Redes de Tensores Lógicos para Interpretação Semântica de Imagens de Veículos

Lucas Martinuzzo Batista
Graduação em Engenharia da Computação
Universidade Federal do Espírito Santo (UFES)
Vitória-ES, Brasil
lcmartinuzzo@gmail.com

Pedro Reisen Zanotti
Graduação em Engenharia da Computação
Universidade Federal do Espírito Santo (UFES)
Vitória-ES, Brasil
pedro.reisen15@gmail.com

***Resumo*—Este documento se trata de um relatório cuja finalidade é avaliar o desempenho das Redes de Tensores Lógicos desenvolvidas por Donadello et al. [1] em duas tarefas de Interpretação Semântica de Imagens, mais precisamente na classificação de *bounding boxes* de imagens de veículos e na detecção de relações *part-of* relevantes entre os objetos presentes nelas. O experimento descrito neste relatório mostra que o uso de *background knowledge* na forma de restrições lógicas pode melhorar a performance de abordagens puramente orientadas a dados, tal como a *Fast* R-CNN [2]. Além disso, ele pode aumentar a robustez do sistema de aprendizado nos casos em que existem erros nos *labels* dos dados de treinamento.**

Palavras-chave—LTN; Fast R-CNN; bounding box; part-of; background knowledge; labels; características; treinamento; teste

I. INTRODUÇÃO

O processo de Interpretação Semântica de Imagens (*Semantic Image Interpretation* – SII) refere-se, de maneira geral, à extração de descrições semânticas estruturadas de imagens. Uma das abordagens para a solução de problemas de SII utiliza uma subdisciplina da Inteligência Artificial e do Aprendizado de Máquina denominada Aprendizado Estatístico Relacional (*Statistical Relational Learning* – SRL), a qual lida com modelos de domínio que apresentam incertezas e com estruturas relacionais complexas. As Redes de Tensores Lógicos (*Logic Tensor Networks* – LTNs) são uma estratégia de SRL que integra redes neurais artificiais com lógica difusa de primeira ordem para permitir aprendizado eficiente a partir de dados ruidosos na presença de restrições lógicas e raciocínio com fórmulas lógicas que descrevem propriedades gerais dos mesmos [3].

As LTNs combinam características visuais e conhecimento simbólico na forma de axiomas lógicos para solucionar dois problemas de SII [4]: a classificação de *bounding boxes* de imagens e a identificação de relações *part-of* entre os objetos que a compõem. Dessa forma, elas visam compensar a falta de correspondência entre características de baixo nível (numéricas) que podem ser observadas em uma imagem e descrições semânticas de alto nível associadas aos objetos presentes nela (*gap* semântico) através da utilização de *background knowledge*.

Neste experimento, as LTNs foram aplicadas sobre um subconjunto de imagens de veículos proveniente do conjunto de dados PASCAL-Part. Dados do Laboratório de Computação de Alto Desempenho (LCAD) não foram utilizados pelo simples fato de que os dados de entrada das LTNs precisam de um pré-processamento de origem desconhecida, o qual não é mencionado no artigo de Donadello et al. [1].

Este relatório está organizado da seguinte maneira: a Seção II menciona outros trabalhos que utilizaram *background knowledge* em tarefas de SII e elucida as diferenças dos mesmos em relação a este. A Seção III discute características de implementação e alguns formalismos das LTNs. A Seção IV descreve em detalhes os experimentos realizados e as técnicas de avaliação utilizadas para o modelo. Por fim, a Seção V apresenta e discute os resultados obtidos nos experimentos.

II. TRABALHOS CORRELATOS

Em seu artigo, Donadello et al. [1] cita outras abordagens que exploraram a ideia de utilizar *background knowledge* para resolver problemas de SII. Essencialmente, ele cita três vertentes dessas abordagens: as baseadas em lógica (*Description Logics* (DL) [4], bases de conhecimento), as baseadas em modelos gráficos probabilísticos (*Markov Logic Networks* (MLNs), *Conditional Random Fields* (CRFs)) e as baseadas em modelos linguísticos. Entretanto, todas elas possuem limitações em relação à capacidade das LTNs, tais como: falta de definições formais, inconsistências e domínios simples ou pouco expressivos, restritos a um conjunto de relações menos complexas que a relação de *part-of*.

III. METODOLOGIA

Para tarefas de SII, Donadello et al. [1] considera uma linguagem de lógica de primeira ordem cuja assinatura é $\Sigma_{SII} = <C, P, F>$, em que $C = U_{p \in \text{Imagens}} b(p)$ é o conjunto de identificadores das *bounding boxes* de todas as imagens (conjunto de símbolos de constantes), $F = \emptyset$ (conjunto de símbolos funcionais) e $P = \{P_1, P_2\}$ (conjunto de símbolos de predicados). P_1 é o conjunto de predicados unários (por exemplo, $P_1 = \{\text{Bicicleta, Carro, Motocicleta, Ônibus, Trem}\}$ e $P_2 = \{\text{partOf}\}$.

A partir de Σ_{SII}, é possível descrever fórmulas lógicas que podem ser interpretadas pela lógica difusa para lidar com exceções. Cada objeto no conjunto de interpretação está associado a um vetor *n*-dimensional de números reais. De fato, cada constante *b* que denota uma *bounding box* está associada a um conjunto de características geométricas que descrevem a posição e a dimensão dessa *bounding box* e a um conjunto de características semânticas que descrevem o *score* de classificação retornado pelo detector de *bounding boxes* para cada classe de objetos existente. Para uma *bounding box* $b \in C$ e uma classe de objetos $C_i \in P_1$, a interpretação ou *grounding* de *b* é descrita como $G(b) = \langle class(C_i, b), ..., class(C_{|P1|}, b), x_0(b), y_0(b), x_1(b), y_1(b)\rangle$, em que os últimos elementos são as coordenadas do vértice superior-esquerdo e do vértice inferior-direito de *b* e $class(C_i, b) \in [0,1]$ é o *score* de classificação de *b* em relação à classe de objetos C_i.

A partir da definição acima, Donadello et al. [1] estabelece expressões para o *grounding* de predicados, os quais incluem predicados unários e a relação partOf. Com isso, é possível avaliar o grau de confiança de qualquer fórmula atômica descrita a partir da linguagem Σ_{SII}.

Um conjunto de treinamento pode ser representado por uma *grounded theory* $T_{expl} = \langle K_{expl}, \hat{G}\rangle$, em que K_{expl} é um conjunto de literais fechados do tipo $C_i(b)$ ou $partOf(b, b')$ para cada *bounding box b* rotulada com C_i e para cada par de *bounding boxes* $\langle b,b'\rangle$ conectados pela relação *part-of*. O conjunto $\hat{G}$ é um *grounding* parcial, ou seja, um *grounding* definido para um subconjunto de Σ_{SII}, especificamente para todas as *bounding boxes* de todas as imagens em que tanto as características semânticas $class(C_i, b)$ quanto as coordenadas dessas *bounding boxes* são computadas por um detector de objetos *Fast* R-CNN [2].

Como $\hat{G}$ não está definido para os símbolos de predicado em P, as informações assertivas contidas em T_{expl} a respeito de *bounding boxes* específicas serão utilizadas pelos classificadores para que eles aprendam indutivamente a partir de exemplos positivos e negativos. É possível adicionar um conjunto de axiomas mereológicos M em K_{expl} como uma forma de *background knowledge*, de modo que se defina uma nova *grounded theory* $T_{prior} = \langle K_{prior}, \hat{G}\rangle$, em que $K_{prior} = K_{expl} + M$.

IV. Experimentos

Avaliou-se a performance da abordagem discutida no artigo de Donadello et al. [1] para as duas tarefas de SII discutidas anteriormente: a classificação de *bounding boxes* e a detecção de relações *part-of* entre elas. Em especial, esta tarefa é importante pelo fato de que a relação de *part-of* pode ser usada para representar, através de reificação, uma classe maior de relações. Outras relações poderiam ter sido incluídas nessa avaliação, porém a complexidade de tempo da LTN cresce linearmente com o aumento do número de axiomas lógicos. Também se avaliou a robustez da abordagem em relação a dados de entrada ruidosos (dados com *labels* inexistentes ou errôneos, com discordâncias entre esses labels ou com objetos não localizados, por exemplo).

O conjunto de dados escolhido para treinamento e teste das LTNs foi o PASCAL-Part, que contém 10103 imagens com *bounding boxes* anotadas com tipos de objetos e com relações *part-of* definidas entre pares de *bounding boxes*. Os *labels* desse conjunto pertencem a três principais grupos: animais, veículos, e objetos de interior. É importante mencionar que objetos inteiros dentro de um mesmo grupo podem compartilhar partes entre si, enquanto objetos inteiros de grupos distintos não podem fazê-lo. Como os *labels* originais eram muito específicos, definiu-se um novo conjunto de *labels* mais genéricos para facilitar a atividade de aprendizado. Neste experimento, utilizou-se somente o grupo de veículos, o qual contém 9 *labels* para objetos inteiros (*aeroplane*, *bicycle*, *bus*, *car*, *motorbike*, *train*, *coach*, *locomotive*, *boat*) e 14 *labels* para partes de objetos (*artifact_wing*, *body*, *engine*, *stern*, *wheel*, *chain_wheel*, *handlebar*, *headlight*, *saddle*, *bodywork*, *door*, *license_plate*, *mirror*, *window*). Utilizou-se 80% dos dados de entrada para treinamento e 20% para teste.

A. Classificação de Tipos de Objetos e Detecção de Relações Part-Of

A partir de um conjunto de *bounding boxes* gerado por um detector de objetos *Fast* R-CNN [2], analisou-se o desempenho de duas LTNs distintas, uma treinada apenas com exemplos (T_{expl}) e outra treinada com o auxílio de *background knowledge* (T_{prior}) extraído de uma ontologia mereológica (WordNet). As LTNs utilizaram tensores de $k = 6$ camadas e um parâmetro de regularização de $\lambda = 10^{-10}$. Escolheu-se a T-norma de Lukasiewicz para definir a semântica de conectivos lógicos e a média harmônica foi usada como operador de agregação. Foram executadas 1000 *training epochs* do algoritmo RMS*Prop* disponível no TensorFlow™. Em seguida, os resultados foram comparados com a performance da *Fast* R-CNN para a classificação de tipos de objetos e com a *baseline* da relação de inclusão $ir = area(b \cap b')/area(b)$ para a tarefa de detecção de relações *part-of*. Se $ir \geq 0{,}7$, pode-se afirmar que as *bounding boxes* pertencem a uma relação *part-of*.

Figura 1 – Exemplos de imagens do conjunto de dados PASCAL-Part.

Cada *bounding box b* é classificada em $C_i \in P_1$ se $G(C_i(b)) \geq 0{,}7$. Portanto, uma *bounding box* pode ser classificada em mais de uma classe. Para cada classe, a precisão e a revocação foram calculadas da maneira usual.

B. *Robustez em Conjuntos de Treinamento Ruidosos*

Para avaliar a robustez do modelo e medir o seu desempenho na presença ou ausência de axiomas lógicos, selecionou-se aleatoriamente *k*% das *bounding boxes* do conjunto de treinamento e alterou-se aleatoriamente os seus *labels* de classificação, com $k = \{10, 20, 30, 40\}$. De maneira similar, selecionou-se aleatoriamente *k*% dos pares de *bounding boxes* e alterou-se o valor do *label* das suas relações *part-of*. Para cada valor de K, treinou-se as LTNs T^k_{expl} e T^k_{prior} e avaliou-se os resultados nos dois problemas de SII mencionados anteriormente.

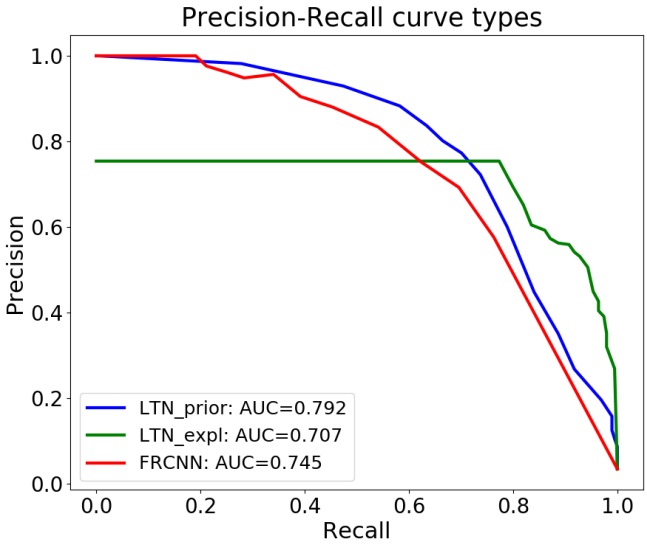
V. RESULTADOS

A. *Classificação de Tipos de Objetos e Detecção de Relações Part-Of*

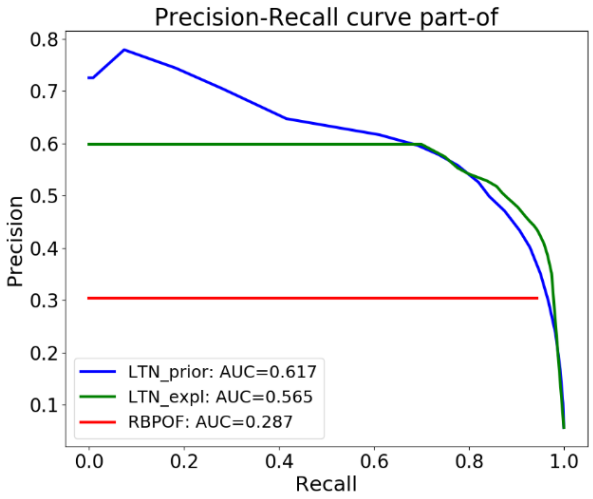
Os resultados estão indicados na Fig. 2, na qual AUC é a área sobre a curva de precisão-revocação. Eles mostram que, tanto para os tipos de objetos quanto para as relações *part-of*, a LTN treinada com conhecimento prévio representado através de axiomas mereológicos obteve melhor performance que a LTN treinada somente com exemplos. Além disso, o conhecimento prévio permitiu à LTN melhorar o desempenho do detector de objetos *Fast* R-CNN [2]. Diferentemente da *Fast* R-CNN, as LTNs fazem escolhas globais que levam em consideração todas as características semânticas e geométricas dos dados juntas. Isso oferece robustez ao classificador LTN ao custo de uma queda na precisão. Entretanto, os axiomas lógicos compensam essa perda.

B. *Classificação de Tipos de Objetos e Detecção de Relações Part-Of*

Como esperado, verificou-se que adicionar ruído nos *labels* de treinamento implica em uma queda de performance. Cada par de barras na Fig. 3 indica a AUC de T^k_{expl} e T^k_{prior} para um percentual *k*% de erros. Entretanto, os resultados indicaram que os axiomas LTN oferecem robustez ao ruído: apesar da queda geral no desempenho, observa-se uma diferença crescente entre a queda de performance da LTN treinada somente com exemplos e a queda de performance da LTN que incluiu *background knowledge*.

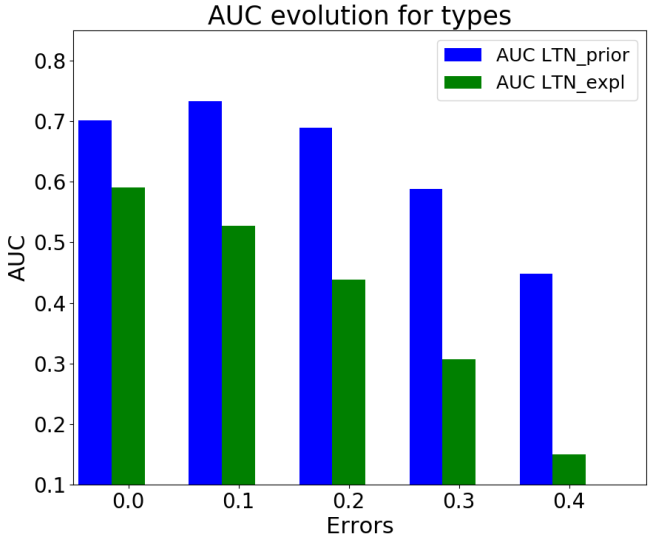


(a) LTNs com conhecimento prévio melhoram a performance da *Fast* R-CNN na classificação de tipos de objetos, alcançando uma AUC de 0,792, em comparação com 0,745.

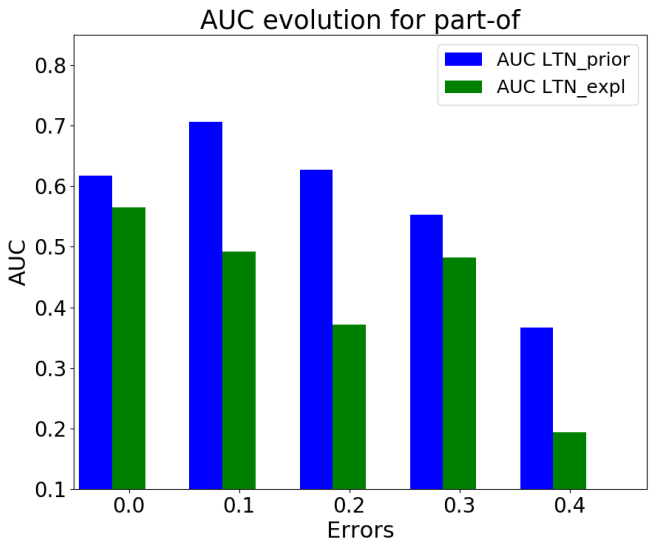


(b) LTNs com conhecimento prévio superam a abordagem baseada em regras, alcançando uma AUC de 0,617, em comparação com 0,287.

Figura 2 – Curvas de precisão-revocação para a classificação de tipos de objetos e para a relação *part-of* entre os objetos.



(a)



(b)

Figura 3: AUC para os tipos de objetos e para a relação *part-of* com ruído crescente nos *labels* dos dados de treinamento. A queda na performance é notavelmente menor para a LTN treinada com *background knowledge*.

BIBLIOGRAFIA

[1] I. Donadello, L. Serafini e A. S. d'Avila Garcez, "Logic Tensor Networks for Semantic Image Interpretation", em Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI-17), 2017.

[2] R. Girshick, "Fast R-CNN", na International Conference on Computer Vision (ICCV), 2015.

[3] L. Serafini e A. S. d'Avila Garcez, "Leaning and reasoning with logic tensor networks", em Proc. AI*IA, páginas 334-348, 2016.

[4] I. Donadello e L. Serafini, "Integration of numeric and symbolic information for semantic image interpretation", Intelligenza Artificiale, 10(1):33-47, 2016.