National Ilan University

DOI:10.6176/BCE.201912_(13).0005
Bulletin of College
of Engineering
http://engineering.niu.edu.tw

國立宜蘭大學工程學刊第 13 期，50-65 頁，2019 年 12 月
Bulletin of College of Engineering National Ilan University No. 13, PP. 50-65, December 2019

基於注意力模型神經網路之行為辨識

鄭又銘[1]　陳文輝[2]

1.國立台北科技大學自動化科技研究所研究生

2.國立台北科技大學自動化科技研究所教授

摘要

近年來各種行動感測器的應用蓬勃發展，尤其是在生活紀錄、健身追蹤及健康監測等領域。這些應用功能有賴於智慧型裝置中的嵌入式行動感測器來辨識人們的行為，雖然最近的研究有很大的進展，但因為人類活動行為廣泛以及變化性高，所以行為辨識仍然是一項具有挑戰性的任務，如何明確區分行為的功能就變得非常重要。在行為辨識領域，傳統上使用人工定義特徵的方式來解決問題，近期的研究使用深度卷積神經網路來提取特徵，但是人類行為是由複雜的時間序列組成，若能提取這種時間動態特徵，更能夠讓行為辨識的準確率提高，本研究應用注意力模型導入至行為辨識的研究當中，藉由最新的行為辨識深度學習模型(DeepConvLSTM)的架構中加入注意力機制(Transformer Encoder)，並使用公開的數據集PAMAP2來評估此模型，研究結果顯示，比較其他現行的方法，辨識準確率可以提升至84.1%。

關鍵字：注意力模型、神經網路、行為辨識

***通訊作者 E-mail: t106618513@ntut.edu.tw**

National Ilan University

DOI:10.6176/BCE.201912_(13).0005
Bulletin of College
of Engineering
http://engineering.niu.edu.tw

國立宜蘭大學工程學刊第 13 期，50-65 頁，2019 年 12 月
Bulletin of College of Engineering National Ilan University No. 13, PP. 50-65, December 2019

Using Attention Models on Neural Networks for Human Activity Recognition

Yu-Ming Cheng[1], Wen-Hui Chen[2]

1. Graduate Student, Graduate Institute of Automation Technology, National Taipei University of Technology.

2. Professor, Graduate Institute of Automation Technology, National Taipei University of Technology.

ABSTRACT

The application of motion sensor is flourishing in recent years, especially in the fields of life recording, fitness tracking and health monitoring. These applications depend on embedded motion sensors in smart phone to record human behavior. Although recent research has made great progress, human activity recognition (HAR) is still a challenging task because of the high variability. Artificial feature is the traditional way to solve the recognition problems. Recent researches have used deep convolutional neural networks to extract features, but human activity is composed of complex time series. It is better to extract temporal dynamic feature that can improve the accuracy of recognition. This paper introduces the attention model for the human activity recognition. We use attention model(Transformer Encoder) to the state-of-the-art deep learning HAR model（DeepConvLSTM） and evaluate the model on benchmark datasets. The research results show that compared with other state-of-the-art methods, the recognition accuracy can be improved to 84.7%.

Keywords: Attention Model, Neural Network, Human Activity Recognition

***Corresponding author E-mail: t106618513@ntut.edu.tw**

一、前言

在具有感測能力的行動裝置快速普及下，人體行為辨識產生了巨大的需求，生活紀錄、醫療保健、健身、工作等應用，都可以從行為辨識中獲得很大的幫助，如何利用技術進行行為辨識將變成很重要的課題，因此許多研究提出很多方法來辨識具有廣泛且變化性高的人類行為。在行為辨識中，分析以及模型化的都是資料都是時序資料，藉由觀察每個感測器中某個時間範圍的數據，來完成建構模型與辨識的基礎，傳統上使用滑動窗口方法(Bulling, Blanke, and Schiele, 2014)，此方法使用固定大小的窗口，來擷取每個感測器某段時間的數據，在許多使用深度學習的人體行為辨識研究中有著至關重要的作用，窗口長度是關鍵參數，會影響整個辨識的過程，因此需要很充足的領域知識，而且這種方法會將資料約束成單一固定的大小，在對具有持續且不同長短時間的行為進行建模時，效果可能並不理想。

另一種替代方法是使用順序模型，遞迴神經網路在行為辨識的應用中也有良好的辨識結果，長短期記憶模型(Long Short-term Memory, LSTM) (Hochreiter and Schmidhuber, 1997) 可以學習具有時間序列的資料。但是假設遙遠的某個資料實際上會影響當前的資料其實是不合理的，因為當前的行為跟很久以前的行為通常是沒有甚麼關聯的。對於模型而言，以前的哪些行為資料才是重要且需要學習的，以及模型是否可以自動擷取出這些資料。因此本研究希望建立一個模型，可以自動學習過去相關的行為數據。

本研究將注意力模型應用於行為辨識，注意力模型可以幫助模型去學習一組輸入數據的權重，這些權重可以表示成對於學習時的相對重要性。在行為分析上，這種模型會學習先前用於分析的感測器讀數樣本的貢獻度，使模型能夠生成時間軸上的資料權重分布，藉由權重可以觀察出模型分類決策的相關資料。

本研究將注意力模型加入至最新的行為辨識模型中，進而觀察注意力模型的潛力，使用行為辨識為標準的資料庫來進行評估。本研究目的為：(1) 建立一個使用感測器數據進行行為辨識之神經網路模型 (2) 比較本研究所使用之注意力模型與現行使用之行

為辨識模型的差異，以及使用資料集評估是否能提高辨識率。

二、文獻探討

深度神經網路在圖形辨識問題上是一項非常強大的工具，近期研究人體行為辨識主要都關注在卷積神經網路(Convolution Neural Network, CNN)與長短期記憶單元(LSTM)。

2-1 卷積神經網路

卷積神經網路(CNN)(LeCun and Bengio, 1998)是參考人體大腦視覺組織所發展出來的一種網路，具有特徵擷取的能力，經由堆疊的卷積運算可以產生抽象化的特徵，這些特徵可以讓模型有能力去進行辨識。卷積神經網路由一個或多個卷積層及池化層所組成，卷積層有一組濾波器可以擷取輸入資料的特徵，而池化層可以壓縮及保留特徵，經過卷積以及池化作用後會接完全連接層來組合所有輸入進來的特徵，在進行分類，這種分層化的組織模型在圖形辨識上有很好的效果。

研究(Yang, Nguyen, and San et al, 2015)使用 CNN 方法應用於行為辨識上，將加速度計的三軸數據分別使用一層卷積層濾波器來擷取特徵，再使用最大池化層來保留以及壓縮特徵，池化效果可以縮小數據的尺寸並減少計算時間，之後將三軸所擷取出來的特徵串接在一起，並輸入至二個完全連接的隱藏層，最後再傳入 Softmax 分類器來進行分類。整個模型使用前向傳播以及反向傳播演算法來計算 CNN 模型的權重，並將目標函數最小化。此研究也使用權值衰減(Weight Decay)、Momentem、以及 Dropout 等正規化方法來改進模型，研究結果顯示這種 CNN 方法優於其他當時的機器學習方法。

2-2 長短期記憶神經網路

長短期記憶神經網路(Hochreiter and Schmidhuber, 1997)是一種遞迴式神經網路，此網路具有記憶單元，可以儲存和輸出相關的資訊，因此適合處理具有時間序列較長的問題。長短期記憶單元利用閘控的概念，根據啟動不同的閘來更新單元的狀態，寫入資料使用輸入閘(Input Gate)、讀取資料使用輸出閘(Output Gate)，將資料清除使用遺忘閘(Forget Gate)，整體架構如圖1所示，數學表示式如式(1)至式(5)所示。

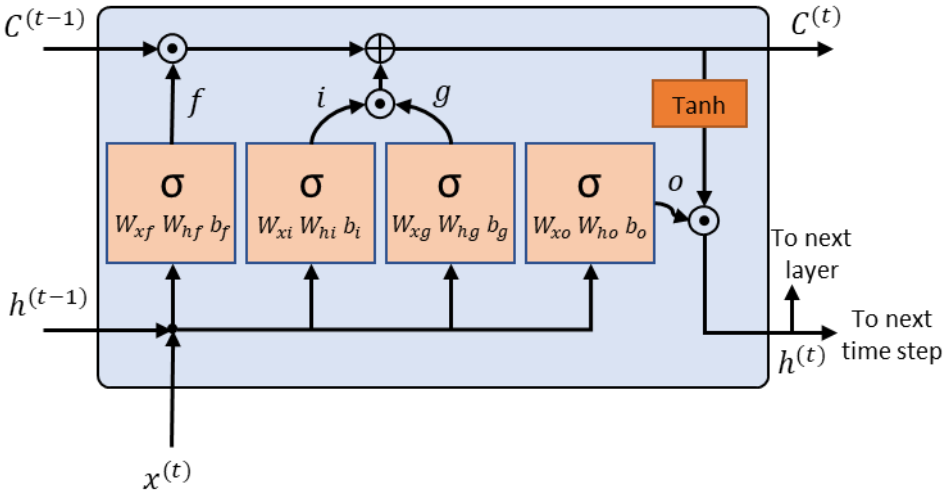

圖 1 長短期記憶單元(Raschka, 2015)

$$f_t = \sigma_f\left(w_{xf}x^{(t)} + w_{hf}h^{(t-1)} + b_f\right) \quad (1)$$

$$i_t = \sigma_i\left(w_{xi}x^{(t)} + w_{hi}h^{(t-1)} + b_i\right) \quad (2)$$

$$o_t = \sigma_o\left(w_{xo}x^{(t)} + w_{ho}h^{(t-1)} + b_o\right) \quad (3)$$

$$c_t = f_t c_{t-1} + i_t \sigma_c\left(w_{xg}x^{(t)} + w_{hg}h^{(t-1)} + b_{og}\right) \quad (4)$$

$$h^{(t)} = o_t \sigma_h(c_t) \quad (5)$$

其中i代表輸入閘，f代表遺忘閘，o代表輸出閘，c代表單元狀態，σ項代表非線性方程式，$x^{(t)}$表示記憶單元在t時間的輸入，w為權重，b為偏誤。長短期記憶單元因為有記憶的功能，會比標準的遞迴神經網路更適合處理具有較長時間的資料。

2-3 DeepConvLSTM神經網路

DeepConvLSTM架構的神經網路(Ordóñez and Roggen, 2016)，結合了卷積層以及長短期記憶網路的功能。卷積層作用在擷取輸入數據的特徵。長短期記憶單元可以為這些擷取出來的特徵，做成具有時間性的模型。DeepConvLSTM與傳統卷積神經網路模型的結構上有些差異，主要區別是在於卷積層後的完全連結層部分，卷積神經網路是使用完全連結層，傳統卷積神經網路架構如圖2所示；而DeepConvLSTM是使用長短期記憶單元來取代完全連結層，這些單元是遞迴式的運作而非完全連接的，因此比較適合使用在具

有時間序列的資料上， DeepConvLSTM神經網路架構如圖3所示。

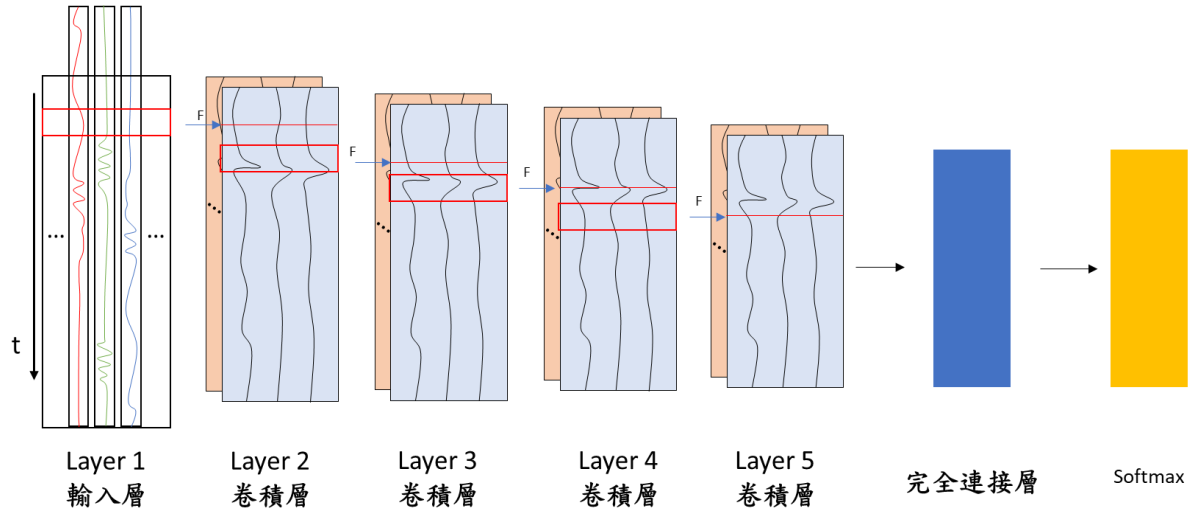

圖 2 傳統卷積神經網路架構

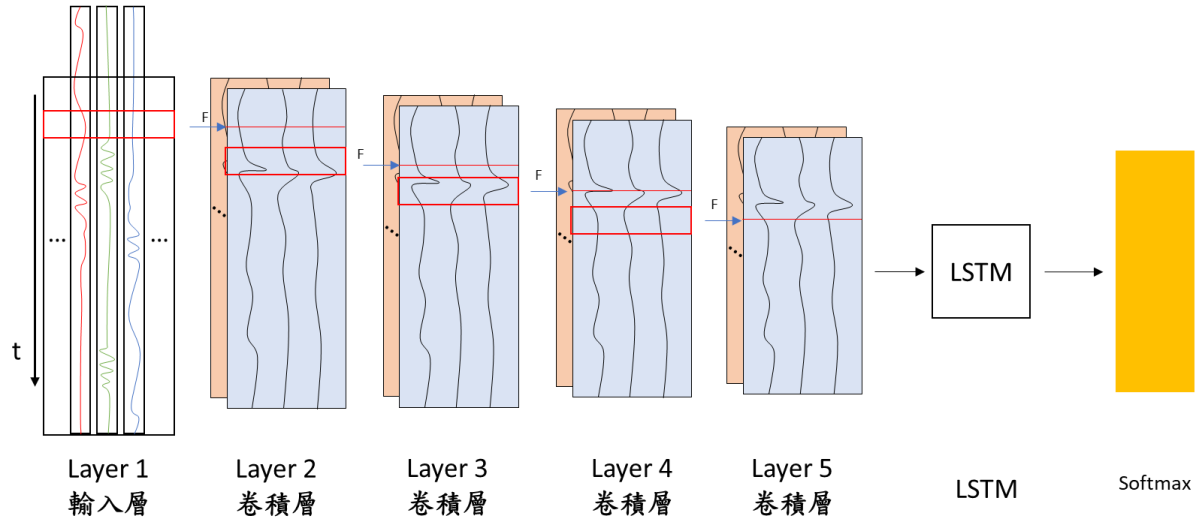

圖 3 DeepConvLSTM架構(Ordóñez and Roggen, 2016)

2-4 注意力模型

在自然語言處理領域，導入了注意力模型來解決語音標記的問題(Kumar, Irsoy, Ondruska et al, 2016)，使用線性層來學習一組權重，會將d維的k個向量映射到一組一維分數，這些分數再傳遞給softmax函數給出一組個數為*k*個權重集。近期有研究使用DeepConvLSTM導入基礎的注意力模型來建立新的人體行為辨識模型(Murahari and Ploetz, 2018)，辨識率相較於DeepConvLSTM有微幅的提升，完整網路架構如圖4所示。

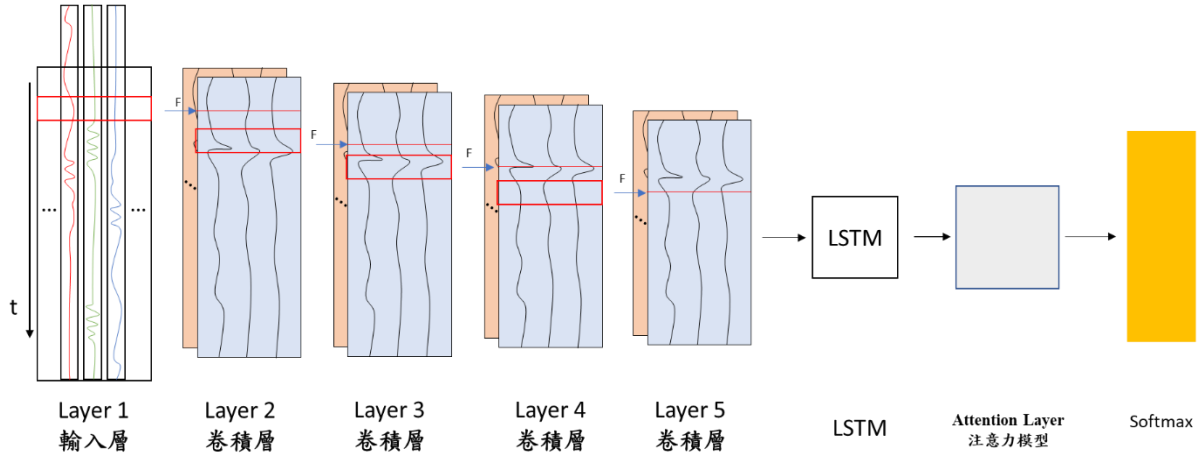

圖 4 注意力模型整合DeepConvLSTM神經網路架構(Murahari and Ploetz, 2018)

2-5 Transformer

Transformer是Google團隊所發表的新形態網路架構(Vaswani, Shazeer, and Parmar et al, 2017)，在Encoder-Decoder的架構中導入了注意力模型，在機器翻譯的任務中，辨識率獲得顯著的提升。

本研究將使用DeepConvLSTM並導入Transformer Encoder，Encoder中含有注意力機制以及LayerNorm機制(Ba, Kiros, and Hinton, 2016)，來建立新的人體行為辨識模型。

三、研究方法

3-1 研究模型

本研究的模型將使用Transformer中的Encoder架構整合於DeepConvLSTM，Encoder Layer中含有二個子層，第一個子層是一個多頭自我注意力機制層，第二個子層是一個全連接前饋網路。在每個子層的輸出都與原先的輸入相加，並使用LayerNorm標準化機制。Transformer Encoder Layer架構如圖5所示。

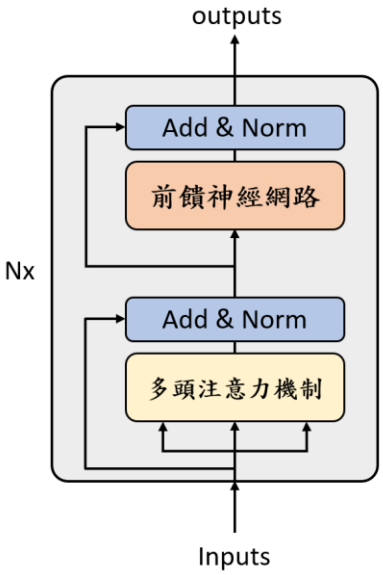

圖 5 Transformer Encoder Layer架構(Ashish et al, 2017)

在自我注意力機制中，會將輸入映射到一組Query、Key以及Value，如果使用多頭型態則會對Q,K,V做多次不同的映射，之後經過自我注意力模型的計算。

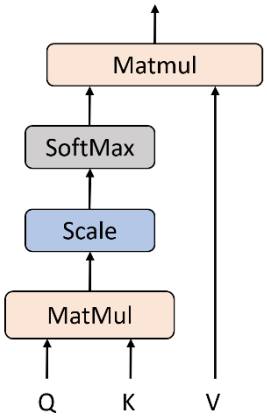

圖 6 自我注意力模型架構(Ashish et al, 2017)

自我注意力模型使用Dot-Product Attention，計算Q與K的點積，並使用softmax函數來獲得V的權重，公式如式(6)。

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (6)$$

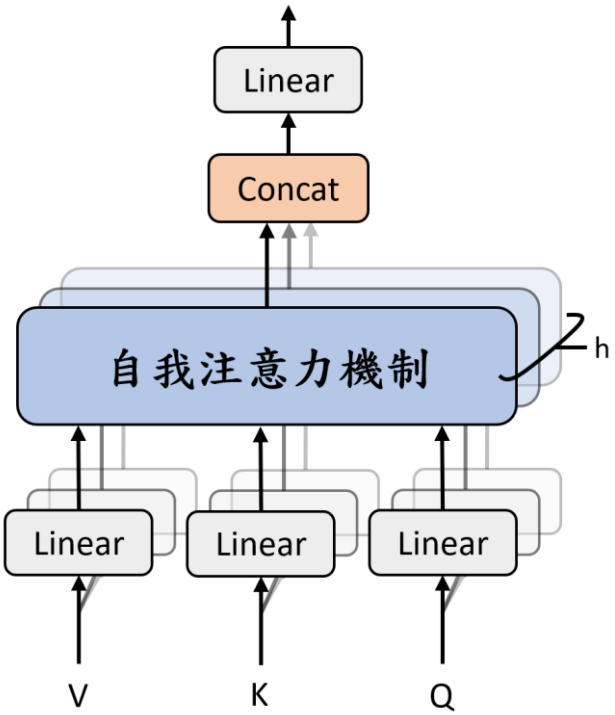

圖 7 多頭注意模型架構(Ashish et al, 2017)

多頭模式下的注意力模型會將V、K以及Q映射h次，可以在不同的位置上進行注意力機制，最後將它們連接起來並經過線性層轉換，詳見式(7)(8)。

$$MultiHead(Q,K,V) = Concat(head_1,\dots,head_h)W^o \tag{7}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

本研究將對h以及Encoder的層數等參數進行最佳化實驗，以尋找辨識率最佳的模型，並與其他研究的神經網路模型來進行比較。

研究模型使用DeepConvLSTM神經網路，並導入Transformer Encoder用於改進注意力模型，DeepConvLSTM部分會先對數據進行特徵擷取以及長短期記憶模型對時間序列的運算，之後的輸出再輸入至Transformer Encoder中進行注意力機制，整體架構如圖

8與表1所示。

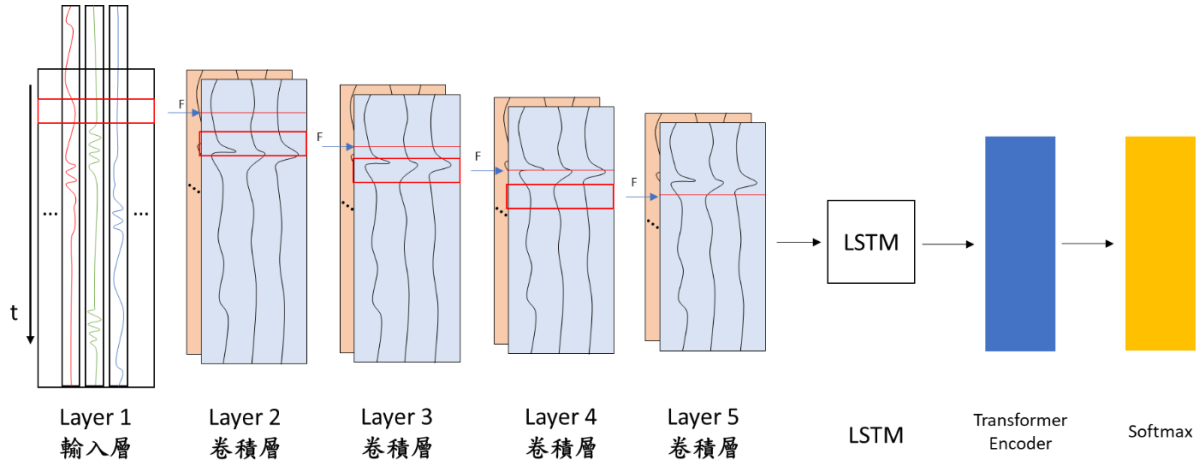

圖 8 研究模型架構

表 1 模型結構與參數

序列	類別	參數與描述
1	輸入層	
2	卷積層 Conv2d	64 filter，size(5,1)
3	線性整流單元 ReLU	
4	卷積層 Conv2d	64 filter，size(5,1)
5	線性整流單元 ReLU	
6	卷積層 Conv2d	64 filter，size(5,1)
7	線性整流單元 ReLU	
8	卷積層 Conv2d	64 filter，size(5,1)
9	線性整流單元 ReLU	
10	長短期記憶單元 LSTM	128 hidden size
12	長短期記憶單元 LSTM	128 hidden size
13	LSTM Dropout	0.6
14	Transformer Encoder	見圖 5
15	Transformer Encoder	見圖 5
16	Softmax 層	

卷積層的設置為64個filter，filter大小設置為(5,1) ，步進值為1，並在各卷積層後

使用線性整流函數(ReLU)。長短期記憶層設置為128個單元，並使用Dropout設置為0.6，之後連結二個Transformer Encoder層，最後的輸出經由Softmax層來進行分類。

3-2 實驗環境與資料集

實驗過程主要使用Pytorch深度學習框架來進行訓練，數據會使用滑動窗口方法來做處理，每幀之間會有50%的重疊。使用Cross-Entropy Loss 訓練模型，模型的優化器(Optimizer)使用RMSprop，學習率設定為0.001，在模型學習上都有對學習衰減率與dropout(Srivastava, Hinton, Krizhevsky et al, 2014)進行優化，衰減率為0.95，dropout值設為0.6，Batch size設置為100，以上實驗設置皆與研究(Murahari and Ploetz, 2018)之設置相同，以進行比較。

本研究使用人類行為辨識領域的標準數據集PAMAP2資料集(Reiss and Stricker, 2012)，包含12種日常活動與運動項目，由九個測試人員在身體上穿戴三個慣性感測器，進行感測與錄製，採樣率為100Hz。本研究將第五個與第六個人員所感測的數據設定為測試資料集，其餘人員的感測數據設定為訓練資料集。

在模型上使用五種架構的模型來進行對比與評估，第一種模型使用CNN、第二種是DeepConvLSTM、第三種模型使用LSTM + Attention、第四種是DeepConvLSTM + Attention，而第五種為本研究建構之神經網路模型，相關模型結構如表1。

3-3 模型驗證

訓練過程中，epochs設為100，每經過一次epoch會使用測試數據集來進行模型驗證，辨識率的設置上使用F1分數(F1 Score)。

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

Precision是精準度，公式如式(10)所示，recall是召回率，公式如式(11)所示。

$$precision = \frac{TP}{TP+FP} \tag{10}$$

$$recall = \frac{TP}{FN+TP} \tag{11}$$

TP(True Positive)，數學表示預測結果為正，實際也為正。FP(False Positive)，數學表示預測結果為正，實際為負。FN(False Negative)，數學表示預測結果為負，實際為正。Recall越高代表對正樣本的識別能力越強，precision越高表示對負樣本的區分能力越強。F1分數則是兩種指標的混合型態，越高則代表模型越穩健。

四、結果與討論

所有實驗使用AMD Ryzen 5 2600處理器、16GB DDR4記憶體以及Nvidia GeForce GTX 2070 圖形顯示卡。辨識結果如表2。

表 2 辨識結果

模型種類	F1 Score
CNN (Ordóñez and Roggen, 2016)	76.6%
DeepConvLSTM (Ordóñez and Roggen, 2016)	80.7%
LSTM + Attention (Murahari and Ploetz, 2018)	75.4%
DeepConvLSTM + Attention (Murahari and Ploetz, 2018)	82.3%
DeepConvLSTM + Transfomer Encoder	**84.1%**

表2的辨識結果顯示，在傳統CNN模型得到的辨識率是76.6%，而使用DeepConvLSTM模型，能夠擷取特徵並對特徵做時間序列的建模，可以得到80.7%的辨識率，LSTM + Attention在辨識率上得到75.4%，使用DeepConvLSTM + Attention，辨識率可以提升至82.3%，而本研究所使用的Transformer Encoder整合DeepConvLSTM，能夠

有效提升辨識率至84.1%。

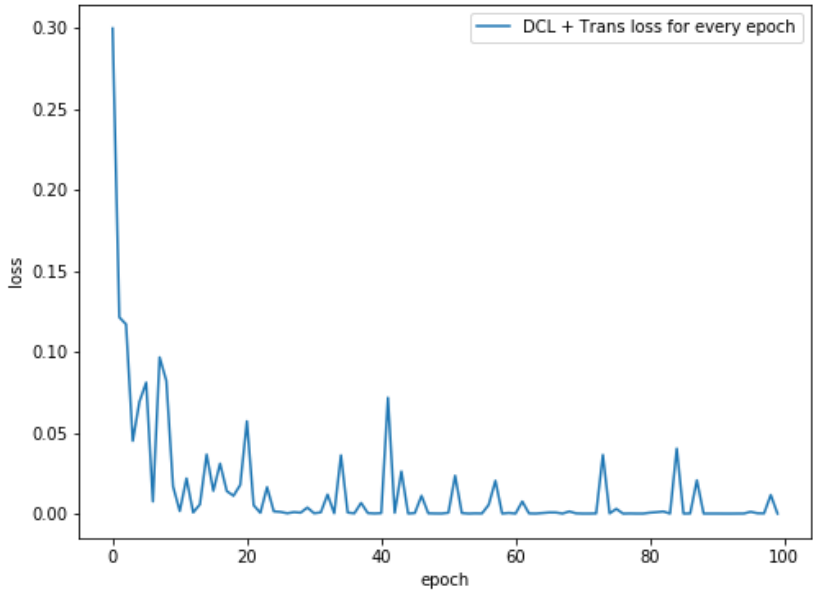

圖 9　Loss值變化曲線

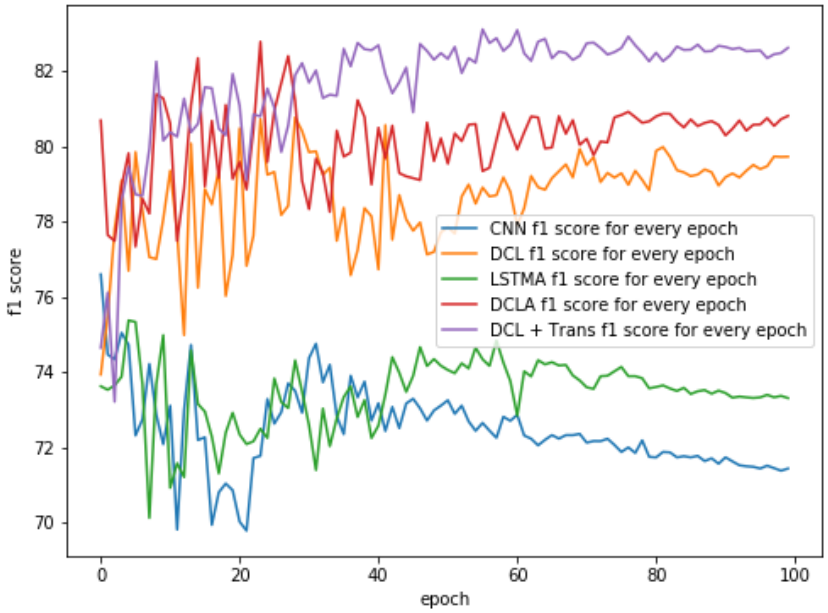

圖 10　F1分數變化曲線

根據圖9顯示，模型在訓練上都能夠有效收斂。從圖10的F1分數變化曲線上，本研

究的模型與其他研究比較，在模型收斂後可以有效提高辨識率。

在Transformer Encoder的模型參數最佳化問題上，研究結果如表3所示。

表 3 最佳化參數結果

Transformer Encoder 層數	多頭數量	F1 Score
1	8	83.3%
	4	83.0%
	2	83.2%
2	8	**84.1%**
	4	83.1%
	2	82.1%

在Transformer Encoder層中的參數最佳化研究中，根據表3顯示，在層數為二以及多頭數量為八時，辨識率達到84.1%為最高，因多頭注意力的功能會對V、K以及Q映射h次，進而對資料的不同位置產生注意力機制，能夠有效地提升辨識率。

五、結論

本研究基於 Transformer Encoder 多頭注意力模型整合 DeepConvLSTM 神經網路架構，提出一種針對感測器訊號進行行為辨識之方法，並使用 PAMAP2 公開資料集來進行模型訓練與驗證，研究結果顯示，辨識率優於之前研究論文所提出的深度神經網路方法，能夠將辨識率 F1 分數提高 2%~8%。

本研究所使用的注意力模型Transformer Encoder能有效地縮小且集中模型需要注意的感測器數據，Transformer Encoder層中的參數最佳化研究結果顯示，在PAMAP2資料集中，使用多頭數量為8以及層數為二時，可以達到最高辨識率84.1%，因

Transformer Encoder中的多頭注意力的功能會對V、K以及Q映射多次，進而對資料的不同位置產生注意力機制，不僅只是依賴LSTM的最後一個隱藏狀態，模型可以更有效的擷取時間序列資料上的關聯性，並提高辨識率。

參考文獻

Bulling, A., Blanke, U., Schiele, B., 2014, “A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors,” *ACM Comput. Surv.* 46, 3, Article 33, Jan. 2014.

Hochreiter, S., Schmidhuber, J., 1997, “Long short-termmemory,” *Neural computation* 9, 8 (1997), pp. 1735-1780.

LeCun, Y., Bengio, Y., 1998, “Convolutional Networks for Images, Speech, and Time Series,” *The Handbook of Brain Theory and Neural Networks*, MIT press: Cambridge, MA, USA, 1998, pp. 255-258.

Yang, J., Nguyen, M.N., San, P.P., Li, X., Krishnaswamy, S., 2015, “Deep Convolutional Neural Networks on Multichannel Time Series for Human Activity Recognition,” *In Proc. IJCAI.*

Raschka, S., 2015, “Python machine learning,” Birmingham, UK: Packt Publishing Print.

Ordóñez, F.J. and Roggen, D., 2016, “Deep convolutional and lstm recurrent neural networks for multimodal wearableactivity recognition,” *Sensors* 16, 1, 2016, 115.

Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulrajani, I., Zhong, V., Paulus, R., Socher, R., 2016, “Ask me anything: Dynamic memory networks for natural language processing,” *In Proc. ICML*.

Murahari, V. S., Ploetz, T., 2018, “On Attention Models for Human Activity Recognition,” *arXiv* preprint arXiv:1805.07648.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017, “Attention is all you need,” *In Advances in Neural Information Processing Systems*, pp. 6000-6010.

Ba, J.L., Kiros, J.R., Hinton, G.E., 2016, “Layer normalization,” *arXiv* preprint arXiv:1607.06450.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014, “Dropout: A simple way to prevent neural networks from overfitting,” *JMLR* 15, 1 (2014), pp. 1929-1958.

Reiss, A., Stricker, D., 2012, “Introducing a new benchmarked dataset for activity monitoring,” *In Proc. ISWC.*