LEARNING ONTOLOGY INFORMED REPRESENTATIONS WITH CONSTRAINTS FOR ACOUSTIC EVENT DETECTION

Akshay Raina, Sayeedul Islam Sheikh and Vipul Arora

Indian Institute of Technology Kanpur, India
{akshayr, sayeedul21, vipular}@iitk.ac.in

ABSTRACT

Acoustic Event Detection (AED) has been of great interest for nearly a decade for diverse applications. Most open datasets contain meta information on the hierarchy of labels, which can be utilized for building robust AED systems. Our study aims at injecting this domain knowledge by enforcing ontology-informed constraints upon the output space. We show that constrained optimization allows a network to confuse less among the child classes and can back off to parent classes when not confident enough. We perform several experiments on different datasets signifying the robustness of the method. The experiments substantiate that the state of the art baselines do not follow ontology constraints, and perform poorer than the proposed method.

Index Terms— Acoustic Event Detection, Constraint-based Learning, Disentangled Representations, Neuro-Symbolic AI

1. INTRODUCTION

One of the rapidly evolving areas in Artificial Intelligence (AI) has been making machine listening systems capable of reproducing the quality of human hearing towards recognizing events in sound recordings [1, 2, 3]. Acoustic Event Detection (AED) has proven to be beneficial to diverse applications, including surveillance [4, 5], smart homes [6], home security [7], recommendation systems [8] and more. Identifying ambiguous and overlapping events has always been a trivial task for human audition. They tend to categorize the sounds with abstract labels, which aid the interpretation and identification of the events. For example, two sound recordings with distinct event labels, say *violin* and *vocal*, can both be categorized into the same label *music*, which can be considered as a level higher in the abstraction or ontology of the labels. These semantic relations among labels, represented by ontologies [9] are well-defined in almost all publicly available datasets for acoustic event detection. However, many existing well-performing systems for the task rarely utilize this information. A formal representation of the structure of classes or their relations within a domain is considered as the *ontology* of that dataset.

Incorporating these hierarchical relations among the labels of a dataset into a deep-learning system allows for better performance, particularly in cases where there is ambiguity between two acoustically similar but semantically distinct events, and the classifier is expected to back off to more general categories, likely a level higher in the ontology.

Let $C_{z,k}$ denote the k^{th} label at z^{th} level in the ontology of a dataset $\mathcal{D} = \{(x^1, y^1), ..., (x^N, y^N)\}$. Here $y^i \in C_1 \times C_2 \times ... \times C_Z$ is the set of labels associated with the audio representation $x^i \in \mathcal{X}$, where Z is the total number of levels. For some level z, $C_z = \{C_{z,k}\}_{k=1}^{L_z}$ denotes the set of labels in z^{th} level in the ontology of $\mathcal{D}$, where L_z is the number of labels in z. Let each label in such C_z be mapped to one label in C_{z-1}, a level higher in the ontology. As also depicted in Fig. 1, for $z = 1$ we have $L_z = 4$ and $C_1 = \{C_{1,k}\}_{k=1}^{L_z=4}$ is the set of labels {*living things*, *mechanical*, *tools*, *street*}. Similarly, C_2 is {*dog bark*, *children playing*, *gun shot*, *air conditioner*, *engine idling*, *jackhammer*, *drilling*, *car horn*, *siren*, *street music* } and each element in C_2 is related to one element in C_1 in the ontology, eg. *dog bark* is related to *living things*. Also, let $\mathscr{C}(z, k)$ denote a set of children labels to the parent label at k^{th} position in the z^{th} level, for instance, $\mathscr{C}(1, 3) = \{$*jackhammer*, *drilling*$\}$.

Utilizing domain knowledge in the form of label-hierarchy has been proven significantly helpful for several tasks, including named entity recognition [10], textual topic selection [11] and more. However, most existing approaches for AED do not exploit these hierarchical relations [1, 2, 3, 12]. Though, there has been light evidence that utilizing the ontology of a dataset leads to more robust systems.

Jiménez et al. [13] used a Siamese Neural Network to generate ontology-preserving embeddings. They enforced scores, for example, being in the same super-class, sub-class, etc., and added a term to the loss function, allowing the network to transform two inputs such that their embeddings are further away if the score is higher. Sun et al. [14] proposed a 2 component ontology aware neural network. The first component, the feed-forward ontology layers, allowed the prediction of one level in the hierarchy using embeddings from other levels, thereby capturing the intra-dependencies

This work was supported by PB/EE/2021128B grant from Prasar Bharati.

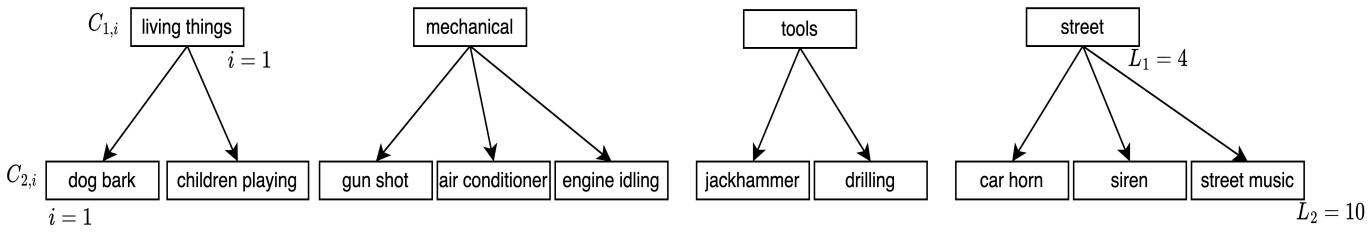

Fig. 1. Ontology of the UrbanSound8K dataset

of labels between different levels of ontology. The second component, graph CNN, was applied to capture the inter-dependency structure of labels within an ontology level. Liu et al. [15] proposed a novel loss function called OBCE loss, which reweights the Binary Cross Entropy function based on the depth of the ontology. Their experiments on Audioset showed considerable improvements in audio event detection task using the OBCE loss. Zharmagambetov et al.[16] proposed a joint model consisting of a representation neural network and a decision tree model based on pre-defined tree-structured ontology for a semi-supervised framework for acoustic scene classification.

An AED system is expected to categorize an input audio recording into one or more event labels in the lowest level of ontology or the finest available labels. However, since different acoustic scenes occur with different environmental conditions in nature, it is likely for a method to confuse among classes in such cases, which may lead to unavoidable false positives. For instance, based on acoustic environmental conditions, a cat meow may perceptually be similar to a baby crying. If knowledge from label-structure of a dataset, or ontology is incorporated into a learning algorithm, it can be expected to not confuse among classes of at least distinct parents in the hierarchy. In this work, we aim at a similar objective using constraint-based learning. We formulate constraints using the relation between the child classes and their parent class, and impose these on the neural network which is thus encouraged to align with the structure of labels.

2. PROPOSED METHODOLOGY

2.1. Constrained Learning

Let a neural network g_θ with parameters θ be trained by minimizing an average loss $\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} l(g_\theta(x^i), y^i)$, where l can be any standard loss function. The goal is to find a set of parameters θ^*, such that $\theta^* = \arg\min_\theta \mathscr{L}(\theta)$. In this work, we aim at injecting domain knowledge in the form of hard constraints restricting the output label space, to achieve a more robust and generalized model.

For label structure as in Fig 1 and some input x^i, let $\hat{y}^z$ be a set of predicted labels for each level z. For $Z = 2$, we can deduce the relation between network predictions for both levels as-

$$\begin{aligned} p(y^1 = \hat{y}_k^1 \mid x^i) &= \sum_{y^2 \in C_2} p(y^1 = \hat{y}_k^1,\ y^2 \mid x^i) \\ &= \sum_{y^2 \in C_2} p(y^1 = \hat{y}_k^1 \mid y^2, x^i)\ .\ p(y^2 \mid x^i) \\ &= \sum_{y^2 \in C_2} p(y^1 = \hat{y}_k^1 \mid y^2)\ .\ p(y^2 \mid x^i) \\ &= \sum_{y^2 \in \mathscr{C}(1,k)} p(y^2 \mid x^i) \end{aligned} \tag{1}$$

Clearly, the ontology is a human-defined label structure for a known dataset. Therefore in nature, one may expect a higher number of child nodes to a label in the ontology. As an example, it is possible that for the hierarchy in Fig 1, $|\mathscr{C}(1,3)| > 3$ in nature. Hence, the equality in Eq 1 for a parent node is valid only when all possible child nodes are included, which may be unrealistic.

Hence, a more general form of Eq 1 would be-

$$p(y^1 = \hat{y}_k^1 \mid x^i) \geq \sum_{y^2 \in \mathscr{C}(1,k)} p(y^2 \mid x^i) \tag{2}$$

Let $\mathfrak{C}_k^i : f_k^i \geq 0$ where $f_k^i = p(y^1 = \hat{y}_k^1 \mid x^i) - \sum_{y^2 \in \mathscr{C}(1,k)} p(y^2 \mid x^i)$ be the k^{th} inequality constraint for the i^{th} example. Here for each example, there are $K = \sum_{z=1}^{Z-1} L_z$ constraints.

The constrained optimisation problem is formulated as-

$$\theta^* = \arg\min_\theta \mathscr{L}(\theta) \ \text{ subject to } \ f_k^i(\theta) \geq 0\ \ \forall\ \ i, k \tag{3}$$

In order to reduce the number of constraints $N * K$ where N is the dataset size, we take inspiration from the trick used by Nandwani et al. [2] and modify the optimisation problem using a Hinge function. We define $H : \mathbb{R} \to \mathbb{R}$ as $H(c) = -c$ for $c < 0,$ and 0 for $c \geq 0$. Now each constraint $f_k^i(\theta) \geq 0$ can be equivlently replaced by $H(f_k^i(\theta)) = 0$. The modified optimisation problem can thus be formulated as-

$$\theta^* = \arg\min_\theta \mathscr{L}(\theta) \ \text{ subject to } \ H(f_k^i(\theta)) = 0\ \ \forall\ \ i, k \tag{4}$$

By definition, $H(c) \geq\ 0\ \forall\ c$. Hence, we can enforce the constraint $H(f_k^i(\theta)) = 0$ by the condition $\sum_i H(f_k^i(\theta)) =$

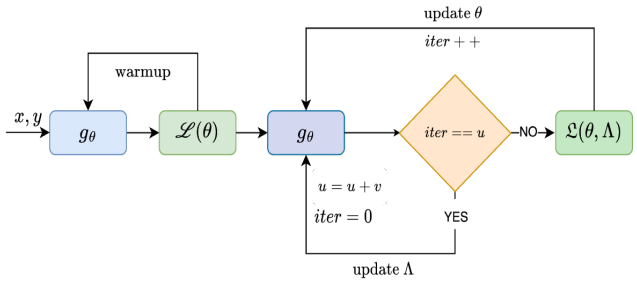

Fig. 2. Training Algorithm of the proposed method.

0. This formulates the following objective, where $h_k(\theta) = \sum_i H(f_k^i(\theta))$.

$$\theta^* = \arg\min_\theta \mathscr{L}(\theta) \text{ subject to } h_k(\theta) = 0 \ \forall \ k \quad (5)$$

Clearly, the number of constraints are now reduced by a factor of N.

2.2. Optimisation

It can be trivially noted that enforcing the constraint $f_k^i(\theta) \geq 0$ on the network may lead towards reducing the predicted probability of the child classes relative to their parent. This can be an inherent cause for confusion among the classes in the finest level of the heirarchy. Therefore, we formulate the $\mathscr{L}(\theta)$ as a weighted sum of the binary cross-entropy and the categorical cross-entropy, where w_1 and w_2 are the scalar weights-

$$\mathscr{L}(.) \stackrel{\text{def}}{=} w_1 \mathscr{L}_{bce}(.) + w_2 \mathscr{L}_{cce}(.) \quad (6)$$

Note that $\mathscr{L}_{bce}$ is applied onto the predictions for all classes, where as $\mathscr{L}_{cce}$ is applied only on the predictions for the children classes. This ensures that the relative confidence of the network between the parent and the children classes is not too high.

The K constraints $\{\mathfrak{C}_1^i(\theta), \mathfrak{C}_2^i(\theta), ..., \mathfrak{C}_K^i(\theta)\}$ are on the predictions for the i^{th} input example and hence a function of the network parameters. The optimisation problem in Eq 5 can be solved by finding a stationary point of the corresponding Lagrangian, where $\Lambda = \{\lambda\}_{k=1}^K$ denotes the K-sized vector of Lagrange multipliers, as-

$$\mathfrak{L}(\theta; \Lambda) = \mathscr{L}(\theta) + \sum_{k=1}^{K} \lambda_k h_k(\theta) \quad (7)$$

The primal form of Eq 7 can be realized as-

$$\min_\theta \ \max_\Lambda \ \mathfrak{L}(\theta, \Lambda) \quad (8)$$

The corresponding dual to be solved can be written as-

$$\max_\Lambda \ \min_\theta \ \mathfrak{L}(\theta, \Lambda) \quad (9)$$

Table 1. F1 scores for different methods on both levels for UrbanSound8K and FSD50K

Dataset	Method	Level 1	Level 2
UrbanSound8K	Baseline-NC	85.7	82.2
	Baseline-CI	84.8	79.7
	Jimenez et al. [13]	87.3	84.8
	Sun et al. [14]	88.0	88.3
	Ours	**88.9**	**88.5**
FSD50K	Baseline-NC	76.58	75.92
	Baseline-CI	70.02	68.11
	Ours	**78.19**	**77.91**

The dual optimisation problem is solved by alternating the gradient descent steps over θ and Λ. Jin et al. [17] presented that the alternating gradient ascent (descent) converges to the local min-max point under several conditions. The training algorithm (Fig. 2) also follows the same.

3. EVALUATION

3.1. Dataset Used

We evaluate the proposed methodology on a variety of databases including- UrbanSound8K (US8K) [18], Free Sound Database 50K (FSD50K) [19]. UrbanSound8K consists of 8732 labelled sound excerpts each with a duration $\leq$ 4s, distributed across 10 classes. For the ontology of the dataset at level $z = 2$, we utilize the existing 10 classes and follow [13] to deduce 4 classes for $z = 1$: {*living_things, mechanical, street, tools*}. All recordings have a single channel 44.1kHz, 16-bit .wav format. The data has been divided into 10 folds and we employ the 10-fold cross-validation approach to assess our methodology's performance.

FSD50K is a collection of 51,197 human-labeled audio clips each with duration in $[0.3s, 30s]$. The dataset has a total of 200 classes drawn from Audioset [20] ontology, with the total duration being over 100 hours. To allow for the labels in lowest level to be perceptually distinct, we prune out the ontology resulting into a 2-level heirarchy of labels for the dataset. We take a subset of the FSD50K dataset with a total of 14778 samples spanning 29 classes. The parent level comprises of 5 classes each with at least 4 children classes. The level $z = 2$ comprises of 24 classes in total.

3.2. Experimental Setup

Preprocessing: All audio recordings are resampled to 32 kHz. The input size (d) for the US8K and FSD50K is set to be 4s and 10s respectively and necessary zero-padding is added to examples where required. For FSD50K, the the extracted features are chunked down to 10s in case $d - \epsilon > 10$, where ϵ is some threshold, and each chunk is stacked along the *batch* dimension. We extract Mel Spectrograms for each input with 128 mel filterbanks, hop length of 400 frames, window size

Table 2. F1 scores, number of constraints violations (#CV) for different methods for UrbanSound8K and FSD50K

Dataset	Method	Overall F1	#CV
UrbanSound8K	Baseline-NC	81.16	1173
	Baseline-CI	67.41	1293
	Ours	**82.80**	**45**
FSD50K	Baseline-NC	74.18	2219
	Baseline-CI	63.91	2496
	Ours	**76.82**	**122**

of 1024 frames, and cut-off frequencies of 50 Hz to 14 kHz. US8K has an imbalanced distribution of classes, thus we used augmentation to upsample the less-frequent classes. During training, we also used the SpecAugment using torchlibrosa as also used in PANNs [21]. We also used class-weights inversely proportional to the number of samples of that class to tackle the problem of class-imbalance.

Network Architecture: The network consists of 3 CNN blocks, followed by a Fully Connected block. Each CNN block consists of 2 Conv2D layers followed by BatchNorm2D and ReLU activation. The kernel size is set as $(3, 3)$ with unity stride and padding. The Fully Connected consists of 3 Linear Layers with the number of neurons in the final layer being the number of nodes in the ontology of the dataset, i.e. $L_1 + L_2$. After every CNN block, a maxpool of (2,2) is applied followed by a dropout layer.

3.3. Results

We conduct several experiments to prove the robustness of our method. As also depicted in Fig. 2, we first train the network for 10 warm-up epochs with $\Lambda = \{0\}_K^{k=1}$. Then θ is updated u times for every Λ update, where u follows an arithmetic progressionm with step size v. This is a critical convergence criteria for a min-max optimisation problem solved using alternating gradient ascent (descent) as presented by Jin et al. [17] and used by Nandwani et al. [10]. The learning rates for both θ and Λ updates are also updated using OneCycleLr [22] and BetaScheduler respectively. This is followed for 100 epochs after the warm-up phase. For baseline we first train the same CNN in similar settings, except that no constraints are used and θ is updated by only using $\mathscr{L}_{bce}$ over the predictions and call this *Baseline-NC*. We prepare another baseline system *Baseline-CI* same as the previous one, except that it is trained only for the classes in the finest level with $\mathscr{L}_{cce}$ loss function. The predicted class in the parent level is decided using Constrained Inference, i.e. given the heirarchy and a predicted label, the corresponding label prediction at parent level can be inferred directly by implication, in the ontology.

We tabulate the F1 scores for both levels for both baselines, two existing works [13, 14] and our method on both datasets in Table 1. Clearly our method outperforms all others with significant margins for both levels. Also note that the baseline with the constrained inference (*Baseline-CI*) performs poor relative to the baseline and our method without any constraints. It indicates that imparting constraints while inference does not aid significantly to the performance in contrast to inject this knowledge while training itself. Table 2 shows the overall F1 score and the number of constraints violated (#CV) obtained by both baselines and our method considering this a multilabel-classification setting. Table 3 further adds to the claim that our method is able to disambiguate among child classes at least of distinct parents by making less false predictions and shows the AUPRC score for the USound8K dataset.

Table 3. AUPRC for different methods on both levels for UrbanSound8K

Method	Level 1	Level 2
Baseline-NC	0.75	0.65
Baseline-CI	0.64	0.58
Ours	**0.81**	**0.72**

The proposed method can also be extended to a semi-supervised case. Note that the second term in Eq. 7 does not require the target values and aims at learning the heirarchy only. Hence, we can compute this term even for unlabeled examples, which has huge potential particularly in cases with labeled data scarcity [10, 23].

4. CONCLUSION

In this paper, we attempt at building a robust system for AED with a focus on leveraging domain knowledge using ontology-informed constraints. We formulate the constraints using the hierarchical relation between parent and child classes in the ontology of the dataset and use constrained optimisation to train a deep neural network. This allows the network to exhibit a more structured understanding of acoustic events which has potential for improved interpretability and performance. The results indicate that the proposed methods outperform the baselines and existing works with significant margins and violates much lesser constraints on the validation set. Such adaptability aligns our system more closely with real-world applications of AED.

5. REFERENCES

[1] Sangwook Park and Mounya Elhilali, "Time-balanced focal loss for audio event detection," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 311–315.

[2] Vipul Arora, Ming Sun, and Chao Wang, "Deep embeddings for rare audio event detection with imbalanced data," in *ICASSP 2019-2019 IEEE International Con-*

ference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 3297–3301.

[3] Wuyang Liu, Yanzhen Ren, and Jingru Wang, "Attention mixup: An accurate mixup scheme based on interpretable attention mechanism for multi-label audio classification," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[4] Marco Cristani, Manuele Bicego, and Vittorio Murino, "Audio-visual event recognition in surveillance video sequences," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 257–267, 2007.

[5] Giuseppe Valenzise and et al., "Scream and gunshot detection and localization for audio-surveillance systems," in *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. IEEE, 2007, pp. 21–26.

[6] Andrey Temko, Robert Malkin, Christian Zieger, Dusan Macho, Climent Nadeu, and Maurizio Omologo, "Acoustic event detection and classification in smart-room environments: Evaluation of chil project systems," *Cough*, vol. 65, no. 48, pp. 5, 2006.

[7] Danish Chowdhry, Raman Paranjape, and Paul Laforge, "Smart home automation system for intrusion detection," in *2015 IEEE 14th Canadian Workshop on Information Theory (CWIT)*. IEEE, 2015, pp. 75–78.

[8] Pedro Cano, Markus Koppenberger, and Nicolas Wack, "Content-based music audio recommendation," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 211–212.

[9] Balakrishnan Chandrasekaran, John R Josephson, and V Richard Benjamins, "What are ontologies, and why do we need them?," *IEEE Intelligent Systems and their applications*, vol. 14, no. 1, pp. 20–26, 1999.

[10] Yatin Nandwani, Abhishek Pathak, and Parag Singla, "A primal dual formulation for deep learning with constraints," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[11] Hao Wang, Dejing Dou, and Daniel Lowd, "Ontology-based deep restricted boltzmann machine," in *Database and Expert Systems Applications*, Sven Hartmann and Hui Ma, Eds., Cham, 2016, pp. 431–445, Springer International Publishing.

[12] Sangwook Park and et al., "Self-training for sound event detection in audio mixtures," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021.

[13] Benjamin Elizalde, Abelino Jimenez, and Bhiksha Raj, "Sound event classification using ontology-based neural networks," in *NIPS 2018 Workshop*, 2018.

[14] Yiwei Sun and Shabnam Ghaffarzadegan, "An ontology-aware framework for audio event classification," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 321–325.

[15] Haohe Liu, Qiuqiang Kong, Xubo Liu, Xinhao Mei, Wenwu Wang, and Mark D Plumbley, "Ontology-aware learning and evaluation for audio tagging," *arXiv preprint arXiv:2211.12195*, 2022.

[16] Arman Zharmagambetov, , Qingming Tang, Chieh-Chi Kao, Qin Zhang, Ming Sun, Viktor Rozgic, Jasha Droppo, and Chao Wang, "Improved representation learning for acoustic event classification using tree-structured ontology," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 321–325.

[17] Chi Jin, Praneeth Netrapalli, and Michael I Jordan, "Minmax optimization: Stable limit points of gradient descent ascent are locally optimal," *arXiv preprint arXiv:1902.00618*, 2019.

[18] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *22nd ACM International Conference on Multimedia (ACM-MM'14)*, Orlando, FL, USA, Nov. 2014, pp. 1041–1044.

[19] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[20] Jort F Gemmeke and et al., "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017.

[21] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[22] Leslie N Smith and Nicholay Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*. SPIE, 2019, vol. 11006, pp. 369–386.

[23] Jingyi Xu, Zilu Zhang, Tal Friedman, Yitao Liang, and Guy Van den Broeck, "A semantic loss function for deep learning with symbolic knowledge," in *Proceedings of the 35th International Conference on Machine Learning*. 10–15 Jul 2018, vol. 80 of *Proceedings of Machine Learning Research*, pp. 5502–5511, PMLR.