Uncertainty-Aware Graph-Based Multimodal Remote Sensing Detection of Out-Of-Distribution Samples

Iain Rolland[1], Andrea Marinoni[1,2] and Sivasakthy Selvakumaran[1]

[1]Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, United Kingdom
[2]Department of Physics and Technology, UiT the Arctic University of Norway, P.O. box 6050 Langnes, NO-9037, Tromsø, Norway

Abstract

Having the ability to quantify prediction confidence or uncertainty will greatly assist the successful integration of deep learning methods into high-stake decision making processes. Graph-based convolutional neural networks can be trained to perform classification of multimodal remote sensing data using a model output which represents a Dirichlet distribution parameterization. This parameterization can then also be used to obtain measures of prediction uncertainty. By making a correspondence between a multinomial opinion, as described by subjective logic, and a Dirichlet distribution parameterization, a direct mapping between the two can be performed. A multinomial opinion of this kind can produce quantified measures of uncertainty and distinguish uncertainty due to a lack of evidence (vacuity) and uncertainty due to conflicting evidence (dissonance). With an appropriately chosen loss function, the graph-based classifier will converge to provide accurate estimates of uncertainty. The results presented in this paper show that the measures of uncertainty provided by such models are capable of better distinguishing out-of-distribution data samples than probabilistic measures of uncertainty produced by equivalent deterministic neural networks.

Keywords

Multimodal remote sensing, uncertainty estimates, graph convolutional networks, subjective logic, land cover classification

1. Introduction

The capability of algorithms to provide accurate measures of confidence and uncertainty is important if they are to be adopted in real-world scenarios where the stakes can be high [1]. Although deep learning methods are often capable of producing high-accuracy predictions [2, 3], they are generally criticized for being unable to express when to have confidence in the prediction and when the prediction should be presented as uncertain. If deep learning models are to be integrated reliably into real-world decision making processes, it is of vital importance that the methods being used are capable of accurately expressing uncertainty [4].

With remotely-sensed data being available with ever-greater temporal and spatial resolutions, the development of computational processing methods which are capable of robustly handling such large volumes of data will assist countless earth-monitoring applications [5]. Specifically, with data being captured now using a wide range of techniques with complementary strengths, the ability to combine this data into a multimodal analysis will allow each data mode to interact synergistically to provide better results than any individual data mode would produce in isolation. Each data capturing technique will naturally have its own strengths and weaknesses, inherent to the physical properties of the sensing mode [6, 7]. Deterministic classification, while useful, is held back by its inability to express uncertainty. Adoption of such techniques will always be limited by the adopter's trust in the predictions. Uncertainty estimates, however, will greatly assist human trust in models, as it will provide a quantification of confidence that might indicate when a prediction is not to be trusted, and more importantly, when a prediction is given with great certainty [8].

In this paper, we have analyzed how well different measures of model uncertainty perform the task of identifying data points which belong to a distribution other than those observed during training (out of distribution detection). To do so, we have used graph-based neural network architectures that are adapted to provide subjective opinions (as described in the field of belief or evidence theory [9]) through the use of Dirichlet distribution parameterizations [10, 11]. The subjective opinions can be used to measure two intuitive measures of uncertainty: vacuity and dissonance. Vacuity is a measure of the uncertainty related to an absence of observed evidence, i.e. a higher measure of vacuity suggests a lack of supporting evidence for a prediction. Dissonance is a measure of prediction uncertainty arising due to the presence of conflicting evidence. This approach (using graph-based neural networks within a subjective-logic framework) is, to the best of our knowledge, as-yet untested as a method for performing classification of multimodal remote sensing

CDCEO 2021: 1st Workshop on Complex Data Challenges in Earth Observation, November 1, 2021, Virtual Event, QLD, Australia.
imr27@cam.ac.uk (Iain Rolland); andrea.marinoni@uit.no (Andrea Marinoni); ss683@cam.ac.uk (Sivasakthy Selvakumaran)
0000-0002-4137-5605 (Iain Rolland); 0000-0001-6789-0915 (Andrea Marinoni); 0000-0002-8591-0702 (Sivasakthy Selvakumaran)

data. The performance of the adopted technique represents a promising avenue in the search for meaningful uncertainty estimates for this task.

The remainder of this paper is organized as follows: Section 2 describes the uncertainty framework adopted in the methods presented, Section 3 details the construction of the graph-based neural networks used, Section 4 presents an analysis of results and Section 5 summarizes and draws conclusions as well as suggests areas for future work.

2. Uncertainty framework

The proposed uncertainty-aware framework relies on the definition of uncertainty metrics, which in turn are based on subjective logic and a Dirichlet mapping [11]. These steps are detailed in this section, and have been properly adapted to the task of multimodal remote sensing classification.

2.1. Subjective Logic

Subjective Logic (SL), takes an evidence-based approach to decision making [12]. Expressing an opinion using measured quantities of belief allows the distinction to be made between uncertainty due to a lack of evidence (vacuity) and uncertainty due to the presence of conflicting evidence (dissonance). A multinomial opinion, ω, can be expressed as $\omega = (\mathbf{b}, u, \mathbf{a})$, where $\mathbf{b}$ is a belief mass vector, the scalar u is the uncertainty mass and $\mathbf{a}$ is the base rate vector. For a K-class classification problem, $\mathbf{y}$, $\mathbf{a}$ and $\mathbf{b}$ are all vectors of dimension K. A projection of ω onto a probability distribution can be made according to

$$P(y = k) = b_k + a_k u. \tag{1}$$

It follows that since $\sum_{k=1}^{K} a_k = 1$ for the base rate vector, an additivity requirement is described by

$$u + \sum_{k=1}^{K} b_k = 1. \tag{2}$$

2.2. Dirichlet mapping

If $\mathbf{p}$ is a K-dimensional random vector containing the probability of belonging to each output class, and $\boldsymbol{\alpha}$ is the strength vector which parameterizes a Dirichlet distribution, the probability density function of the Dirichlet is given by

$$\mathrm{Dir}(\mathbf{p}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1} \alpha_k} \prod_{k=1}^{K} p_k^{\alpha_k - 1}, \tag{3}$$

where $\Gamma()$ is the gamma function. The distribution's expected value is given by

$$\mathbb{E}\left[\mathrm{Dir}(p_k|\boldsymbol{\alpha})\right] = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}. \tag{4}$$

If we allow the uncertainty mass and base rates to be given by

$$u = \frac{K}{\sum_{k=1}^{K} \alpha_k} = \frac{K}{S} \tag{5}$$

and

$$a_k = 1/K, \forall k \tag{6}$$

respectively, where S refers to the Dirichlet strength, then by equating the probability projection of (1) with the expected value of the Dirichlet distribution given by (4), the expression for the belief mass can be obtained as

$$b_k = \frac{\alpha_k - 1}{S}. \tag{7}$$

This provides us with everything needed in order to map from a Dirichlet distribution to a SL opinion and vice versa.

2.3. Uncertainty measures

From the definitions of the evidential uncertainties presented in [9], the measures of vacuity and dissonance have been adopted. The measure of vacuity uncertainty is simply given by the uncertainty mass, i.e.

$$vac(\omega) \equiv u = \frac{K}{S}, \tag{8}$$

and the measure of dissonance uncertainty is given by

$$diss(\omega) = \sum_{i=1}^{K} \left(\frac{b_i \sum_{j \neq i} b_j \mathrm{Bal}(b_j, b_i)}{\sum_{j \neq i} b_j} \right), \tag{9}$$

where $\mathrm{Bal}()$ is a function which gives the relative balance between two belief masses, defined by

$$\mathrm{Bal}(b_j, b_i) = \begin{cases} 1 - \frac{|b_i - b_j|}{b_i + b_j}, & \text{if } b_i + b_j \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

The entropy of the node-level multinomial distributions provided by the models is also computed to represent a form of uncertainty. This is done in order to provide a comparitive metric against which the evidential uncertainties can be compared.

3. Graph network architecture

The multimodal data can be represented using a graph, where each of the N nodes in the graph represents a pixel in the image. The graph's adjacency matrix, $\mathbf{A} \in \mathbb{R}^{N \times N}$,

is used to represent edges between nodes deemed similar. A set of features, $\mathbf{X} \in \mathbb{R}^{N \times C}$, is used to assign a vector description of each graph node, where C denotes the number of input features. The graph's degree matrix, $\mathbf{D} \in \mathbb{R}^{N \times N}$, is a diagonal matrix with elements given by $\mathbf{D}_{ii} = \sum_j \mathbf{A}_{ij}$.

The graph convolutional networks (GCNs) used are of the form proposed by [10], where the graph convolutional layer is given by

$$\mathbf{Z}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{Z}^{(l)}\boldsymbol{W}^{(l)}\right), \tag{11}$$

where $\mathbf{Z}^{(l)}$, $\mathbf{Z}^{(l+1)}$ and $\boldsymbol{W}^{(l)}$ are the inputs, outputs and weights of the l^{th} layer respectively, and $\sigma()$ is a non-linear activation function. For brevity, the tilde operator is used to represent the inclusion of self-connection edges in the graph, i.e. $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$ and $\tilde{\mathbf{D}} = \mathbf{D} + \mathbf{I}$.

3.1. Subjective models

An adaptation to the GCN architecture used by [10] must be made in order to obtain the subjective opinions that will be used to obtain measures of vacuity and dissonance uncertainty. The adaptation made means that the model will output node-level Dirichlet distribution parameters, such that the output will provide a probability distribution over multinomial class probabilities for each node. To do so, the softmax output activation function used in the output layer of the GCN is substituted for a ReLU function. In this way, the model is trained to output non-negative evidence contributions, $\mathbf{E} \in \mathbb{R}^{N \times K}$, where $\mathbf{E}_i = \boldsymbol{\alpha}_i - \mathbf{1}$ and $\boldsymbol{\alpha}_i$ refers to the K-dimensional concentration parameters of the i^{th} node. In order to train such a model, the loss function is made up of two components: a squared error term, which is minimized in order to classify a greater proportion of the nodes correctly, and a variance term, which is minimized to incentivize the model to provide confident predictions where possible. This loss, $\mathcal{L}(\boldsymbol{\theta})$, is given by

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= \sum_{i\in\mathbb{L}}\sum_k \left[(p_{ik} - y_{ik})^2 + \text{Var}(p_{ik})\right], \\ &= \sum_{i\in\mathbb{L}}\sum_k \left[(p_{ik} - y_{ik})^2 + \frac{\alpha_{ik}}{S_i^2}\left(\frac{S_i - \alpha_{ik}}{S_i - K}\right)\right],\end{aligned} \tag{12}$$

where $i \in \mathbb{L}$ refers to the fact that the loss is computed using a sum only over nodes in the training set, $\mathbb{L}$. Models trained with such an output activation and loss function will be denoted using the 'S-' prefix in order to indicate they provide subjective predictions, e.g. S-GCN.

3.2. Convergence assistance techniques

In order to assist the convergence of subjective models, two additional assistance techniques have been used: teacher knowledge distillation and the use of a Dirichlet prior. These have been shown to allow subjective models to provide better uncertainty estimates [11].

3.2.1. Teacher knowledge distillation

By training a non-subjective model in advance, its outputs, $\hat{p}_{ik}$, can be used in order to encourage the subjective model to converge to node Dirichlet distributions with $\mathbb{E}[p_{ik}]$ which are close to the teacher's deterministic estimates. This is achieved using an additional term in the loss function,

$$\mathcal{L}_{\text{T}}(\boldsymbol{\theta}) = \sum_i \sum_k \left(\hat{p}_{ik} \log \frac{\hat{p}_{ik}}{\mathbb{E}[p_{ik}]}\right), \tag{13}$$

which corresponds to the summation of Kullback-Leibler (KL) divergence terms between the teacher output probability and the expected value of the subjective model's Dirichlet distribution for each node. Using $D_{\text{KL}}(\,||\,)$ to compute the KL divergence, this is stated equivalently as $\sum_i D_{\text{KL}}(\hat{p}_{ik} \parallel \mathbb{E}[p_{ik}])$. Notice that this sum is computed over all nodes as opposed to just the nodes in $\mathbb{L}$. Models trained using a teacher are denoted using the '-T' suffix e.g. a S-BGCN-T model would indicate that a pre-trained GCN was used as a teacher in order to assist the training convergence of a subjective graph convolutional model.

3.2.2. Dirichlet prior

A second convergence assistance technique which can be used involves the use of a Dirichlet prior, $\hat{\boldsymbol{\alpha}}$. The exact method chosen to provide $\hat{\boldsymbol{\alpha}}$ will depend on the nature of the problem but we will assume nodes which are nearby in the graph are more likely to belong to the same output class than nodes which are far apart, a property known as homophily [13]. Using this assumption, we can use the computed distances on the graph to assign contributions of evidence from observed node labels to the other nodes in the graph using a function of our choosing. If d_{ij} denotes the shortest path distance between a given node, indexed by i and an observed node, indexed by j, then the amount of evidence contributed to suggest that the i^{th} node belongs to the k^{th} class is given by

$$h_{ik}(y_j, d_{ij}) = \begin{cases} \frac{\exp\left(\frac{-d_{ij}^2}{2\sigma^2}\right)}{\left(2\pi\sigma^2\right)^{1/2}}, & \text{if } y_{jk} = 1, \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

where σ is a scale parameter which controls the order of distance magnitude over which evidence will propagate in the prior. The total evidence to suggest the i^{th} node belongs to the k^{th} class, e_{ik} can be found by summing these contributions over the nodes in the training set, such that the element in the prior is given by

$$\hat{\boldsymbol{\alpha}}_{ik} = 1 + e_{ik} = 1 + \sum_{j\in\mathbb{L}} h_{ik}(y_j, d_{ij}). \tag{15}$$

Table 1
Loss function components and their weighting coefficients for different model types

Model name	$\mathcal{L}_{\text{total}}(\boldsymbol{\theta})$
S-BGCN	$\mathcal{L}(\boldsymbol{\theta})$
S-BGCN-T	$\mathcal{L}(\boldsymbol{\theta}) + \lambda_{\text{T}}\mathcal{L}_{\text{T}}(\boldsymbol{\theta})$
S-BGCN-K	$\mathcal{L}(\boldsymbol{\theta}) + \lambda_{\text{K}}\mathcal{L}_{\text{K}}(\boldsymbol{\theta})$
S-BGCN-T-K	$\mathcal{L}(\boldsymbol{\theta}) + \lambda_{\text{T}}\mathcal{L}_{\text{T}}(\boldsymbol{\theta}) + \lambda_{\text{K}}\mathcal{L}_{\text{K}}(\boldsymbol{\theta})$

The KL divergence between the Dirichlet distribution of the prior and the model output is given by the term

$$\mathcal{L}_{\text{K}}(\boldsymbol{\theta}) = \sum_i D_{\text{KL}}(\text{Dir}(\mathbf{p}_i|\boldsymbol{\alpha}_i) \parallel \text{Dir}(\hat{\mathbf{p}}_i|\hat{\boldsymbol{\alpha}}_i)), \quad (16)$$

which can, in turn, be incorporated into the total loss function. Models trained using a prior are denoted using the '-K' suffix.

Table 1 shows how these convergence assistance techniques can be weighted and combined in various permutations to provide a total loss function, $\mathcal{L}_{\text{total}}(\boldsymbol{\theta})$, as well as the model name abbreviations used to denote which combination has been used. The 'B' in the model names of Table 1 refers to the fact that dropout inference has been used as a Bayesian approximation. The coefficients λ_{T} and λ_{K} are used to control the relative importance of the teacher network and the Dirichlet prior respectively against the importance of the subjective loss function given in (12). These have been considered as hyperparameters which are to be tuned during training.

4. Results and analysis

4.1. Data

A subsection of the 2018 IEEE GRSS Data Fusion Challenge dataset [14] ws selected for the purposes of validating the described methods. The ground truth labels in this dataset describe 20 different urban land cover/land use classes (i.e. $K = 20$) as well as an unlabelled state, described as Unclassified. The modes of input data represent measurements from three sensor types: LiDAR, optical and hyperspectral (HS). The LiDAR data was provided at 0.5 m resolution, the same resolution as the ground truth labels (GT). In order to simplify analysis, the optical data (which was provided at 0.05 m resolution) and the HS data (which was provided at 1.0 m resolution) were bilinearly resampled to obtain 0.5 m resolution across inputs and outputs.

The graph was constructed with each 0.5 m $\times$ 0.5 m pixel representing a node in the graph. Each node has a 52-dimensional feature vector describing it (produced by stacking 3 optical channels, 48 HS channels and 1 LiDAR channel). The graph edges are computed using a k-nearest neighbors algorithm with two nodes receiving an edge connecting them if either node was one of the k nodes which were nearest the other. This produces a graph which is both undirected and unweighted. The graph, which contains approximately 2.16 million nodes, was computed with $k = 15$.

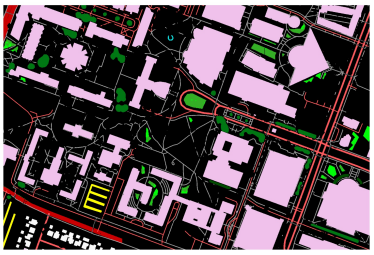

Figure 1: Ground truth data with colors depicting land cover classes. This data represents a subset of the 2018 IEEE GRSS Data Fusion Challenge dataset.

In order to measure an uncertainty output's ability to separate OOD nodes, a receiver operating characteristic (ROC) curve and a precision-recall (PR) curve can be computed. The area under the ROC curve and PR curve (AUROC and AUPR respectively) can be used as a single numerical representation of the detection performance, where an area of 1.0 would represent a perfect discriminator for both metrics.

4.2. Network training and hyperparameters

Models were implemented and trained using the TensorFlow library [15] on a personal laptop computer with Intel Core i7 CPU and 16 GB of RAM. In order to handle the imbalance of classes in the dataset, sample weighting was used. Samples were given weights which were inversely proportional to the number of total samples of each class in the training set. This allows the losses related to nodes from under-represented classes to have an increased influence over parameter updates and vice versa.

All GCN-based models were constructed using a dropout layer (dropout probability 0.5), a graph convolutional layer, as described in (11), a second dropout layer (dropout probability 0.5) and a second graph convolutional layer with the relevant output activation function. The kernel weights of the first graph convolutional layer were regularized using an L_2 penalization. Where dropout inference has been used, the number of samples taken was 100.

Hyperparameters including the learning rate, the L_2 regularization coefficient and the number of GCN layer output features, F, were selected via a grid-search

Table 2
OOD detection: Ability of each uncertainty type to detect OOD nodes (measured by the AUROC and AUPR metrics). Values shown represent the mean ± standard deviation.

Model	AUROC			AUPR		
	Vacuity	Dissonance	Entropy	Vacuity	Dissonance	Entropy
S-BGCN-T-K	$\mathbf{0.882 \pm 0.085}$	0.605 ± 0.197	0.878 ± 0.089	$\mathbf{0.318 \pm 0.289}$	0.132 ± 0.184	0.316 ± 0.306
S-BGCN-T	0.588 ± 0.147	0.664 ± 0.133	0.578 ± 0.186	0.128 ± 0.187	0.137 ± 0.192	0.143 ± 0.208
S-BGCN	0.586 ± 0.147	0.666 ± 0.132	0.580 ± 0.191	0.127 ± 0.186	0.139 ± 0.190	0.145 ± 0.209
S-GCN	0.580 ± 0.145	0.650 ± 0.120	0.586 ± 0.181	0.125 ± 0.185	0.130 ± 0.191	0.143 ± 0.207
S-MLP	0.767 ± 0.152	0.805 ± 0.114	0.787 ± 0.125	0.245 ± 0.214	0.233 ± 0.170	0.219 ± 0.201
GCN	-	-	0.538 ± 0.188	-	-	0.116 ± 0.179

method. Where used, λ_T and λ_K were also found using a grid-search.

Learning was performed for a maximum of 400 epochs, but was stopped early if the validation loss failed to decrease further for 60 consecutive epochs. If stopped early, model weights were returned to the settings which provided the lowest validation set loss upon the termination of training.

Each test was performed for different random dataset splits and model weight initializations to obtain mean and standard deviation measures of performance.

A benchmark has been provided by training 'standard' GCNs which provide prediction entropy as a form of uncertainty estimate.

4.3. Out of distribution detection

It would be reasonable to expect that uncertainty should be higher when the model is asked to make a prediction using an input which does not resemble the inputs upon which it was trained. The relative inability of neural networks to successfully extrapolate beyond the support of the training data is a well-known weakness of these methods [16]. By training models using only a subset of the classes provided by the GT, with the other classes acting as out of distribution (OOD) samples, the OOD detection ability of the uncertainty metrics can be measured. The AUROC and AUPR can be calculated for each uncertainty output provided by each model type, in order to determine the relative performance of the respective metrics for this task.

In the results presented, two classes were randomly selected to act as OOD. This was repeated 10 times, with two new randomly sampled classes selected for each training and evaluation loop in order that the variation in OOD detection performance due to the nature of the classes selected as OOD could be averaged out and the mean and standard deviation computed. Each model type was assessed over the same 10 sampled OOD class pairs for fairness. The AUROC and AUPR values measured can be found in Table 2.

For the task of OOD detection, the S-BGCN-T-K model is the highest ranked model. Its measure of vacuity uncertainty provided the best distinguishing metric, with mean AUROC and AUPR of 0.882 and 0.318 respectively, closely followed by performance from the measure of entropy (AUROC and AUPR of 0.878 and 0.316 respectively). The performance of the S-BGCN-T-K model stands out above the performance of other models trained. This highlights the importance of the convergence assistance techniques used, particularly the use of a meaningful prior.

The fact that vacuity is the uncertainty measure which best distinguishes OOD nodes reflects intuition. Since vacuity measures the absence of evidence for a prediction, it is natural to expect that it would better distinguish OOD nodes for which the model ought to have little evidence to support its classification.

5. Conclusion

In this paper we have adapted a novel classification method capable of providing uncertainty estimates to the task of multi-class classification of multimodal remote sensing data. The adopted framework, based upon the theory of Subjective Logic, provides measures of vacuity and dissonance uncertainty. Of the types of uncertainty assessed, the measure of vacuity was the best metric to perform identification of OOD samples. Experimental results have shown the performance of the S-BGCN-T-K model in the task of OOD detection to be improved against baseline methods. This represents a promising avenue for uncertainty-aware learning in the task of multimodal remote sensing classification.

The presented results illustrate the importance of convergence assistance techniques as a means for improving the quality of uncertainty estimates, particularly through the use of a prior. This can be seen by comparing the S-BGCN-T-K OOD detection performance with equivalent models which do not use a prior, e.g. S-BGCN-T.

Future work should consider the generalisation poten-

tial of this method by assessing performance on other challenging remote sensing classification datasets. The analysis could also be extended to assess whether the presented uncertainty measures could be used to detect model misclassifications. Additionally, there is scope for research into how the choice of method for computing the $\hat{\boldsymbol{\alpha}}$ prior affects the quality of uncertainty estimates, either by varying the scale parameter, σ, or considering different prior computation methods entirely.

Acknowledgments

This work is funded in part by Centre for Integrated Remote Sensing and Forecasting for Arctic Operations (CIRFA) and the Research Council of Norway (RCN Grant no. 237906), the Automatic Multisensor remote sensing for Sea Ice Characterization (AMUSIC) Framsenteret 'Polhavet' flagship project 2020, the Isaac Newton Trust, and Newnham College, Cambridge, UK.

References

[1] B. Goodman, S. Flaxman, European Union regulations on algorithmic decision-making and a 'right to explanation', AI Mag. 38 (2017) 50–57. doi:10.1609/aimag.v38i3.2741.

[2] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[3] L. Zhang, L. Zhang, B. Du, Deep learning for remote sensing data: A technical tutorial on the state of the art, IEEE Geosci. Remote Sens. Mag. 4 (2016) 22–40.

[4] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human Factors 46 (2004) 50–80. URL: https://doi.org/10.1518/hfes.46.1.50_30392. doi:10.1518/hfes.46.1.50_30392.

[5] M. Chi, A. Plaza, J. A. Benediktsson, Z. Sun, J. Shen, Y. Zhu, Big data for remote sensing: Challenges and opportunities, Proceedings of the IEEE 104 (2016) 2207–2219.

[6] S. Chlaily, M. D. Mura, J. Chanussot, C. Jutten, P. Gamba, A. Marinoni, Capacity and limits of multimodal remote sensing: Theoretical aspects and automatic information theory-based image selection, IEEE Trans. Geosci. Remote Sens. 59 (2021) 5598–5618. doi:10.1109/TGRS.2020.3014138.

[7] A. Marinoni, S. Chlaily, E. Khachatrian, T. Eltoft, S. Selvakumaran, M. Girolami, C. Jutten, Enhancing ensemble learning and transfer learning in multimodal data analysis by adaptive dimensionality reduction, CoRR abs/2105.03682 (2021). URL: https://arxiv.org/abs/2105.03682. arXiv:2105.03682.

[8] S. Chakraborty, et al., Interpretability of deep learning models: A survey of results, in: IEEE Smart World Congr. DAIS - Work. Distrib. Anal. Infrastruct. Algorithms Multi-Organization Fed., 2017, pp. 1–6. doi:10.1109/UIC-ATC.2017.8397411.

[9] A. Josang, J.-H. Cho, F. Chen, Uncertainty characteristics of subjective opinions, in: 2018 21st Int. Conf. Inf. Fusion, Cambridge, U.K., 2018, pp. 1998–2005.

[10] T. N. Kipf, M. Welling, Semi-supervised classification with graph convolutional networks, in: 5th Int. Conf. Learn. Representations, Toulon, France, 2017.

[11] X. Zhao, F. Chen, S. Hu, J.-H. Cho, Uncertainty aware semi-supervised learning on graph data, in: Advances Neural Inf. Process. Syst., volume 33, 2020, pp. 12827–12836.

[12] A. Jøsang, Subjective Logic - A Formalism for Reasoning Under Uncertainty, Artificial Intelligence: Foundations, Theory, and Algorithms, Springer, 2016. doi:10.1007/978-3-319-42337-1.

[13] Q. Huang, H. He, A. Singh, S.-N. Lim, A. Benson, Combining label propagation and simple models out-performs graph neural networks, in: 9th Int. Conf. Learn. Representations, 2021.

[14] S. Prasad, B. Le Saux, N. Yokoya, R. Hansch, 2018 IEEE GRSS Data Fusion Challenge - Fusion of Multispectral LiDAR and Hyperspectral Data, 2018. URL: https://dx.doi.org/10.21227/jnh9-nz89. doi:10.21227/jnh9-nz89.

[15] M. Abadi, et al., TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL: https://www.tensorflow.org/, Software available from tensorflow.org.

[16] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in Neural Inf. Process. Syst., volume 30, Long Beach, CA, USA, 2017, pp. 6402–6413.