BiLSTM with CNN Features For HAR in Videos

Carlos Ismael Orozco[1], María Elena Buemi[2], and Julio Jacobo Berlles[2]

[1] Departamento de Informática, FCE. Universidad Nacional de Salta, Argentina
[2] Departamento de Computación, FCEyN. Universidad de Buenos Aires, Argentina
ciorozco.unsa@gmail.com, {mebuemi,jacobo}@dc.uba.ar

Abstract. El reconocimiento de acciones en videos es actualmente un tema de interés en el área de visión por computadora debido a sus potenciales aplicaciones tales como indexación en multimedia, vigilancia en espacios públicos, entre otras. En este trabajo proponemos una arquitectura CNN–BiLSTM. Primero, una red neuronal convolucional VGG16 previamente entrenada extrae las características del video de entrada. Luego, un BiLSTM clasifica el video en una clase en particular. Evaluamos el rendimiento de nuestro sistema utilizando la precisión como métrica de evaluación, obteniendo 40.9% y 78.1% para los conjuntos de datos HMDB-51 y UCF-101 respectivamente.

Keywords: Reconocimiento de acciones · CNN · BiLSTM.

1 Introducción

Dada una lista de posibles acciones y un video en el que muestra a un actor llevando a cabo una de ellas, el objetivo de los sistemas de reconcimiento de acciones humanas (HAR) es reconocer la acción siendo ejecutada dentro del video. El problema en cuestión es de gran interés en el área de visión por computadora debido a sus potenciales aplicaciones tales como: indexación multimedia, recuperación de información, monitoreo y control de pacientes, vigilancia automatizada en espacios públicos, entre otros. Un gran número de trabajos han sido propuestos, como por ejemplo:

Enfoques clasicos: Liu et al. [1] proponen un marco para la detección y el reconocimiento de las acciones humanas. Para lograr una estimación sólida de la región de interés, utilizan una combinación de flujo óptico junto con un detector de bordes Harris 3D para obtener información de espacio-tiempo a partir del video. Luego, con el cálculo de las características locales SIFT y STIP, entrenan un fondo de modelo universal (UBM) para la tarea en cuestión. Wang et al. [2] proponen un enfoque de trayectoria densa. Toman puntos densos en cada frame del video y los rastrean según la información de desplazamiento del flujo óptico.

Enfoques CNN-3D: Ji et al. [3] proponen un enfoque 3D-CNN. Extracción de las características de los datos en dos dimensiones: espacial y temporal, capturando así información de movimiento en las transmisiones de video. Esta arquitectura CNN genera múltiples canales de información de marcos de video adyacentes y realiza convolución y submuestreo por separado en cada canal. La representación de la característica final se obtiene combinando todos los canales.

Las Redes Neuronales de Corta y Larga Memoria (LSTM) [4] se utilizan para aprender dinámicas temporales complejas. En esta línea de investigación, Orozco et al. [5] proponen una arquitectura CNN–LSTM. Primero, una CNN previamente entrenada [6] extrae las características de los fotogramas del video de entrada. Luego, una red LSTM se encarga de la clasificación. [7] propone apilar una capa LSTM en la parte superior de las capas ConvNet 2D para incorporar modelado de información de movimiento explícito, ajustando capas convolucionales y recurrentes de manera conjunta de un extremo a otro directamente desde cuadros RGB (sin procesar) y sus componentes de flujo óptico.

El objetivo de este trabajo consiste en implementar un sistema de reconocimiento de acciones en video. Para ello proponemos el uso de una arquitectura CNN–BiLSTM. Una red neuronal convolucional extrae las características del video mientras que una red neuronal BiLSTM clasifica el video en una categoría determinada. El trabajo esta organizado de la siguiente manera: en la sección 2, se describe la estructura general del sistema; en la sección 3, se describe la base de datos empleada, la métrica de evaluación, los experimentos realizados y los resultados obtenidos. Finalmente, en la sección 4 se presentan las conclusiones y trabajo a futuro.

2 Nuestra Propuesta

Las redes LSTM bidireccionales (BiLSTM) [8] son una extensión de los modelos LSTM [4] en los que se aplican dos LSTM a los datos de entrada. En la primera vuelta, se aplica una LSTM con la secuencia normal de los fotogramas de entrada ($\overrightarrow{h}_t$ forward). En la segunda vuelta, se aplica una LSTM con la secuencia invertida de los fotogramas de entrada ($\overleftarrow{h}_t$ backward). Para calcular la salida y_t en el paso t se consideran ambas salidas, es decir:

$$\overrightarrow{h}_t = LSTM(x_t, \overrightarrow{h}_{t-1}) \qquad \overleftarrow{h}_t = LSTM(x_t, \overleftarrow{h}_{t-1}) \tag{1}$$

$$y_t = ReLU(\overrightarrow{h}_t, \overleftarrow{h}_t) \tag{2}$$

La Figura 1 muestra un esquema general del sistema en sus diferentes etapas. Los parámetros de la red se optimizan minimizando la función de pérdida de entropía cruzada utilizando el descenso de gradiente estocástico con la regla de actualización RMSProp [9].

- Entrada: el video v es normalizado en un total de 40 fotogramas.
- $Encoder$: Usamos la arquitectura convolucional VGG16 propuesta por [6]. Para cada $x_t \in v$ codificamos el fotograma en un cuboide X_t de tamaño $7 \times 7 \times 512$ resultante de la capa de submuestreo VGG16.
- $BiLSTM$: propuesto por [8] tiene como comportamiento natural, recordar información durante largos períodos de tiempo.
- MLP_1: formado por una capa de Dropout con parametro 0.5 y una capa densa con un nodo por cada clase.
- : Indica la dimensión de salida.

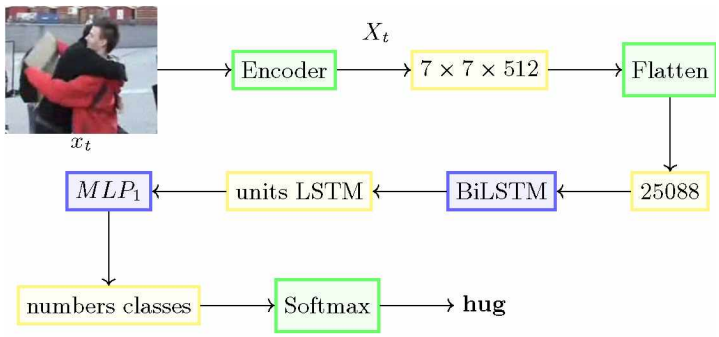

Fig. 1: Arquitectura CNN–BiLSTM propuesta.

3 Experimentos y Resultados

3.1 Base de datos

HMDB-51 Human Motion dataset propuesto por [10] tiene 6766 videos que pertenecen a una de las siguientes 51 clases: clap, drink, hug, jump, somersault, etc. También proporciona tres divisiones de prueba de entrenamiento, cada una de las cuales consta de 5100 videos, 3570 para entrenamiento y 1530 para test, es decir, una proporción de 70/30 por clase. Evaluamos la precisión promedio sobre estas tres divisiones.

UCF-101 dataset propuesto por [11] Estas 101 categorías se pueden clasificar en 5 tipos (interacción humano-objeto, solo movimiento corporal, interacción humano-humana, tocar instrumentos musicales y deportes). La duración total de estos videos es de más de 27 horas. Todos los videos se recopilan de YouTube y tienen una velocidad de 25 FPS con una resolución de 320×240. Para la división de los conjuntos de entrenamiento y test, seguimos la configuración propuesta en el artículo original [11].

3.2 Resultados

Nuestro sistema fue implementado en Python usando la librería Tensorflow [12] sobre una computadora Intel CORE i7-6700HQ con 16GB de memoria DDR3 y sistema operativo Ubuntu 16.04. Los experimentos se llevaron a cabo sobre una GPU NVIDIA Titan Xp montada en un servidor con características similares.

La Tabla 1 resume los resultados obtenidos por nuestro sistema, comparado con otros enfoques citados en la bibliografía.

4 Conclusiones

En este trabajo implementamos un sistema de reconocimiento de acciones en video, empleando una red neuronal CNN–BiLSTM. Primero, una VGG16

Table 1: Resultados de la clasificación de videos.

Autores	Referencia	Dataset	
		HMDB-51	UCF-101
Kuehne et al.	[10]	23.0%	-
Kliper-Gross et al.	[13]	29.2%	-
Jiang et al.	[14]	40.7%	-
Sharma et al.	[15]	41.3%	-
Li et al.	[16]	63.0%	-
Wang et al.	[17]	64.3%	-
Ye et al.	[18]	-	85.4%
Zhang et al.	[19]	-	86.4%
CNN–LSTM		40.7%	75.8%
CNN–BiLSTM		40.9%	78.1%

extrae las caracterísitcas del video. Luego una red neuronal BiLSTM clasifica la clase a cual pertenece. La misma fue implementada en Python usando la librería Tensorflow, entrenada y testeada utilizando la base de datos HMDB-51 y UCF-101. Evaluamos el rendimiento de nuestra propuesta utilizando la métrica de evaluación de precisión. Obtenemos $\sim 41\%$ y $\sim 78\%$ para entrenamiento y test respectivamente. El uso de la BiLSTM permite mejorar el aprendizaje de las dependencias a largo plazo y, por tanto, mejorará la precisión del modelo como se demuestra en [20].

Como trabajo futuro vamos a implementar los mecanismos de atención propuestas por [21].

Agradecimientos

Los autores agradecen a NVIDIA por la donación de una GPU TITAN Xp para el Departamento de Informática. Facultad de Ciencias Exactas. Universidad Nacional de Salta. Argentina.

References

1. D. Liu, M. Shyu, and G. Zhao, "Spatial-temporal motion information integration for action detection and recognition in non-static background," in *2013 IEEE 14th International Conference on Information Reuse Integration (IRI)*, pp. 626–633, Aug 2013.
2. H. Wang, A. Kläser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *CVPR 2011*, pp. 3169–3176, June 2011.
3. S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.
4. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

5. C. I. Orozco, M. E. Buemi, and J. J. Berlles, "Cnn-lstm architecture for action recognition in videos," in *I Simposio Argentino de Imágenes y Visión (SAIV 2019)-JAIIO 48 (Salta)*, 2019.
6. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
7. J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4694–4702, 2015.
8. M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
9. Y. Dauphin, H. de Vries, and Y. Bengio, "Rmsprop and equilibrated adaptive learning rates for non-convex optimization," in *NIPS*, 2015.
10. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.
11. K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, vol. abs/1212.0402, 2012.
12. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.
13. O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *European Conference on Computer Vision (ECCV)*, Oct. 2012.
14. Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *European Conference on Computer Vision*, pp. 425–438, Springer, 2012.
15. S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," *CoRR*, vol. abs/1511.04119, 2015.
16. Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "Videolstm convolves, attends and flows for action recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.
17. Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li, "Hierarchical attention network for action recognition in videos," *arXiv preprint arXiv:1607.06416*, 2016.
18. Y. Ye and Y. Tian, "Embedding sequential information into spatiotemporal features for action recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1110–1118, 2016.
19. B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with deeply transferred motion vector cnns," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2326–2339, 2018.
20. P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future in protein secondary structure prediction," *Bioinformatics*, vol. 15, no. 11, pp. 937–946, 1999.
21. D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014.