Detecção de automóveis em condições de iluminação variadas com uma câmera de videovigilância

Elian Laura
Círculo de Investigação
Universidade Nacional San Agustín
Arequipa, Perú
Email: elian.laura.riv@gmail.com

Juan Carlos Gutiérrez
Escuela de Ciência da Computação
Universidade Nacional San Agustín
Arequipa, Perú
Email: jcgutierrezc@gmail.com

***Abstract*—In this work we evaluated different models of convolutional neural networks (CNN) for automobile detection. We obtained 27 models from the combination of three hyperparameters: technique to initialize weights, subsampling function and activation function. We use the accuracy as a measure factor to find the best model. Finally we made a comparison of the best CNN model with a cascade classifier and a support vector machine. Our dataset is created from a video surveillance camera under various lighting conditions such as noon light, afternoon light, and night with a camera in infrared mode. The results shows that a CNN gets the best result for the automobile detection which is important for practical applications.**

***Keywords*-Image processing; convolucional networks; automobile detection in images.**

***Resumo*—Neste trabalho diferentes modelos de redes neurais convolucionais (CNN) para detecção de automóveis são testados, 27 modelos são obtidos a partir da combinação de três hiperparâmetros: A técnica inicialização de pesos, a função de sub-amostragem e a função de activação. A percentagem de precisão é o factor de medição para encontrar o modelo com melhor desempenho. Foi realizado uma comparação do melhor modelo de CNN com um classificador cascade e uma máquina de vetores de suporte. Os conjuntos de amostras são obtidas a partir de uma câmera de videovigilância em condições de iluminação variadas, luz do meio-dia; luz da tarde; e noite. Porém às amostras de noite são obtidas no modo infravermelho. Os resultados demonstram que uma CNN obtive o melhor resultado para detecção de automóveis com uma câmera de videovigilância, o que é importante em aplicações práticas.**

***Palavras-chave*-Processamento de imagens; redes convolucionais; detecção de automóveis em imagens.**

I. Introdução

O presente póster apresenta o progresso do projeto de pesquisa sobre detecção automática de automóveis. As contribuições do projecto são: (1) Um modelo de CNN com hiperparâmetros ajustados para o reconhecimento de automóveis no tempo real, com uma câmera de videovigilância. Têm sido realizados a combinação de três hiperparâmetros: Técnica de inicialização de pesos, função de sub-amostragem(*subsampling*) e função de ativação. Destes encontrou-se um modelo com o melhor percentagem de precisão nas imagens de automóveis capturados perto do meio-dia, também testou-se com imagens tiradas em diferentes momentos do dia. (2) Uma técnica para controlar a iluminação variada. Como a câmera está em constante vigilância em uma única posição na vía pública o cambio de iluminação durante o dia é progressivo, mas é diferenciável entre o dia e a noite. A detecção automática debe ser em tempo real, para aplicações praticas e é utilizado uma câmara de videovigilância focalizando imagens na rua fora de uma universidade.

A. Trabalhos relacionados

Por muitos anos a detecção de automóveis foi resolvido com os algoritmos AdaBoost em cascade, a máquina de vetores de suporte e redes neurais, demonstrando bons resultados quando são mixturados com técnicas de extração de características. Nos últimos anos, a abordagem de aprendizagem profunda estão melhorando os resultados por sua elevada abstracção das características da imagem.

Wang *et al.* [1] propõe um detector de veículos com base em *deep belief network* (DBN), com uma arquitetura de duas camadas ocultas apresentou a menor taxa de erro. O autor usa imagens da parte traseira dos veículos, da base de dados Caltech1999 [2], e é complementada com imagens próprias do autor. O trabalho de Li *et al.* [3] propõe adaptar um detector de veículos a um dominio diferente utilizando uma rede neuronal convolucional, capaz de detectar o mesmo objeto em outro domínio. Dois detectores teve que ser treinado, um para detectar veículos de perfil e outro para veículos laterales. A CNN também é usada para extrair vetores de características, logo são treinados com outro classificador como AdaBoost Cascade (C-Haar) ou máquina de vetores de suporte (Support Vector Machine - SVM).

Os trabalhos sobre detecção automática de imagens com *deep learning* faz uso de um conjunto de imagens que contêm os automóveis na vista frontal / traseira ou lateral com iluminação constante, usando uma câmera montada em outro veículo da cena. Alguns autores experimentam em cenas públicas, mas não tem bons resultados. Por isso a importância do presente projeto.

B. Visão geral técnica

A proposta do sistema automático, esboçada na Figura 1, Apresenta duas etapas: um detector de movimento e um detector de automóveis baseado na aprendizagem profunda.

Temos disponível uma câmera de vídeo de vigilância localizada na rua, cerca de 4m. sobre na calçada, do lado de fora de uma universidade.

Na primeira etapa do sistema automático detecta qualquer objeto da cena que está em movimento. Se emprega a técnica *Motion History Image* (MHI) o qual estabelecera segmentos ou caixas delimitadoras dos automoveis na imagem que debem ter uma dimensão mínima e máxima, evitando assim segmentos que não se assemelham ao tamanho de um automóvel.

Na segunda etapa tem-se um classificador binário previamente treinado com a técnica de aprendizado profundo, a fim de classificar os segmentos de imagem em 2 classes: automóveis e não automóveis. O classificador é um modelo de rede neural convolucional (CNN) otimizado para detecção de automóveis com hiperparâmetros.

O treinamento é feito com imagens obtidas a partir de vídeos gravados ao meio-dia com uma iluminação clara. E os testes são feitos com imagens muito diferenciável em iluminação com referencia as imagens de treinamento.

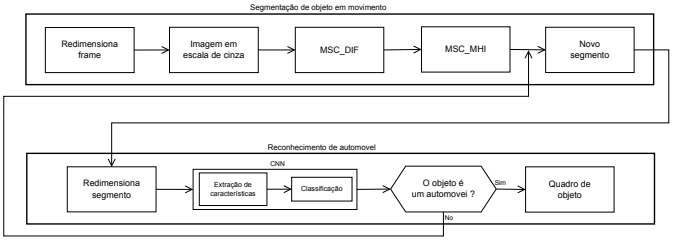


Figura 1. Diagrama de fluxo do sistema automático de detecção de automóveis.

II. Técnicas de fundo

A. Detecção de movimento com MHI

Tem-se uma sequência de imagens de cinzento, a partir do qual é gerada uma máscara de diferença (MDIF) entre cada par de imagens consecutivas. As intensidades dos pixeles de MDIF determina a sua informação de movimento em outra máscara que chamaremos máscara MHI (MMHI). A informação do movimento obtive-se do *timestamp*, este é o registro atual do tempo em milisegundos. Então por cada pixel armazena-se o valor *timestamp* ou o mesmo valor. Se o movimento de um pixel é antigo então estabelece-se em cero.

$$MHI(x,y) = \begin{cases} \textit{timestamp} & \text{if } MDIF(x,y) \neq 0 \\ 0 & \text{if } MDIF(x,y) = 0 \ \textbf{\&} \ MMHI < (\textit{timestamp} - \text{duration}) \\ MMHI(x,y) & \text{other cases} \end{cases} \quad (1)$$

Assim MMHI é obtido para posteriormente analisar o historico do movimento. A partir daqui tem-se a possibilidade de calcular o gradiente e orientação em cada pixel.

B. Rede Neural Convolucional (CNN)

A CNN suprime atenção na cantidade de capas concentrando na profundade, e aprendendo de forma hierarquica características cada vez mais complexas fazendo previsões mais precisas. CNN é constituída por 2 partes: 1) o extractor de características automática, que consiste de uma camada de convolução e outra de *subsampling*, 2) o classificador, que é uma rede neural totalmente conexa. O objetivo de uma camada convolucional é aprender a representação de características através da convolução entre cada mapa de características e um filtro contendo os valores aprendidos (pesos). Na camada sub-amostragem através de uma função de resolução reduz as características . A arquitectura de uma CNN termina com um ou mais camadas completamente conexas, igual que um perceptron multicamada com sua função de activação.

C. Hiperparâmetros de CNN

A seguir são descritas três hiperparâmetros duma CNN usados na busca do melhor modelo para uma detecção automática de automóveis.

1) Inicialização de pesos

Distribuição Uniforme	$x \sim U(a,b)$ intervalo [a,b]
Distribuição de Gauss	$x \sim N(\theta, \delta)$ média θ, desvio padrão δ
Algoritmo Xavier	$r = \sqrt{\frac{6}{n_{in}+n_{out}}}$ conexões que entram no neurônio, (n_{in}) conexões emergentes do neurônio (n_{out})

2) Função de Activação

Função Sigmóide	$\sigma(x) = \frac{1}{(1+e^x)}$ restrita entre 0-1
Função ReLU	$relu(x) = max(x, 0)$ Com discontinuade em 0
Função PRELU	$prelu(x_i) = max(0, x_i) + a_i min(0, x_i)$ a_i é aprendido pelo o canal i

3) Função subsampling

Subsampling máximo (MAX)	$s_j = \max_{i \in R_j} a_i$ seja R a região
Subsampling promedio (AVE)	$s_j = \frac{1}{\lvert R_j \rvert} \sum_{i \in R_j} a_i$
Subsampling estocástico (STO)	$p_i = \frac{a_i}{\sum_{k \in R_j} a_k}$ probabilidades p de cada região j

III. Proposta

Na Figura 2 apresenta-se a arquitetura de rede neural convolucional usada para o reconhecimento dos automóveis, baseado na arquitetura LeNet-5 proposta por Yann LeCun [4]. A camada de convolução (*convolutional layer*) é chamado CLX, a camada de *subsampling* é chamado SLx, e uma camada completamente conexa (*full connection*) será abreviada como FCLx. A imagem de entrada é 44px de largura e 28px

de altura, porque as imagens de automóveis têm uma posição diagonal, como é mostrado na tabela I.

Na camada CL1 a convolução é efectuada em imagens de 40x24 px, a redução no tamanho é devido ao kernel de filtro. Em seguida, a camada subsampling SL2 reduz a imagem com uma função subsampling, a camada CL3 obtém por segunda vez as imagens convolvidos que serão reduzidas em camada SL4, a camada CL5 realiza uma convolução e simultaneamente age como uma camada totalmente conexa ao através de uma função de ativação. Ela é ligada para a próxima camada totalmente conexa FCL6. Finalmente uma função de regressão *softmax* gera uma distribuição de probabilidade dos valores de saída que indicam o valor preditivo para a imagem de entrada, sendo rotulado '1' sim foissem automóveis e '0' para a imagem de não automóveis.

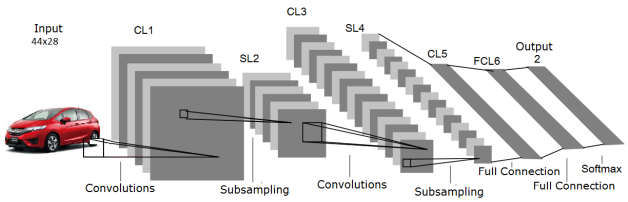


Figura 2. Arquitetura LeNet-5 para o detecção de automóveis. Figura baseada na proposta de LeCun *et al.* [4].

IV. Implementação

A biblioteca Caffe [5] é um *framework* que engobla a implementação de modelos de aprendizagem em profundidade. Seu código fonte está publicada na https://github.com/BVLC/caffe. Entre os modelos Caffe - CNN desenvolvidos tem-se LeNet-5, IMAGEnet, GoogleNet, AlexNet, entre outros. Ele pode ser executado na CPU ou GPU. A execução da bliblioteca Caffe para o presente experimento foi conduzido num computador de 8-core, 8GB de RAM, 2,7 Ghz, contendo uma placa gráfica nVidia GeForce GT 750.

V. Experimentos

A. Amostras de treinamento e teste

O conjunto de amostras positivas e negativas foram segmentadas a partir de vídeos obtidos com uma câmera de videovigilância [6]. As imagens de treinamento são vistos na tabela I, onde os automóveis são apresentados na frente, traseira e perspectiva oblíqua. Eles são 5000 imagens de amostras utilizadas no treinamento, sendo 2.500 amostras positivas e 2.500 negativas. Estas imagens foram tiradas a partir de gravações feitas apoximadamente ao meio-dia com luz clara.

O conjunto de amostras para testes também foram obtidos com a câmara de videovigilância, a partir de gravações que foram feitas em circunstâncias com iluminação diferente, têm sombra e uma iluminação um pouco mais opaca ao contrário das imagens de treinamento.

Os conjuntos de amostras são rotulados com dois valores, 0 se imagem é de fundo, e 1 se for um carro, esses rótulos

Tabela I
Imagens para o treinamento do classificador CNN.

Amostras positivas	Amostras negativas
	

são denomidas etiquetas reais. As etiquetas de saída são as resultantes do modelo de detecção, elas são comparados com etiquetas reais. O resultado da comparação é verdadeiro se ambas etiquetas coincidem, e falso caso contrário. A soma dos verdadeiros valores indica a precisão de amostras correctamente detectados. Estes resultados de detecção são analisados com uma fórmula de *accuracy* 2, onde, VP = verdadeiro positivo; VN = verdadeiro negativo; P = positivo; N = negativo.

$$Precisão = \frac{VP + VN}{P + N} \tag{2}$$

Primeiro experimento: Tabela III mostra 27 experimentos realizados com a combinação de três hiperparâmetros, referidos na seção II-C, a fim de encontrar o melhor modelo CNN para detecção de automóveis.

Tabela III tem as seguintes colunas: Modelo CNN (M-CNN), função de ativação (FA), inicialização de pesos(IP) função de sub-amostragem (FS) e a última coluna é o percentual de acerto na detecção das amostras de teste.

No total, 6.000 imagens de amostras, chamado CD11, foram submetidos a teste para encontrar o modelo CNN com melhor percentagem, sendo 3000 amostras positivas e 3000 amostras negativas. As amostras CD11 foram tomadas a partir de cenas com diferente iluminação que podem ser vistos na Figura 3a e na Fig. 3b.

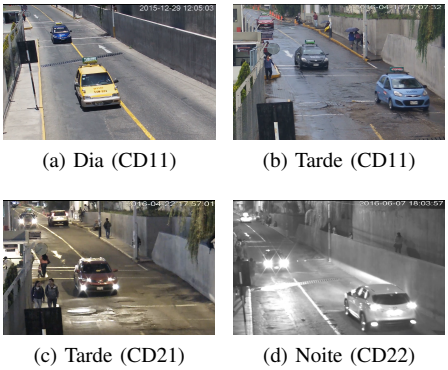
Na Tabela III podemos ver que os modelos M12 e M16 obtem uma precisão maior a 91%. Ambos modelos têm o hiperparâmetro de função de ativação ReLU. A diferença de precisão é de 0,38%, isto coloca M12 como melhor, sendo sua função de inicialização de pesos o Xavier e função de sumsampling AVE. Uma outra opção descrita neste artigo é M16 usando ReLU, onde a distribuição de Gauss foi usada para a inicialização de pesos com o desvio padrão de 0.01 e a função de submuestro MAX. O resultado de maior precisão nos experimentos reflete que a utilização de ReLU fornece um desempenho satisfatório, como propõe Krizhevsky *et al.* [7].

Segundo experimento: Como parte do projeto de pesquisa propomos uma técnica de manipulação de amostras com iluminação variada. Nosso segundo experimento envolve a coleta de dois conjuntos de imagens com iluminação diferente que o conjunto de imagens do primeiro experimento. Um conjunto de dados, chamada CD21 é capturado em horas da entardecer, e o segundo conjunto de dados que chamamos de CD22 foi coletado de uma gravação de noite com câmera infravermelha. As duas situações mencionadas, CD21 e CD22, podem ser vistos na imagem 3. A precisão obtida em CD21

Tabela II
COMPARAÇÃO NA PRECISÃO OBTIDA COM CNN E OUTRAS TÉCNICAS

Cenas	CNN-M12	C-Haar	SVM
CD11	91.38%	88.05%	81.25%
CD21	88.00%	81.95%	59.68%
CD22	66.00%	74.59%	64.54%

e CD22 com o modelo M12 é mostrado na Tabela II onde CNN-M12 representa o modelo M12 obtido no primeiro experimento, e também apresenta-se C-Haar e SVM onde eles foram treinados com o mesmo conjunto de amostras que CNN-M12.



(a) Dia (CD11) (b) Tarde (CD11)

(c) Tarde (CD21) (d) Noite (CD22)

Figura 3. Cenas que mostram a variação da iluminação, com o qual as imagens foram obtidas para os experimentos.

VI. RESULTADOS E DISCUSSÃO

As percentagens de precisão apresentados na tabela II demonstram que o modelo CNN é melhor para uma detecção automática em comparação com C-Haar e SVM. Como foi mencionado no segundo experimento, CD21 representa o conjunto de amostras colhidas em uma cena com pouca luz, Fig 3c. O percentual obtido pela CNN-M12 é de 88% que excedam as percentagens obtidas por C-Haar e SVM. No caso de CD22, Fig 3d, as amostras são colhidas numa cena nocturna onde os faróis de automóveis tornan mais dificil a detecção, neste segundo experimento o modelo CNN-M12 recebe o resultado inferior que C -Haar, mas é melhor que SVM.

A. Limitação

Como foi descrita na seção V Foram realizados experimentos em diferentes cenários de iluminação, a fim de medir a porcentagem de precisão que pode alcançar o sistema automatizado proposto com uma câmera de videovigilância por um dia inteiro. Para cenários com baixa iluminação planeja-se implementar um algoritmo de normalização de iluminação para esclarecer cenas e facilitar o processo da detecção dos automóveis. Os testes são realizados no dominio dos exteriores da universidade, espera-se melhorar o modelo CNN para atender outros tipos de veículos e com mais casos de iluminação.

Tabela III
RESULTADOS DA PRECISÃO OBTIDA POR CADA MODELO DE CNN

M-CNN	FA	IP	FS	Precisão
M1	Sigmóide	Xavier	Max	82.38%
M2	Sigmóide	Xavier	Sto	55.70%
M3	Sigmóide	Xavier	Ave	83.90%
M4	Sigmóide	Uniforme	Max	89.02%
M5	Sigmóide	Uniforme	Sto	58.37%
M6	Sigmóide	Uniforme	Ave	82.88%
M7	Sigmóide	Gauss	Max	85.83%
M8	Sigmóide	Gauss	Sto	70.90%
M9	Sigmóide	Gauss	Ave	83.48%
M10	ReLU	Xavier	Max	88.25%
M11	ReLU	Xavier	Sto	52.03%
M12	ReLU	Xavier	Ave	**91.38%**
M13	ReLU	Uniforme	Max	85.85%
M14	ReLU	Uniforme	Sto	50.25%
M15	ReLU	Uniforme	Ave	90.33%
M16	ReLU	Gauss	Max	**91.00%**
M17	ReLU	Gauss	Sto	60.15%
M18	ReLU	Gauss	Ave	87.45%
M19	ReLU	Xavier	Max	84.68%
M20	PReLU	Xavier	Sto	52.15%
M21	PReLU	Xavier	Ave	89.27%
M22	PReLU	Uniforme	Max	76.72%
M23	PReLU	Uniforme	Sto	50.02%
M24	PReLU	Uniforme	Ave	89.63%
M25	PReLU	Gauss	Max	89.02%
M26	PReLU	Gauss	Sto	65.60%
M27	PReLU	Gauss	Ave	82.22%

VII. CONCLUSÃO

O modelo CNN tem sido comparada com outras técnicas e revelam ser superior em duas situações de iluminação bem diferenciáveis, com a terceira situação de iluminação, CD22, que foi com a câmera no modo infravermelho, CNN apresenta um percentual inferior da técnica cascade Haar, mas supera a técnica SVM. Os testes revelam que CNN pode clasificar objetos de diferentes condições de iluminação, y a função ReLU demonstra as melhores percentagens. Espera-se realizar mais testes para determinar o melhor modelo de classificação das amostras do conjunto de dados CD22. O projeto ainda continua seu desenvolvimento para detecção automática dos automóveis em diversas condições de iluminação com uma câmera de videovigilância.

REFERÊNCIAS

[1] H. Wang, Y. Cai, and L. Chen, "A vehicle detection algorithm based on deep belief network," *The scientific world journal*, vol. 2014, 2014.

[2] Caltech, "www.vision.caltech.edu/html-files/archive.html," 1999. [Online]. Available: www.vision.caltech.edu/html-files/archive.html

[3] X. Li, M. Ye, M. Fu, P. Xu, and T. Li, "Domain adaption of vehicle detector based on convolutional neural networks," *International Journal of Control, Automation and Systems*, pp. 1–12, 2015.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.

[6] Dahua-Technology, "http://www.dahuasecurity.com/es/," 2010. [Online]. Available: http://www.dahuasecurity.com/es/

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.