PIMMS: Permutation Invariant Multi-modal Segmentation

Thomas Varsavsky[1(✉)], Zach Eaton-Rosen[1,2(✉)], Carole H. Sudre[1,2,3(✉)], Parashkev Nachev[4(✉)], and M. Jorge Cardoso[1,2(✉)]

[1] CMIC, University College London, London, UK
{ucabtmv,z.eaton-rosen,carole.sudre.12}@ucl.ac.uk
[2] School of Biomedical Engineering and Imaging Sciences, King's College London, London, UK
m.jorge.cardoso@kcl.ac.uk
[3] Dementia Research Centre, University College London, London, UK
[4] Institute of Neurology, University College London, London, UK
p.nachev@ucl.ac.uk

Abstract. In a research context, image acquisition will often involve a pre-defined static protocol and the data will be of high quality. If we are to build applications that work in hospitals without significant operational changes in care delivery, algorithms should be designed to cope with the available data in the best possible way. In a clinical environment, imaging protocols are highly flexible, with MRI sequences commonly missing appropriate sequence labeling (e.g. T1, T2, FLAIR). To this end we introduce PIMMS, a Permutation Invariant Multi-Modal Segmentation technique that is able to perform inference over sets of MRI scans without using modality labels. We present results which show that our convolutional neural network can, in some settings, outperform a baseline model which utilizes modality labels, and achieve comparable performance otherwise.

1 Introduction

Over the years, public medical imaging datasets have emerged which enable researchers to benchmark the performance of their algorithms [1]. Data is mostly acquired from patients who have volunteered to be part of a clinical research study and are subject to a strict study protocol. If the study involves the acquisition of Magnetic Resonance Imaging (MRI) scans, the study protocol might dictate the scanner choice as well as the acquisition parameters to be used [4]. In the real unconstrained clinical setting however, MRIs are more likely to be acquired from different machines under different acquisition protocols and parameters. There is no guarantee that a particular sequence will be available, no guarantee on the number of available modalities, no guarantee that modalities will be unique (e.g. same sequence acquired with different orientations and contrasts), and no guarantee that any of the modalities will be labeled appropriately for

D. Stoyanov et al. (Eds.): DLMIA 2018/ML-CDS 2018, LNCS 11045, pp. 201–209, 2018.
https://doi.org/10.1007/978-3-030-00889-5_23

algorithmic use. If hospitals are to benefit from advances in neuroimaging, algorithms that can cope with this lack of available modalities are necessary. We argue that an algorithm which is to be deployed in this setting should have two key properties: (1) permutation invariance, i.e permuting the order of the input images should not affect the output and (2) robustness to missing modalities. To this end we propose a segmentation model, with neural networks as building blocks, which can learn with limited data and segment scans without MR modality labels. In this work we focus on the task of segmenting white matter hyperintensities (WMH). In studies involving WMH segmentation the most common modalities are T1, T2 and T2-FLAIR which provide complementary information about the imaged tissue. Although T1 and T2 modalities are created from different underlying physical signals (longitudinal and transverse relaxation time respectively) the scans produced will almost always be a combination of both (hence the name attribute - weighted). By varying the acquisition parameters, such as the echo and relaxation times, these underlying physical signals are observed in different proportions [3]. Modality labels are a discrete approximation of a continuous acquisition parameter landscape and we use this as inspiration for the model we present.

In order to address missing modalities, research has focused mostly on generative models where missing MRI scans are synthesized or imputed [2,8]. In the work of [6] the authors handle missing modalities without using generative models of MR modalities. Instead of synthesizing the missing modalities, their model, Hetero-modal Image Segmentation (HeMIS), is trained to handle missing input modalities. More details about HeMIS can be found in Sect. 2. Although HeMIS is successful at dealing with missing modalities, it assumes that the MR modalities in a test case will be labeled. The authors of [10] tackle the issue of generalizing to unseen protocols and scanners. In order to be robust to different scanners and protocols, they propose a tuning of the batch normalization parameters of a CNN. However, their method still requires approximately four scans with their associated segmentations from the unseen protocol to perform well.

We introduce a model that learns to build intermediate representations of the images as a linear combination of the available inputs which are more continuous than their original labels. The proposed model does not assume the modality is known and has the ability to generalize to unseen scanners/protocols, taking in N unordered input scans with no modality labels to produce accurate segmentation masks. We provide results on a variety of datasets featuring WMH with large variability in scanner type and acquisition parameters and show that our model is both permutation invariant and robust to missing modalities. We demonstrate that it can perform comparatively well with an algorithm which utilizes the modality labels having never seen an image from that particular protocol. Furthermore, our model can outperform the baseline method (HeMIS) in the case where it has seen MR modality labels of the same protocol it is being tested on.

2 Methods

HeMIS. In HeMIS each available modality, $x_1, \ldots, x_M$, is embedded with a modality specific function $\phi_m(x_m) \in \mathbb{R}^{D \times K}$ denoted the "back-end" to produce embeddings. An "abstraction" layer then operates on these embeddings by computing the mean and variance across their K dimensions and concatenating the two resulting vectors $\phi_\alpha = [\hat{\mathrm{E}}(\boldsymbol{\phi}(\mathbf{x})), \hat{\mathrm{Var}}(\boldsymbol{\phi}(\mathbf{x}))]$, where $\mathbf{x} \in \mathbb{R}^{D \times M}$ M is the number of modalities and D is the spatial dimensions of the input. Let ϕ_α be a fixed dimensional tensor which represents an input of variable size. This forms the input to the final portion of the network referred to as the "frontend" which will output a semantic segmentation map. The network is trained using a Dice loss, first proposed in [11] as a loss function for training neural networks.

During training, random modalities are set to zero, encouraging robustness to missing modalities. HeMIS, shown in Fig. 1, forms part of our architecture.

Our Approach. We propose a method which at test time takes in an arbitrary number of N scans (denoted X) which do not have corresponding MR modality labels and produces a permutation invariant representation that is also robust to missing modalities. In theory this common representation could be applied to a variety of tasks. In this paper we focus on white matter hyperintensity segmentation.

The inputs are fed into an MR modality classifier f_{mod} which outputs a distribution over modalities for a given scan as its prediction. These modality scores $\mathcal{S} \in \mathbb{R}^{M \times N}$ are combined with the inputs, X, to produce modified inputs denoted as $\hat{X} \in \mathbb{R}^{D \times M}$. In the attention literature a distinction is drawn between "soft" and "hard" attention [14]. Soft attention generally involves a probabilistic weighted sum whilst a hard attention is a categorical argmax over the inputs. With this in mind, we explore two methods for performing $X \rightarrow \hat{X}$: f_{soft} and f_{hard}. The function f_{soft} is defined as,

$$f_{soft}(X, \mathcal{S}) = \sum_{n=1}^{N} \mathcal{S}_{mn} x_n = \hat{x}_m \tag{1}$$

Each component $\hat{x}_m$ of the modified input $\hat{X}$ is formed by taking a weighted sum of each input x_n according to the probabilities provided by $\mathcal{S}$. f_{hard} is defined as,

$$f_{hard}(X, \mathcal{S}) = \sum_{n=1}^{N} \mathbb{1}(\arg\max_{m^*} \mathcal{S}_{m^* n} = m) x_n = \hat{x}_m \tag{2}$$

The modified input $\hat{X}$ now consists of a finite number of modalities. The mapping $f : X \rightarrow \hat{X}$ is illustrated in the blue block in Fig. 1.

Each MR modality is designed to capture fundamentally different physical properties which justifies having individual feature extractors, ϕ_m, for each $\hat{x}_m$

modality representation. The output of these modality-specific feature extractors is collected into one tensor by taking the mean and the variance across modalities and concatenating the result to give $\phi_\alpha \in \mathbb{R}^{D \times K}$ where K is given by the choice of filter depth in ϕ_m. This feeds into a final network, ϕ_{seg} which produces a segmentation prediction. This use of modality specific models, pooling and a separate segmentation network is the same as HeMIS and is illustrated in the grey block in Fig. 1.

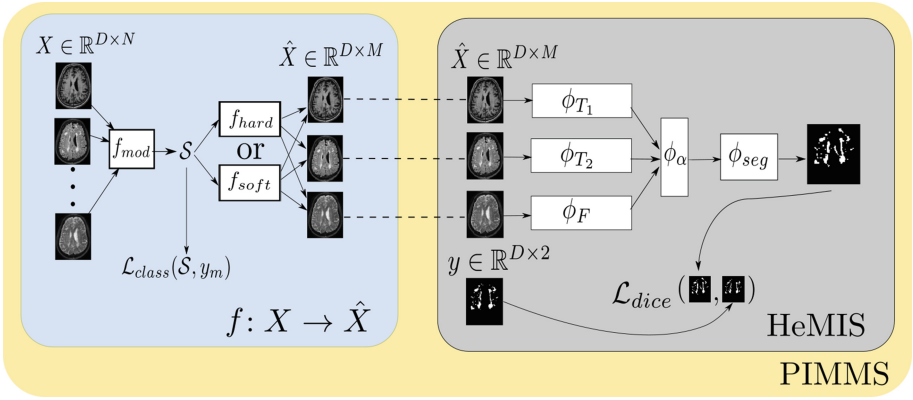

Fig. 1. Diagram showing the network architecture. During training the inputs are $X \in \mathbb{R}^{D \times N}$ and the corresponding ground truth binary segmentation $y \in \mathbb{R}^{D \times 2}$. A function f_{mod} takes each scan as input and outputs a modality score $\mathcal{S}$ which produces the representation $\hat{X} \in \mathbb{R}^{D \times M}$. The weights of ϕ_{T_1}, ϕ_{T_2}, ϕ_F and ϕ_{seg} are learned by differentiating with respect to $\mathcal{L}_{seg}$ and the weights of f are learned by differentiating with respect to $\mathcal{L}_{class}$. y_m is a one-hot encoded modality label.

A convolutional neural network was used for f_{mod}. A network with 36 layers using skip connections and ReLU non-linearities inspired by the residual network (ResNet) proposed in [7] is used. The network was trained with the categorical cross-entropy loss which we refer to as $\mathcal{L}_{class}$. where y_{mi} is a one-hot encoded modality label and $\mathcal{S}_{mi}$ is the modality score. Each of the branches ϕ_m as well as ϕ_{seg} were two convolutional layers with ReLU non-linearities (more details in Sect. 2). The parameters of ϕ_m and ϕ_{seg} were found by minimizing $\mathcal{L}_{seg}$ which is the binary Dice Loss.

For two of our variants these losses were trained separately (or "offline"). However, we also trained an "online" variant where the parameters of the modality classifier are learned using a multi-objective loss function. This loss is defined as, $\mathcal{L}_{tot} = \mathcal{L}_{seg} + \lambda \mathcal{L}_{class}$, where λ is some choice of weighting or parametrized weighting function. Although the loss consists of multiple objectives this should not be considered "multi-task learning". There is no conditional independence between the tasks and no representation sharing — instead this can be seen as a differentiable attention mechanism. The four variants trained are summarized below,

HeMIS - $X \rightarrow \hat{X}$ using labels, f_{mod} trained separately from $\phi_{seg}, \phi_{T_1}, \phi_{T_2}$ & ϕ_F
Soft - f_{soft} used to create $\hat{X}$, f_{mod} trained separately from $\phi_{seg}, \phi_{T_1}, \phi_{T_2}$ & ϕ_F,
Hard - f_{hard} used to create $\hat{X}$, f_{mod} trained separately from $\phi_{seg}, \phi_{T_1}, \phi_{T_2}$ & ϕ_F,
Online - f_{soft} used to create $\hat{X}$, f_{mod} trained jointly with $\phi_{seg}, \phi_{T_1}, \phi_{T_2}$ & ϕ_F.,

Implementation Details. It is important to note that the network architecture takes in 2D patches from the image as was done in [6]. Specifically we take patches of size 100×100 from 3D scans which have all been resampled to $1\,mm \times 1\,mm \times 1\,mm$. This theoretical framework permits any spatial dimension D and future work will train and run inference in full 3D.

All results were obtained using the NiftyNet framework [5], which is a wrapper around TensorFlow designed for medical imaging. f_{mod} uses a standard ResNet design with nine blocks per resolution, each with three convolutions and ReLU activations. The network is trained using the Adam optimizer with a learning rate of 3×10^{-4}. A batch size of 64 was used on this network and weight decay regularization of 1×10^{-4}.

For each ϕ_m and ϕ_{seg} the implementation details from [6] were recreated. Two convolutional layers with 48 filters, 5×5 kernel sizes, zero-padding and ReLU activation were used followed by a max pooling layer with kernel size (2, 2) and a stride of 1 this preserves the spatial resolution of the image. For ϕ_{seg} two convolutional layers were used, one with 16 filters, 5×5 kernel sizes, zero padding and ReLU activation the last convolutional layer had 2 filters, a kernel size of 21×21, zero padding and a softmax activation which provided the per class predictions. We also utilized the pseudo-curriculum learning approach from HeMIS. Random modalities are set to zero but the chance of setting only one or no modalities to zero is higher. The online model was harder to train than the offline ones. The joint training lead to odd dynamics between the classification loss and the segmentation loss. To help stabilize the training an exponential decay weighting was used on the classification loss in order to encourage training it towards the start and remove its importance later on so that the model could experiment with representations which do not match the provided labels and not be punished by $\mathcal{L}_{class}$. Our best performing "online" model used $\lambda(i) = e^{-\gamma i}$ where i is the current iteration and γ is a decay constant hyperparameter set to 1×10^{-4}.

This same ResNet architecture was used as f_{mod} in the online case in order to make a fair comparison in terms of number of parameters. However, in the online setting, the batch size had to be reduced as a practical consideration as the combination of both modality and backend models proved too large to fit in GPU memory. All experiments were run on a single NVIDIA Titan Xp.

3 Experiments and Results

Data used in this work comes from a variety of sources, chosen to try and capture the acquisition variability observed in a practical setting due to multiple MRI

scanners/protocols. A subset of 973 subjects each with T1 and FLAIR scans were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [9]. The data in this study was collected from multiple scanners, but used the same protocol for setting the acquisition parameters. We therefore deem this dataset one of relatively low variance between subjects. We also utilise data collected from the longitudinal SABRE study [13]. The data contains one cohort of 586 subjects with T1, T2 and FLAIR obtained using the same scanner (low variance) and another of 1263 with T1, T2 and FLAIR obtained from multiple scanners with multiple settings (high variance). Additionally we use a dataset of 626 patients with T1 and FLAIR obtained from multiple scanners using multiple field strengths. As no manual annotations were available for this large collection of MRI scans, the outputs of BaMoS [12], a fully unsupervised WM lesion segmentation algorithm, were quality controlled by an experienced human rater and subsequently used as silver-standard training labels. Additionally, we evaluate our trained models on a manually annotated dataset from the MICCAI 2017 White Matter Hyperintensity Challenge [1].

The split between training, validation and test sets was chosen in order to measure the ability of our method at generalizing to unseen scanners and protocols. Three separate holdouts were created, defined as follows,

Silver Protocol Holdout - ADNI: 973 subjects with silver standard labels.
Gold Protocol Holdout - MICCAI2017: 60 subjects with human rater labels.
Mixed Holdout - Random 10% subset of the full data minus Silver/Gold.

Overall there was a 80/10/10 split between training, validation and test using the 2474 subjects that are not in the gold or silver protocol holdouts. All four models described in Sect. 2 were trained with this subset.

Table 1. Dice scores of the different models on different combinations of available modalities. Modalities present are denoted by • and those that are missing are denoted by ○. Bold numbers are results which outperform the baseline model, HeMIS, with statistical significance $p < 0.01$ as provided by a Wilcoxon test. Presentation of table inspired by the one in [6]

Mixed Holdout									
Modalities			Dice Score				Avg. Symmetric Distance			
T_1	T_2	F	HeMIS	Soft	Hard	Online	HeMIS	Soft	Hard	Online
•	•	•	0.47	**0.51**	0.48	**0.54**	0.71	0.65	0.71	1.9
•	•	○	0.3	**0.39**	0.3	0.24	2.32	**1.92**	2.36	4.21
○	•	•	0.26	**0.32**	**0.26**	**0.4**	0.77	0.82	0.76	3.32
•	○	•	0.44	**0.45**	**0.45**	**0.52**	0.61	0.63	0.62	2.06
•	○	○	0.1	**0.1**	0.1	**0.19**	3.42	3.76	3.51	4.48
○	•	○	0.08	**0.08**	0.07	**0.09**	4.07	4.13	4.53	7.48
○	○	•	0.16	**0.18**	**0.16**	**0.41**	0.56	0.61	**0.54**	3.31

Modalities			Dice Score				Avg. Symmetric Distance			
T_1	T_2	F	HeMIS	Soft	Hard	Online	HeMIS	Soft	Hard	Online
Silver Protocol Holdout										
•	○	•	0.48	0.46	0.46	0.44	0.68	0.72	1.12	3.52
•	○	○	0.11	0.11	0.08	**0.21**	0.79	0.79	1.63	5.17
○	○	•	0.25	0.16	0.24	**0.5**	0.69	**0.68**	0.8	2.77
Gold Protocol Holdout										
•	○	•	0.59	0.64	0.62	0.61	0.76	**0.57**	0.72	1.18
•	○	○	0.41	0.35	0.42	**0.47**	0.8	**0.44**	**0.77**	2.18
○	○	•	0.38	0.38	0.26	**0.45**	1.01	1.75	20.75	3.63

For the mixed holdout it was found that the classification accuracy was 99% between all three modalities. For unseen protocols the accuracy was lower, 88% for ADNI and 87% for MICCAI17 which showed that the inter-scanner variance was harder to model than the inter-subject variance. For each of the holdout sets, results are presented on all possible subsets of the available modalities. The quantitative and qualitative results are shown in Table 1 and Fig. 2, respectively. The brains shown are selected from the 95%, 50% and 20% percentile of Dice score on the dataset holdout for a model shown all available modalities. We note that the samples of very high Dice score are often the ones with large lesions which the algorithm has managed to capture well and there is poor performance when the contrast settings are significantly different.

We utilise the Wilcoxon signed-rank test to test whether the Dice scores from each of our models outperforms the baseline (HeMIS). Bold values in Tables 1 denotes that the model is better than HeMIS with a statistical significance of $p < 0.01$. We compare ground truths and predictions using the Dice score as well as the average symmetric distance in order to provide a geometric evaluation.

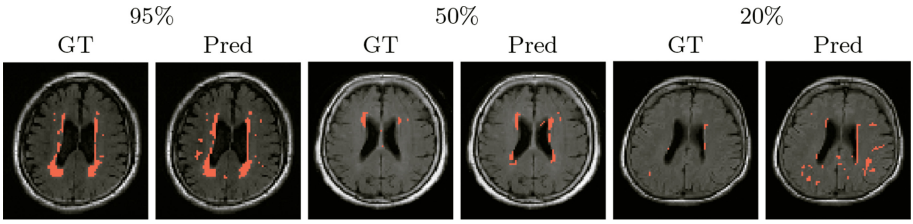

Fig. 2. Qualitative results showing white matter lesion segmentations on the mixed holdout set. Images show the ground truth on the left and the network predictions on the right. Red shows the predicted segmentation. The results were chosen to highlight the 95th, 50th, and 20th percentile in terms of Dice score for a model which is trained on all available scans but does not use modality labels.

4 Discussion

The "hard" setting converges to HeMIS as the accuracy of the modality classifier tends to 1. This is observed in practice. Note that the results of HeMIS are similar to"hard" in the mixed holdout set where the modality classifier has had access to the test set distribution and consistently worse in the Silver Protocol holdout. It does comparatively better on the Gold Protocol as the modality classifier has better performance on these scans than on Silver. The "soft" version matches or improves on the performance of HeMIS and"hard" on the mixed holdout, but does not outperform HeMIS on other holdouts. The fact that "soft" outperfoms"hard" is evidence towards our hypothesis that mixing the input images can lead to better representations which improve performance on a visual task.

This can be interpreted as a coarse attention mechanism as the transformation from X to $\hat{X}$ is linear with few degrees of freedom.

The "online" model outperforms the baseline in the mixed holdout set with statistical significance in 6/7 cases when using the Dice score. Although the median average symmetric distance (ASD) is higher, the average is lower in 4/7 cases with a much lower 95 percentile. There is some improvement over the baseline model even in the protocol holdout but the gains seen in Dice score are not reflected in the ASD. Qualitatively this is explained by the"online" method overpredicting the positive class leading to a higher Dice score and yet missing lesions altogether leading to a larger ASD. This gives us insights as to how we can improve the model.

Future work will extend the "online" model to an unsupervised setting in terms of scan labels. This is appealing not only due to the lack of modality labels currently available in certain hospital databases but also in order to go *beyond* the information contained in the modality label and towards a representation which is more true to the underlying physical structure.

5 Conclusion

We have presented PIMMS, a segmentation algorithm for MRI scans which simultaneously addresses the problem of missing modalities and missing modality labels in a clinical setting. We present three variants which all include a convolutional neural network and are trained to perform modality classification in a supervised setting. We argue that by mixing the input modalities in ratios other than those provided by the labels we can achieve better performance. This could be due to more accurately capturing the underlying distribution of physical quantities, but future work is needed to make this claim. Evidence is presented with statistical significance which suggests that a model which mixes inputs can perform better than one which does not with all other factors kept identical.

The results show that the modality classifier almost replicates modality labels when trained and tested on the same protocol while the categorical accuracy reaches 88% when protocols differ at training and testing times. Our model serves as a proof of concept for a system that could utilize all the MR scans associated with a patient in a hospital and provide accurate segmentation predictions.

Acknowledgements. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. Zach Eaton-Rosen is supported by the EPSRC Doctoral Prize. Carole H. Sudre is supported by the Biomedical Junior Fellowship from Alzheimer's Society. Parashkev Nachev is funded by the Wellcome Trust and the UCLH NIHR Biomedical Research Centre. Jorge Cardoso is funded by Wellcome Trust.

References

1. White Matter Segmentation Grand Challenge at MICCAI 2017
2. Chartsias, A.: Multimodal MR synthesis via modality-invariant latent representation. IEEE Trans. Med. Imaging **37**(3), 803–814 (2017)
3. Fischl, B., et al.: Sequence-independent segmentation of magnetic resonance images. Neuroimage **23**, S69–S84 (2004)
4. Ghafoorian, M., et al.: Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. CoRR abs/1610.04834 (2016)
5. Gibson, E., et al.: Niftynet: a deep-learning platform for medical imaging. arXiv preprint arXiv:1709.03485 (2017)
6. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: hetero-modal image segmentation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 469–477. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_54
7. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 630–645. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_38
8. Iglesias, J.E., Konukoglu, E., Zikic, D., Glocker, B., Van Leemput, K., Fischl, B.: Is synthesizing MRI contrast useful for inter-modality analysis? In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 631–638. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_79
9. Jack, C.R.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging **27**(4), 685–691 (2008)
10. Karani, N., et al.: A lifelong learning approach to brain MR segmentation across scanners and protocols. arXiv:1805.10170 (2018)
11. Milletari, F., et al.: V-Net: fully convolutional neural networks for volumetric medical image segmentation. In: 3DV, pp. 565–571. IEEE (2016)
12. Sudre, C.H.: Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. IEEE TMI **34**(10), 2079–2102 (2015)
13. Tillin, T.: Southall and brent revisited: cohort profile of Sabre, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins. IJEpid **41**(1), 33–42 (2010)
14. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)