP was defined as follows:

$$\Delta\beta(t) = \mu \cdot sign\left(\rho_{HSP} - \frac{\sqrt{N} - \|\mathbf{W}(t)\|_1 / \|\mathbf{W}(t)\|_2}{\sqrt{N}-1}\right), \quad (4)$$

where ρ_{HSP} is the HSP-based target sparsity level (i.e. 0-1; 0 being minimum sparsity and 1 being maximum sparsity).

2.4. Node-wise weight sparsity regularization

The layer-wise regularization of the weight sparsity scheme in Eq. (4) can be modified into the node-wise weight sparsity regularization as follows:

$$\Delta\beta_i(t) = \mu \cdot sign\left(\rho_{i,HSP} - \frac{\sqrt{N} - \|\mathbf{W}_{(i,:)}(t)\|_1 / \|\mathbf{W}_{(i,:)}(t)\|_2}{\sqrt{N}-1}\right), \quad (5)$$

where i represents the hidden node index.

2.5. The GB-RBM training

The numbers of the input nodes and hidden nodes of the GB-RBM were 91,282 and 20, respectively. The GB-RBM was trained using Eq. (2) and Eqs. (3-5) depending on each of the three scenarios of the weight sparsity regularization. Also, the GB-RBM without weight sparsity regularization was trained to evaluate the efficacy of the sparsity regularization schemes. The hyperbolic tangent was used as the activation function of the hidden node. The maximum number of epoch was 1,000. The mini-batch size was 200 and a momentum factor was 0.6 [13]. The MATLAB implementation[3] of the GB-RBM algorithm was modified with an extension of the weight sparsity regularization schemes denoted in Eqs. (3-5). The graphic processing unit installed hardware (Intel i7-4790 3.6 GHz; 8 cores; NVIDIA

[3] github.com/rasmusbergpalm/DeepLearnToolbox

GeForce GTX980; 32 GB RAM; Ubuntu 15.10) and the MATLAB (R2015b) computing environment were used.

2.5. Performance evaluation

The convergence property of the weight sparsity regularization scheme and converged weight sparsity levels were measured using the kurtosis and Gini index across three learning rates (i.e. α = 0.002, 0.001, or 0.0005). Uniformly distributed random weights were used to initialize the GB-RBM and the GB-RBM training was repeated 10 times using each of 10 sets of the random initial weights. Three target PNZ and HSP levels (i.e. ρ_{PNZ} = 0.1, 0.2, 0.3; ρ_{HSP} = 0.6, 0.7, 0.8) were used. The 10 sets of the 20 trained weight feature vectors for each of the 20 hidden nodes of the GB-RBM were subject to the quantitative evaluation using the ICASSO [14]. Finally, the reproducibility of the extracted weight feature vectors from each of the layer-wise PNZ-/HSP-based and node-wise HSP-based weight regularization schemes was evaluated using the cluster quality index [14].

3. RESULTS

Fig. 1 shows the learning curves of the layer-wise weight sparsity regularization schemes. Overall, the weight sparsity levels (bold lines) reached to the target levels and mean-squared errors (MSEs; dotted lines) were well converged. The weight sparsity levels from the HSP-based measurement were stable across the epochs, whereas these from the PNZ-based measurement showed moderate fluctuations while converging to the target sparsity levels particularly when the MSE reached to the minimum values.

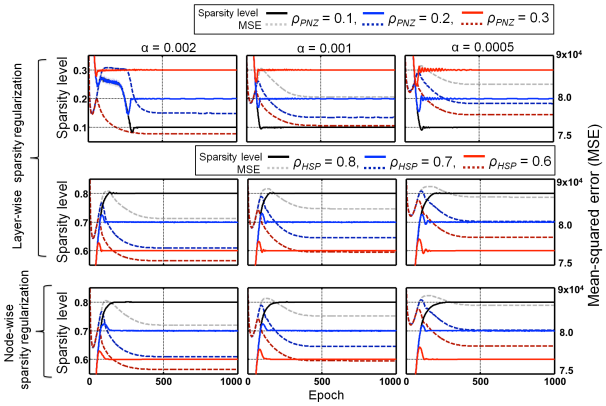

Fig. 1. Learning curves of (1) the weight sparsity level (bold line) and (2) MSE (dotted lines).

Fig. 2 shows the kurtosis values of the weights during the GB-RBM training. Overall, the kurtosis values were stabilized in approximately 400 epochs. The HSP-based sparsity measurement showed the robust kurtosis values across the learning rates. However, when the target PNZ-based sparsity levels were 0.1 and 0.2, the converged kurtosis values from the high learning rate (i.e. α = 0.002) were greater than these from the two reduced learning rates (i.e. 0.001 or 0.0005).

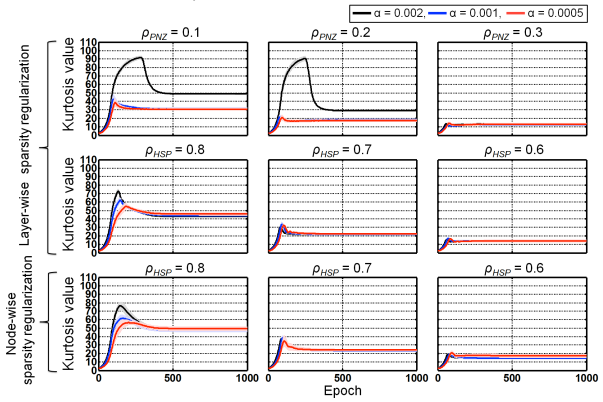

Fig. 2. Learning curves of the kurtosis value. Average (bold line) and standard deviation (shaded area) of kurtosis obtained from the GB-RBM trained using the 10 randomly initialized weights for each of the three learning rates were shown.

Fig. 3 shows the learning curves of the Gini indices of the learned weights. Overall, the Gini indices were stabilized after 500 epochs. The Gini indices from the HSP-based regularization appear to be more robust across the learning rates compared to these from the PNZ-based regularization.

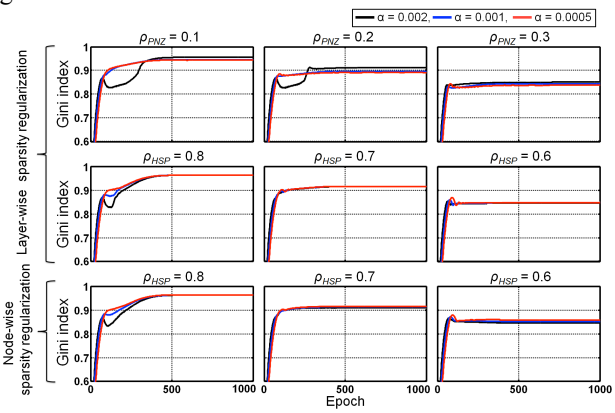

Fig. 3. Learning curves of the Gini index. Average (bold line) and standard deviation (shaded areas) of the Gini indices obtained from the GB-RBM trained using the 10 randomly initialized weights for each of the three learning rates were shown.

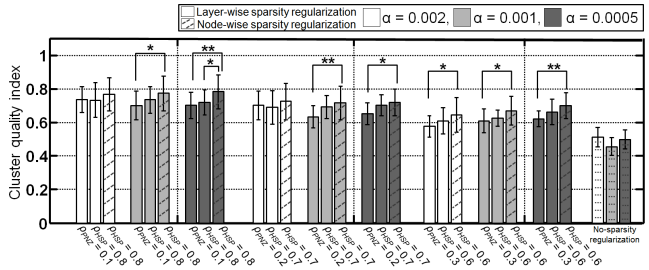

Fig. 4. The cluster quality indices (mean and standard deviation) obtained from the ICASSO using 10 sets of 20 weight feature vectors trained from the GB-RBM (** and * indicate uncorrected p-value < 0.01 and 0.05 from the two-sample t-test, respectively).

Fig. 4 shows the cluster quality indices estimated from the ICASSO using the 10 sets of the 20 trained GB-RBM weight vectors. Overall, the cluster quality indices were enhanced when the sparsity level increases. Notably, by employing the HSP-based measurement, the cluster quality indices were greater from the node-wise regularization than the layer-wise and no-sparsity regularizations.

Fig. 5 shows the representative weight features learned from the HSP-based node-wise regularization scheme. The trained weight features (cluster #02) representing the default mode networks (DMN) showed the greatest cluster quality index (i.e. 0.92) across the 10 sets of the GB-RBM training. The language-related networks (#04), motor networks (#06), lateral parietal network (#08), and cerebellum areas (#10 and #17) showed the higher cluster quality indices than the remaining clusters.

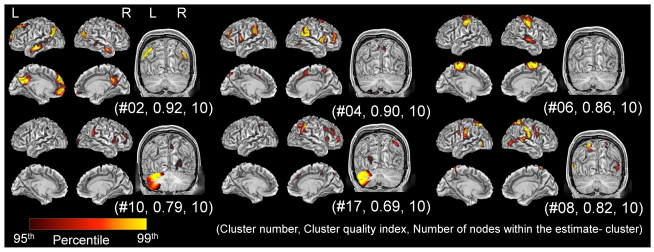

Fig. 5. The representative weight features trained from the HSP-based node-wise regularization (target sparsity level = 0.8 and α = 0.0005; the sign of each weight vector was changed if an element with maximum magnitude is negative; the 95th percentile value or above were visualized using the Connectome Workbench v1.2.3[4]).

4. DISCUSSION

4.1. Summary

In this study, we evaluated the results from the GB-RBM training with an explicit weight sparsity regularization scheme depending on (1) the PNZ-based or HSP-based sparsity level and (2) the layer-wise or node-wise weight sparsity regularization scheme. The reproducibility of the extracted weight features from these regularization schemes was quantitatively compared using the cluster quality index estimated from the ICASSO [14]. The target sparsity levels were well converged from both the PNZ-based and HSP-based weight sparsity definition although the sparsity measurement from the PNZ appears to be sensitive to a learning rate scale. Also, the HSP-based node-wise weight sparsity regularization showed the enhanced reproducibility of the extracted weight features compared to the HSP-based layer-wise weight sparsity regularizations.

Interestingly, the obtained weight features represented the popular spatial patterns of the rfMRI networks such as the DMN and motor networks. The visual networks (data not shown) showed a high cluster quality index (i.e. 0.78), however the number of features of the corresponding cluster was 14 greater than the number of GB-RBM training (i.e. 10) and thus this warrants a future investigation. It is also worth to investigate that eight out of 20 features represented the cerebellum networks in the slightly shifted areas and the weight features in the sub-cortical areas were not obtained in this study. This may be related to the fact that the 8mm spatial smoothing applied to the cerebellum and sub-cortical areas was performed in the volumetric space.

4.2 Future works

The GB-RBM was employed in this study compared to the AE in the previous study [5] and the weight features of the GB-RBM seem to show the greater reproducibility than these from the AEs (data not shown). A future study is warranted to systematically compare the weight features obtained from the GB-RBM to alternative unsupervised learning methods such as independent component analysis [15] and AEs. The application of the greedy layer-wise trained GB-RBM (i.e. deep belief network, or DBN) with the weight sparsity regularization scheme to the rfMRI data would constitute an interesting future study as the DBN could extract the hierarchically organized feature representations of the task-based fMRI data [2]. Consequently, the reproducibility of the hierarchically organized weight feature representations of the rfMRI data deserves future investigation. Furthermore, the weight sparsity regularization can be optimized from the constrained optimization scheme via the Lagrangian multiplier [16] rather than the grid search from the candidate weight sparsity levels in the nested cross-validation framework [1, 2]. It would be straightforward to apply the presented findings to our earlier study presenting the efficacy of a DNN with the PNZ based regularization scheme toward the classification of the schizophrenia and healthy subjects [1].

5. CONCULSION

In this study, the performance of the weight sparsity regularization schemes was evaluated across various scenarios depending on whether the weight sparsity was calculated from the PNZ or HSP measurement during the GB-RBM training using the whole brain rfMRI volumes as input. The convergence property to reach to a target sparsity level and consistency of the converged sparsity level seems to be superior from the HSP-based measurement than the PNZ-based measurement. Also, the node-wise weight sparsity regularization scheme presented the enhanced reproducibility for each of the trained weight features than the layer-wise weight sparsity and sparsity regularization-free schemes. It would be straightforward to extend our reported methods/findings to the stacked-GB-RBM training and to extract the hierarchically organized features from the rfMRI data. As a result, DNN applications with the weight sparsity regularization scheme to classify and/or to predict neurological/neuropsychiatric disorders may be feasible by circumventing the curse-of-dimensionality issue [1, 2, 16].

[4] http://www.humanconnectome.org/software/connectome-workbench.html

6. REFERENCES

[1] Kim, J., Calhoun, V.D., Shim, E., and Lee, J.-H.: 'Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia', NeuroImage, 2016, 124, pp. 127-146

[2] Jang, H., Plis, S.M., Calhoun, V.D., and Lee, J.-H.: 'Task-specific feature extraction and classification of fMRI volumes using a deep neural network initialized with a deep belief network: Evaluation using sensorimotor tasks', NeuroImage, 2016

[3] Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J.A., and Calhoun, V.D.: 'Deep learning for neuroimaging: a validation study', Frontiers in neuroscience, 2014, 8, pp. 229

[4] Hoyer, P.O.: 'Non-negative matrix factorization with sparseness constraints', The Journal of Machine Learning Research, 2004, 5, pp. 1457-1469

[5] Kim, H.C., and Lee, J.H.: 'Evaluation of weight sparsity control during autoencoder training of resting-state fMRI using non-zero ratio and Hoyer's sparseness'. Proc. PRNI 2016: the 6th International Workshop on Pattern Recognition in Neuroimaging, Trento, Italy 2016 pp. Pages

[6] Hinton, G.E., and Salakhutdinov, R.R.: 'Reducing the dimensionality of data with neural networks', Science (New York, N.Y.), 2006, 313, (5786), pp. 504-507

[7] Hjelm, R.D., Calhoun, V.D., Salakhutdinov, R., Allen, E.A., Adali, T., and Plis, S.M.: 'Restricted Boltzmann machines for neuroimaging: an application in identifying intrinsic networks', NeuroImage, 2014, 96, pp. 245-260

[8] Hinton, G.E., Osindero, S., and Teh, Y.-W.: 'A fast learning algorithm for deep belief nets', Neural computation, 2006, 18, (7), pp. 1527-1554

[9] Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., and Jenkinson, M.: 'The minimal preprocessing pipelines for the Human Connectome Project', NeuroImage, 2013, 80, pp. 105-124

[10] Glasser, M.F., Coalson, T.S., Robinson, E.C., Hacker, C.D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C.F., Jenkinson, M., Smith, S.M., and Van Essen, D.C.: 'A multi-modal parcellation of human cerebral cortex', Nature, 2016, 536, (7615), pp. 171-178

[11] Witten, I.H., and Frank, E.: 'Data Mining: Practical machine learning tools and techniques' (Morgan Kaufmann, 2005. 2005)

[12] Hurley, N., and Rickard, S.: 'Comparing measures of sparsity', Information Theory, IEEE Transactions on, 2009, 55, (10), pp. 4723-4741

[13] Bishop, C.M.: 'Neural networks for pattern recognition', 1995

[14] Correa, N., Adali, T., and Calhoun, V.D.: 'Performance of blind source separation algorithms for fMRI analysis using a group ICA method', Magn Reson Imaging, 2007, 25, (5), pp. 684-694

[15] McKeown, M.J., Makeig, S., Brown, G.G., Jung, T.P., Kindermann, S.S., Bell, A.J., and Sejnowski, T.J.: 'Analysis of fMRI data by blind separation into independent spatial components', Human brain mapping, 1998, 6, (3), pp. 160-188

[16] Bengio, Y., Goodfellow, I.J., and Courville, A.: 'Deep learning', An MIT Press book in preparation. Draft chapters available at http://www.iro.umontreal.ca/~bengioy/dlbook, 2015