ARQUITECTURA MEJORADA DE DEEP LEARNING PARA REIDENTIFICACION DE PERSONAS EN CAMARAS DE VIGILANCIA

Jonathan Durand Espinoza
Maestría en ciencias de la computación, Universidad Católica San Pablo
Arequipa, Perú
E-mail: jonathan.drnd@gmail.com

En este trabajo, proponemos una red neuronal profunda de extremo a extremo para aprender simultáneamente características de alto nivel y una correspondiente métrica de similitud para la re-identificación de personas en camaras de videovigilancia. La red neuronal toma un par de imágenes RGB sin procesar como entrada y emite un valor de similitud que indica si las dos imágenes de entrada representan a la misma persona. Una capa oculta calcula las diferencias entre las CNN siamesas el cual es empleado para capturar la relacion entre las imagenes. Esta operación consiste en buscar una característica robusta de las imágenes de entrada. Al aumentar la profundidad a 10 capas y usar filtros de convolución muy pequeños (3 x 3), nuestra arquitectura consigue una mejora notable en las configuraciones de las técnicas en el estado del arte. Utilizamos la optimizacion RootMean-Square adaptativo (RMSProp) que está integrado en nuestra arquitectura, lo que es beneficioso para las redes profundas. Nuestro método supera sistemáticamente el estado de la técnica en dos conjuntos de datos grandes (CUHK03 y Market-1501), y un conjunto de datos de tamaño mediano (CUHK01). Finalmente proponemos un metodo que reemplaza al standard maxpooling para acelerar la convergencia y la precision.

Introducción

El proposito de la reidentificacion personas es realizar el matching de las personas observadas en diferentes vistas de camaras. Hay importantes aplicaciones en videovigilancia como el tracking entre varias camaras, deteccion de acciones/eventos multicamara.

La re-identificación de una persona es esencialmente para medir la similitud de pares de imágenes peatonales de tal manera que a un par se le asigne un alto puntaje de similitud en caso de representar la misma identidad y una puntuación baja si se muestran diferentes identidades.

Esto normalmente implica construir una representación de características sólidas y una medida de similitud apropiada para estimar puntuaciones de similitud precisas. Para este fin, muchos métodos se centran en la representación de features y aprendizaje de la función de la distancia que se diseñan por separado o conjuntamente para ocuparse del problema de reidentificacion de personas. Los features de bajo nivel como el color y la textura se pueden utilizar para este propósito. Algunos estudios han obtenido representaciones de características más distintivas y fiables, incluyendo simetría, partición horizontal y obtencion de rasgos sobresalientes. Sin embargo, todavía es difícil diseñar un tipo de característica que sea discriminatoria e invariable a cambios severos en términos de desalineación, oclusion, brillo a través de vistas de cámara desiguales.

Las Redes Neuronales Convolucionales (CNNs) han demostrado ser muy exitosas en problemas de reconocimiento de imágenes y en diversas aplicaciones de vigilancia incluyendo detección de peatones y tracking. Nuestro objetivo es mejorar la arquitectura del estado del arte para lograr una mayor precisión. Las principales contribuciones de nuestro trabajo son:

1. Presentamos una red profunda usando una arquitectura pequeña (3x3) filtro de convolucion, el cual mejora el performance que utilizando una red de (5x5) como la referencia [1]

2. Utilizamos el metodo de descenso de gradiente, RMSPROP el cual permite una convergencia mas rapida de nuestra red.

3. Se ha realizado varios experimentos, ajustando parametros en dataset publicos como CUHK01, CUHK03 y Viper obteniendo buenos resultados.

Figura 1– Ejemplo de multiples dataset de reidentificacion de personas

El problema a resolver, son de peatones que son vistos a traves de multiples camaras, estas imagenes pueden verse afectadas por oclusiones, brillo, iluminacion, cambios de perspectiva. En estos casos las camaras al estar en ambientes abiertos, y cercanamente posicionadas, las personas estan con la misma vestimenta, como se observa en el dataset CUHK03 (el cual posee una mayor cantidad de imagenes comparados con otros dataset).

Figura 2– Ejemplo de pares positivos

I. Trabajos Relacionados

Muchos estudios recientes en reidentificacion de personas generan robustos representaciones de features que describen la apariencia de peatones bajos ciertas condiciones [4],[5],[6]

Hay 4 algoritmos de aprendizaje de reidentificacion de personas que han sido propuestos. Yi [3] utiliza una red siamesa CNN con estructura simetrica que comprende dos subredes independientes y emplea la funcion coseno como metrica. Li [1] Diseñó una red diferente, que comienza con una sola capa de convolución con max pooling, seguida por una capa de emparejamiento que multiplica las respuestas de las características convolucionales de las dos entradas en una variedad de desplazamientos horizontales, es decir utiliza una metrica para comparar 2 imagenes. Nuestra arquitectura es similar al del paper JointRe-id [2] el cual utiliza una capa oculta, el cual realiza la comparacion entre 2 CNNs que corresponden a cada imagen que se requiere comparar si corresponden a una misma persona.

Nuestra arquitectura difiere sustancialmente de estas redes anteriores, es mas profunda con filtros de convolucion 3x3 ya que se obtienen mejores resultados que al realizar convoluciones de 5x5 como en la referencia [2].

II. Arquitectura

Durante el entrenamiento tanto el dataset CUHK01 y CUHK03 tienen las dimensiones 160x60 imagenes RGB, en otros dataset se realizo el resize de las imagenes para que tengan estas dimensiones.

En este trabajo proponemos una arquitectura de red neuronal profunda que formula el problema de la re-identificación de la persona como clasificación binaria. Dado un par de imágenes de entrada, la tarea consiste en determinar si las dos imágenes representan o no a la misma persona. La Figura 2 ilustra la arquitectura de nuestra red. Brevemente Nuestra red consta de las siguientes capas distintas: dos capas de convolución ligada con 2 max pooling, 1 capa que realiza el merge de las CNN siamesas y finalmente, las capas fullconected para producir la estimación final de si las imágenes de entrada son de la misma persona o no. Cada una de estas capas se explica en las subsecciones siguientes.

a) Convolucion, Maxpooling

Las 2 primeras capas son de convolucion y Maxpooling. Dado 2 imagenes I y J observados por diferente camara en tamaños de 160x60, canales RGB, aplicamos filtros de convolucion de 3x3. La capa de maxpooling sirve para reducir la dimensionalidad a la mitad de nuestra red, este es aplicado en cada pixel alreededor de un vecindario. La funcion de activacion, se realizaron varias pruebas, en el que se concluyó en que la funcion RELU no proporciona un buen resultado, por lo que se ha optado por utilizar una funcion de activacion no lineal como la funcion de tangente hiperbolica, tanh(3x/2).

b) Merge-diferencia

En este capa, nosotros recibimos como entrada 2 CNN que provienen de cada imagen, lo que se desea es combinar de tal manera de que sea posible determinar las similitudes y diferencias de las caracteristicas obtenidas del par de imagenes.

Sea f y h las CNN de la primera y segunda imagen respectivamente.

$$K_i(x,y) = f_I(x,y)\mathbb{I}(3,3) - \mathcal{N}[h_J(x,y)]$$

I(3,3) es una matriz de 3x3 que contienen 1s

f I(x,y) corresponde al valor obtenido dado el pixel (x,y) del feature map en referencia en la primera CNN correspondiente a la primera imagen de dimensiones 32x32x7

hI(x,y) corresponde al valor obtenido dado el pixel (x,y) del feature map en referencia en la segunda CNN correspondiente a la primera imagen de dimensiones 32x32x7

N[hI(x,y)] corresponde a una matriz de 3x3 que se encuentra centrado en el pixel (x,y) del feature map en referencia.

Como podemos observar luego de aplicar la funcion merge, obtenemos la matriz K que representa las diferencias entre las 2 imagenes de entradas, esta matriz tiene como dimensiones (32x3)x(7x3)x(7x3). A partir de esto, reducimos la dimensionalidad utilizando la capa full connected de un vector de 4096 caracteristicas a 512 y finalmente reducir a una clasificacion binaria que me debe indicar si las 2 imagenes de entrada corrsponde a la misma persona o no corresponde.

Al comparar con la arquitectura [2] nuestra arquitectura tiene una menor cantidad de variables de aprendizaje ya que utiliza bloques de (3x3) en vez de (5x5), esto es compensado en la cantidad de capas ocultas que utiliza nuestro modelo. A pesar de esto se observa que el tiempo de ejecucion del programa es menor.

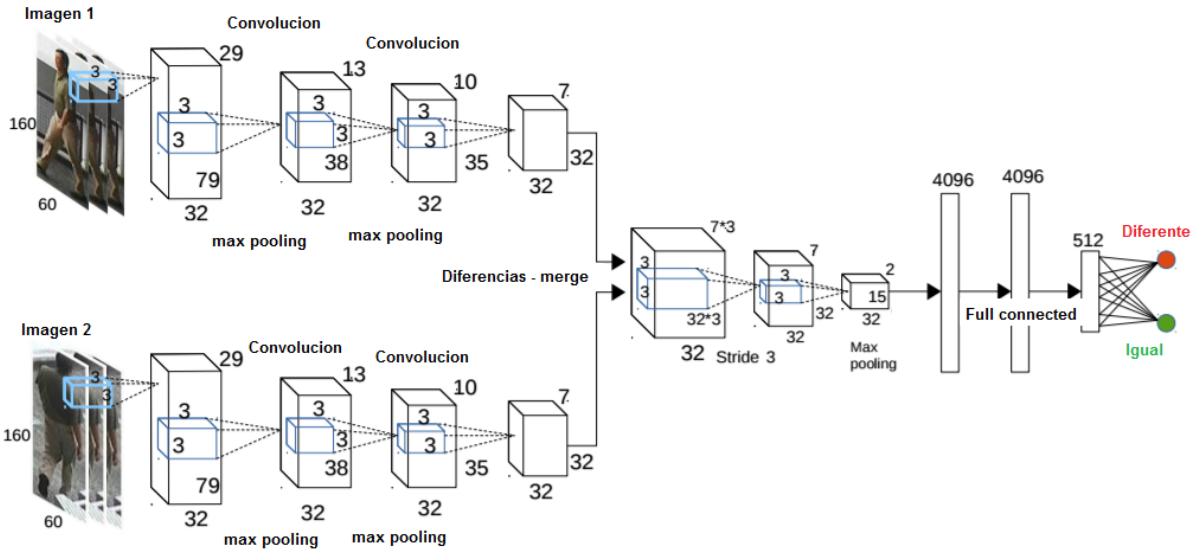

Figura 3– Arquitectura propuesta

Estrategia utilizada

Utilizamos la función tangente hiperbólica como función de activación no lineal en nuestros modelos. Los datos de entrenamiento se dividen en mini-batches. Tenga en cuenta que empleamos RMSProp en lugar de comúnmente utilizado descenso de gradiente estocastico (SGD) como un medio de actualización de gradientes de parámetros.

III. Implementacion

Respecto a la implemntacion, el codigo se ha realizado en Python, utilizando como framework de deepLearning Tensorflow. Por facilidad de implementacion se utilizo el backend keras y Theano para el uso del GPU. En el entrenamiento se utilizaron 10000 epocas el cual tomo 48horas en convergen en NVIDIA Geforce GTX 860.

El codigo se encuentra en el siguiente link: https://github.com/jonathandrnd/SistemasInteligentes/PaperDeepLearningReidPerson/

Respecto a la convergencia el RMSProp el es mas estable y relativamente mas rapido que SGD. Esto se debe principalmente a que SGD por sí mismo depende únicamente del lote dado de instancias de la presente iteración. Por lo tanto, tiende a tener pasos de actualización inestables por iteración y la convergencia toma más tiempo o incluso se atasca en mínimos locales. Por el contrario, RMSProp mantiene un promedio de sus magnitudes de gradiente recientes y divide el siguiente gradiente por este promedio de modo que los valores de gradiente flojo se normalizan. En consecuencia, RMSProp funciona mejor en las actualizaciones de degradado en pasos de diferentes lotes

IV. Resultados

Al ejecutar el algoritmo obtenemos que en la etapa de entrenamiento obtenemos un 85 % para indicar si un par de imagenes corresponden a la misma persona o no. Para la comparacion de nuestro algoritmo con el estado del arte utilizamos las metricas rank1, rank5 y rank10. Rank5, quiere indicar de que dado una imagen, comparamos con toda la base y si entre las 5 mejores probabilidades que obtenemos al comparar con otra imagen que indique de que corresponde a la misma persona entonces es una respuesta valida.

En la siguiente grafica mostramos la comparacion de los resultados obtenidos.

Metodo	r=1	r=5	r=10
JointRe-id	65	88.7	93.12
SDALF	9.9	41.21	56
LDM	26.45	57.69	72.04
Nuestro	68	89.3	94.2

Figura 4– Resultados Metricas Deep Learning-CUHK03

En la tabla anterior observamos los diferentes valores obtenidos para el dataset CUHK03, se tomo este dataset como referencia ya que posee en promedio 6 imagenes por person en diferentes camaras, asimismo cuenta con 13164 imagenes de prueba a diferencia del CUHK01 que cuenta con 971 imagenes los cuales no son suficientes para la data de entrenamiento.

V. Conclusiones

En este trabajo, evaluamos redes convolucionales muy profundas (hasta 6 capas ocultas) para la re-identificación de personas. Se demostró que la profundidad de representación es beneficiosa para la exactitud de reconocimiento de personas y que el rendimiento de estado de los conjuntos de datos de re-id de persona incluyendo CUHK03, CUHK01 y Viper puede lograrse usando una arquitectura basada en concordancia efectiva, con notablemente mayor profundidad. Nuestros resultados experimentales justifican la importancia de la coincidencia de identidad de profundidad en persona.

VI. Referencias

1 W. Li, R. Zhao, T. Xiao, and X. Wang, "Deepreid: Deep filter pairing neural network for person re-identification," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2014.

2 E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2015.

3 D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Deep metric learning for person re-identification," in Proc. Int. Conf. Pattern Recogn., 2014.

4 N. Gheissari, T. B. Sebastian, and R. Hartley, "Person reidenti- fication using spatiotemporal appearance," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2006

5 R. Zhao, W. Ouyang, and X. Wang, "Learning mid-level filters for person re-identfiation," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2014

6 R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2013.

7 M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in Proc. IEEE Conf. Comp. Vis. Patt. Recogn., 2010

8 D. Gray and H. Tao, "Viewpoint invariant pedestrian recognition with an ensemble of localized features," in Proc. Eur. Conf. Comp. Vis., 2008

9 X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in Proc. IEEE Int. Conf. Comp. Vis., 2007

10 Z. Li, S. Chang, F. Liang, T. Huang, L. Cao, and J. Smith. Learning locally-adaptive decision functions for person verification. In CVPR, 2013