Il Deep Learning per le applicazioni di sorveglianza

Gianluca Francini, Enrico Magli
TIM, Politecnico di Torino
gianluca.francini@telecomitalia.it, enrico.magli@polito.it

Abstract

L'identificazione automatica degli oggetti ripresi dalle telecamere cittadine è una delle tecnologie chiave per la realizzazione delle *smart city*. Per il riconoscimento, la tecnologia tradizionale si basa su apparati costosi, che ne limitano l'adozione a pochi punti della città. In questo documento presentiamo alcuni lavori basati sul Deep Learning del Joint Open Lab TIM/Politecnico di Torino, aventi l'obiettivo di consentire un riconoscimento robusto delle immagini anche in presenza di telecamere a basso costo.

1 Introduzione

Molti dei servizi che sono alla base delle *smart city* si basano sulla raccolta di informazioni relative al flusso di persone e mezzi e, per essere efficaci, è necessario che tale raccolta avvenga su tutto il territorio cittadino e non sia applicata esclusivamente a aree specifiche, come le zone a traffico limitato che si trovano nei centri storici.

Con le tecnologie tradizionali, collocare nel territorio cittadino un elevato numero di telecamere può essere molto costoso, soprattutto se ciò che è ripreso dalle telecamere è destinato a essere processato da un sistema automatico. Questo perché gli apparati che eseguono dei riconoscimenti automatici, come le telecamere che controllano i varchi automobilistici, devono essere collocati in posizioni precise rispetto a quanto osservato, sono costituiti da lenti e sensori costosi e richiedono dei dispositivi addizionali, come gli illuminatori a infrarossi che sono impiegati per il riconoscimento delle targhe.

I costi necessari per posizionare delle telecamere nel territorio cittadino possono essere ridotti notevolmente grazie alle nuove tecnologie di telecomunicazione. È ora possibile collocare delle telecamere che si appoggiano alla rete radiomobile per l'invio dei dati, garantendo dei flussi video di alta qualità e eliminando il costo dell'allestimento di un collegamento fisico con la centrale operativa, come nel servizio LTE Public Safety[1] offerto da TIM. Questo tipo di dispositivi sarà sempre più diffuso e vedrà un'ulteriore accelerazione nell'adozione con l'avvento della tecnologia 5G, che introdurrà ulteriori miglioramenti in temini di banda disponibile, riduzione della latenza e introduzione di nuova funzionalità.

Nonostante i vantaggi in temini di costi che si ottengono usando la rete radiomobile, rimane il problema di dover usare telecamere costose, fino ad ora necessarie al fine dell'elaborazione automatica delle immagini. Questo è ancora un vincolo che ostacola significativamente la creazione di infrastrutture basate su molti sensori ed è ciò che ci ha spinti a lavorare nel Joint Open Lab[2] TIM/Politecnico di Torino a lavorare sulle tecnologie Deep Learning applicate all'analisi delle immagini.

Qualche anno fa si è verificato un evento storico nel campo dell'analisi delle immagini. La competizione ImageNet Large Scale Visual Recognition Challenge [Berg *et al.*, 2010], il cui l'obiettivo è quello di classificare centinaia di oggetti presenti in più di un milione di immagini, è stato vinto da AlexNet [Krizhevsky *et al.*, 2012], una rete neurale profonda. AlexNet ha ridotto del più del 10% l'errore di classificazione minimo sino a quel momento ottenuto e ha di conseguenza dato il via alla rivoluzione del Deep Learning, rivoluzione che si è estesa a molti campi di analisi al di là di quello delle immagini.

La qualità dei risultati che il Deep Learning è in grado di fornire ci ha spinti ad applicarlo al campo dell'analisi delle immagini acquisite da telecamere posizionate nel contesto urbano, al fine di identificare automaticamente alcuni degli elementi che costituiscono la scena inquadrata.

[1] https://www.telecomitalia.com/tit/it/notiziariotecnico/edizioni-2017/n-2-2017/capitolo-7.html

[2] https://www.telecomitalia.com/tit/en/sustainability/strategy-objectives/TIM-model/innovation/JOL.html

2 Il Deep Learning per le applicazioni di sorveglianza

Nel Joint Open Lab abbiamo lavorato a tre tecnologie basate sulle reti neurali profonde: un riconoscitore di targhe automobilistiche, un classificatore della tipologia dei veicoli e un identificatore di pedoni. Nell'impostazione del lavoro abbiamo considerato due requisiti. Il primo è che si avesse un'elevata qualità del riconoscimento con telecamere a basso costo, comprese quelle dei dispositivi mobili come cellulari e tablet. Il secondo è che la complessità delle reti neurali fosse sufficientemente contenuta, in modo da poter elaborare le immagini in tempi ristretti e poter gestire più flussi con un singolo sever. Di seguito saranno illustrate sinteticamente tre tecnologie sviluppate congiuntamente da studenti di dottorato, ricercatori TIM e docenti del Politecnico.

1.1 Riconoscimento delle targhe

Le reti neurali profonde sono costituite da molti neuroni e di conseguenza da un elevato numero di parametri, il cui valore deve essere determinato durante la fase di addestramento. Per poter applicare il Deep Learning è quindi necessario avere a dispozione un elevato numero di immagini annotate, in modo da ottenere un sistema che sia robusto e che eviti il problema dell'*overfitting*. È difficile ottenere un dataset costituito da centinaia di migliaia di immagini a causa dell'elevato tempo necessario nell'acquisizione e annotazione manuale dei dati e ai problemi logistici nella collocazione delle telecamere. Questo sforzo deve poi essere moltiplicato per il numero di tipologie di targhe che si vogliono riconsocere. Per ovviare a questo problema, abbiamo provato a creare dei dataset di targhe sintetiche, scoprendo che possono sostituire egregiamente le immagini naturali [Björklund *et al*., 2017] [Rizvi *et al*., 2017]. Il procedimento precede la costruzione di un template della targa avente una sequenza casuale di lettere e cifre (rispetando gli eventuali vincoli del formato). La targa sintetica è alterata in modo da simulare vari cambiamenti visivi a cui può essere sottoposta, come riflessi e obreggiature. Ne vengono alterati i colori e infine è distorta prospetticamente e applicata su un'immagine casuale, selezionata da un dataset di circa un milione di immagini, come mostrato in figura 1.

Figura 1: Immagini sintetiche di targhe automobilistiche

Lo sfondo casuale non influenza negativamente le prestazioni del riconoscimento, testato su un insieme di immagini reali annotate manualmente e consente di generare centinaia di migliaia di esempi.

L'approccio sviluppato si basa su due reti neurali convoluzionali [LeCun *et al*., 2004]. La prima serve a rilevare la presenza di una o più targhe nell'immagine, identificandone la posizione e alternado l'immagine per compensare la distorsione prospettica. La seconda rete processa la porzione di immagine occupata dalla targa e ne riconosce la sequenza di lettere e cifre. Il sistema ottenuto è in grado di eguagliare o superare l'accuratezza ottenuta con lo stato dell'arte, valutata su dataset pubblici come l'Application Oriented License Plate database [Hsu *et al*., 2013] e il Chinese plates PKU dataset [Yan *et al*., 2017].

1.2 Classificazione dei veicoli

Numerosi approcci nello stato dell'arte prevedono che le immagini usate nell'addestramento siano annotate in modo complesso. In particolare, richiedono che in ogni immagine siano identificate manualmente delle porzioni di veicolo che siano altamente distintive e che possiedano una relazione spaziale fissa tra di esse, come ad esempio il frontale dell'automobile e le ruote.

L'approccio adottato in questo lavoro [Ghassemi *et al*., 2017] non richiede questo tipo di annotazione e si basa su una rete neurale di tipo residuale che identifica un insieme di *attention window*, cioè porzioni del veicolo che il sistema considera significative, come mostrato nella seguente figura.

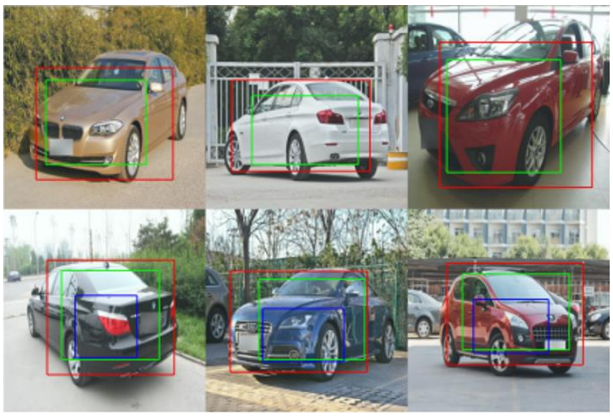

Figura 2: Esempi di attention window identificate: due nella riga superiore, tre in quella inferiore

Il contenuto di ogni attention window è successivamente elaborato da un insieme di reti neurali residuali [He *et al*., 2016] il cui output è fuso e utilizzato come input di due classificatori, uno usato per determinare la marca del veicolo e uno per il modello.

Il sistema sviluppato fornisce prestazioni superiori allo stato dell'arte in temini di accruratezza sia nell'identificazione del produttore sia del modello del veicolo, valutati sui dataset pubblici Stanford Cars Dataset [Krause *et al*., 2013] e The Comprehensive Cars (CompCars) dataset [Yang *et al*., 2015].

1.3 Identificazione dei pedoni

L'identificazione dei pedoni è un problema complesso da risolvere, dato che le persone possono avere aspetti molto variegati a causa della postura e del tipo di vestiario. Vi sono inoltre due aspetti critici addizionali. Spesso i pedoni sono parzialmente occlusi da altri pedoni o da oggetti (automobili, pali, fermate del bus, ecc.), inoltre esiste un'alta variabilità nella dimensione che occupano nella scena, a causa della differente distanza dalla quale possono trovarsi rispetto alla telecamera.

Per ottenere un sistema di identificazione robusto anche in presenza di queste criticità e che nel contempo sia in grado di elaborare numerosi fotogrammi al secondo, l'approccio sviluppato è stato basato su una Feature Pyramid Network [Lin *et al*., 2017], una rete neurale che analizza l'immagine creando una piramide di più scale, basata su una rete neurale convoluzionale di tipo ResNet 101. L'output di questo modulo è rappresentato da un insieme di *feature map* che sono inviate a una Region Proposal Network, che ha l'obiettivo di identificare quali feature map sono più significative per la rilevazione dei pedoni. Le regioni identificate sono successivamente fornite in input a un classificatore basato su una architettura Faster R-CNN [Ren *et al*., 2015], la quale ha il compito di eseguire la classificazione finale e il raffinamento della posizione del pedone mediante regressione. In figura 3 è mostrata la catena di elaborazione completa, contenuta in una sigola rete neurale.

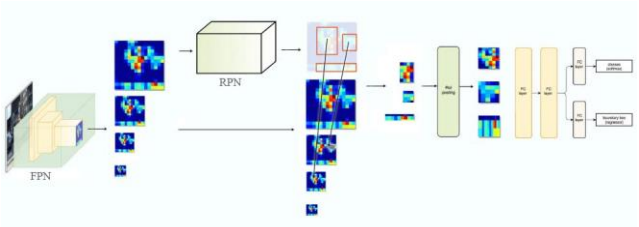

Figura 3: Rete neurale basata su Feature Pyramid Network e Faster R-CNN

Riferimenti bibliografici

[Berg *et al.*, 2010] A. Berg, J. Deng, and L. Fei-Fei. Large scale visual recognition challenge 2010. http://image-net.org/challenges/LSVRC/2010, 2010.

[Krizhevsky et al., 2012] Krizhevsky, Alex & Sutskever, Ilya & E. Hinton, Geoffrey. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems*. 25. 10.1145/3065386, 2012.

[Björklund et al., 2017] Tomas Björklund, Attilio Fiandrotti, Mauro Annarumma, Gianluca Francini, Enrico Magli. Lightweight License Plate Recognition using Neural Networks Trained on Synthetic Images. *IEEE Transactions on Multimedia* (ISSN 1520-9210). 2017.

[Rizvi et al., 2017] Syed Tahir Hussain Rizvi, Denis Patti, Tomas Björklund, Gianpiero Cabodi and Gianluca Francini. Deep Classifiers-Based License Plate Detection, Localization and Recognition on GPU-Powered Mobile Platform. *Future Internet* (ISSN 1999-5903). 2017.

[LeCun et al., 2004] Y. LeCun, F.J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *Computer Vision and Pattern Recognition (CVPR)*. 2004.

[Ghassemi et al., 2017] Sina Ghassemi, Attilio Fiandrotti, Gianluca Francini, Enrico Magli. Fine-Grained Vehicle Classification using Deep Residual Networks with Multiscale Attention Windows, *IEEE 19th International Workshop on Multimedia Signal Processing*. 2017.

[Hsu et al., 2013] G. Hsu, J. Chen and Y. Chung, Application-Oriented License Plate Recognition, in *IEEE Transactions on Vehicular Technology*, vol. 62, no. 2, pp. 552-561, Feb. 2013.

[Yan et al., 2017] K. Yan, Y. Tian, Y. Wang, W. Zeng and T. Huang. Exploiting Multi-grain Ranking Constraints for Precisely Searching Visually-similar Vehicles. *IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 562-570. 2017

[He et al., 2016] K. He, X. Zhang, J. Sun. Deep Residual Learning for Image Recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016): 770-778. 2016.

[Krause et al., 2013] Krause, J., Deng, J., Stark, M., Fei-Fei, L. Collecting a large-scale dataset offine-grained cars. *CVPR-FGCV2*. 2013.

[Yang et al., 2015] Linjie Yang, Ping Luo, Chen Change Loy, Xiaoou Tang. A Large-Scale Car Dataset for Fine-Grained Categorization and Verification. *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[Lin et al., 2017] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, Serge J. Belongie. Feature Pyramid Networks for Object Detection. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 2017.

[Ren et al., 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*. 2015.