Boosting Inertial-Based Human Activity Recognition With Transformers

AI VIETNAM - RESEARCH TEAM

September 24, 2023

Date of publication:	04/02/2021
Authors:	Yoli Shavit, Itzik Klein
Sources:	2021 IEEE
Source code:	https://github.com/yolish/har-with-imu-transformer
Keywords:	Human activity recognition, smartphone location recognition, inertial sensors, pedestrian dead reckoning, convolutional neural networks, Transformers, sequence analysis
Summary by:	Quy Nguyen Dinh

1. **Purpose/outputs:**

   - The paper "Boosting Inertial-Based Human Activity Recognition With Transformers" focuses on improving human activity recognition based on inertial sensor data. The main goal of this research is to introduce a Transformer-based activity recognition model, providing a more general and efficient learning framework.
   - In the context where current machine learning methods for activity recognition from inertial data often use Convolutional Neural Network (CNN) or Long Short-Term Memory (LSTM) architectures, Transformers have been proven to outperform these architectures for sequence analysis tasks.
   - This paper has experimented and evaluated the model on multiple datasets, with over 27 hours of inertial data recordings collected by 91 users, representing various user activity scenarios with varying difficulty. The results show that the proposed method consistently achieves better accuracy and generalizes better across all tested datasets and scenarios.

2. **Contributions:**

   - The paper proposes a framework for classifying inertial data using Transformers. This proposed method is the first to introduce a Transformer-based architecture as a general framework for general activity recognition in both Human Activity Recognition (HAR) and Smartphone Location Recognition (SLR) tasks.
   - The paper presents a design for a Transformer network architecture that handles inertial measurements, along with a publicly available implementation.
   - The paper evaluated the proposed framework for three commonly used classification task: HAR, SLR, and a combination of the two - Smartphone and Human Activity Recognition (SHAR). It demonstrated a consistent improvement in accuracy and better generalization across datasets.

3. **Inputs:**

   It includes 3 datasets:

   - **SLR dataset:** Includes five different smartphone locations. This dataset was created by combining six different SLR datasets. The location of the smartphone is at least in one of five locations: Texting, Pocket, Swing, Talking, and Body, while the user is walking.
   - **HAR dataset:** Includes six different human activities: Walking, Jogging, Sitting, Standing, Stairs down, and Stairs up.
   - **SHAR dataset:** Includes 21 classes, containing data with combined SLR and HAR class labels, referred to as SHAR. For example, the class 'Walking Pocket' refers to a scenario of a person walking (HAR) with the smartphone placed in their pocket (SLR).

   In total, the three datasets SLR, HAR, and SHAR contain 27.76 hours of recordings made by 91 people. Each dataset contains many different files. Each file has the name of the user who made the recording and a description of its type (e.g., user1 walking texting), and can have a different duration. When creating the unified dataset, all files from all users were merged into a single file.

4. **Methodology:**

   - **Problem Definition:** The paper presents a Transformer-based activity recognition model, providing an improved and general framework for learning activity recognition tasks.
   - **Input Data Format:** The Inertial Measurement Unit (IMU) measures the specific force $\mathbf{f} \in \mathbb{R}^3$ and the angular velocity vector $\mathbf{w} \in \mathbb{R}^3$ over time. These two outputs are combined such that each sample $\mathbf{S} \in \mathbb{R}^{k \times 6}$ represents a sequence of k measurements - applying a window of size k.
   - **Network Architecture:**

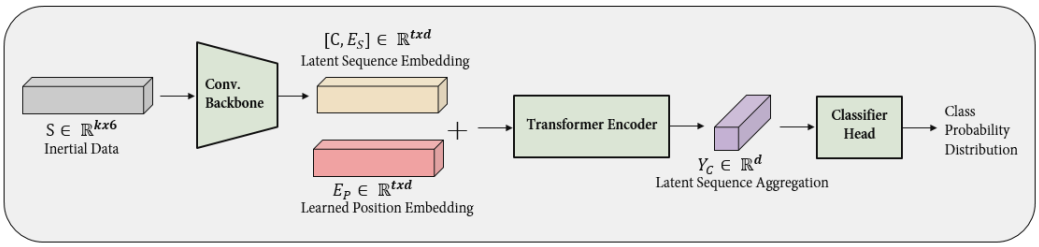


Figure 1: Proposed network architecture

A series of four 1D convolution layers with GELU non-linearity are applied to each sample $\mathbf{S} \in \mathbb{R}^{k \times 6}$ to create an embedded sequence $\mathbf{E_S}$ (latent features). A class token $\mathbf{C} \in \mathbb{R}^d$ is added before the embedded sequence. Each $\mathbf{E_{Pi}} \in \mathbb{R}^d$ for each position $\mathbf{P_i}$ in the sequence is learned and added to the representation of the hidden sequence. Hence, the input to the Transformer Encoder block is: $\mathbf{Z_0} = [\mathbf{C}, \mathbf{Es}] + \mathbf{E_S} \in \mathbb{R}^{t \times d}$ with $t = k + 1$.

A standard Encoder structure is applied, consisting of L layers, each layer includes a Multi-Head Attention (MHA) and a Multi-Layer Perceptron (MLP). In the MLP block, there are two fully connected (FC) layers with hidden dimensions $2 * d$ and GELU non-linearity. For three sequences of length t and dimension d, specifically a query $\mathbf{Q} \in \mathbb{R}^{t \times d}$, a key $\mathbf{K} \in \mathbb{R}^{t \times d}$ and a value $\mathbf{V} \in \mathbb{R}^{t \times d}$, each head h compute a weighted aggregation of $\mathbf{V}$ according to $\mathbf{Q}$:

$$\mathbf{h(Q,\ K,\ V)} = \text{softmax}\left(\frac{\mathbf{Q_h K_h^T}}{\sqrt{d_k}}\right)\mathbf{V_h} \in \mathbb{R}^{t\times d'}, \text{ in that:}$$

$$\mathbf{Q_h} = \mathbf{Q W_h^Q} \in \mathbb{R}^{t\times d'}$$

$$\mathbf{K_h} = \mathbf{K W_h^K} \in \mathbb{R}^{t\times d'}$$

$$\mathbf{V_h} = \mathbf{V W_h^V} \in \mathbb{R}^{t\times d'}$$

The resulting representation is a weighted aggregation of the sequence at each position, based on its relative importance to other positions.

For each layer $l, l = 1..L$ in the Encoder block, with LN being LayerNorm, we have:

$$\mathbf{Z'_l} = sMHA(LN(\mathbf{Z_{l-1}})) + \mathbf{Z_{l-1}} \in \mathbb{R}^{t\times d}$$

$$\mathbf{Z_l} = MLP(LN(\mathbf{Z'})) + \mathbf{Z'} \in \mathbb{R}^{t\times d}$$

The output of the Encoder block at the position of the class token represents a temporally aware aggregation of the input sequence:

$$\mathbf{Y_C} = \mathbf{Z_L[0]} \in \mathbb{R}^d$$

$\mathbf{Y_C}$ is provided as an input for a classifier head, consisting of LN and FC layers with GELU non-linearity and Dropout, reducing the dimension to $\frac{d}{4}$. A second FC layer maps $\frac{d}{4}$ to the number of classes. A Log SoftMax is applied on the output vector in conjunction with Negative Log Likelihood (NLL) loss to learn a multi-label classification task.

5. **Results:**

- **Experimental Setup:**

  The paper compares the proposed method with a CNN model that has been proven to perform well on various SLR tasks. Each dataset is split into a train and test set, with 85% of the samples selected for the train set. Both models (the CNN model and the proposed model) use the Adam optimization algorithm, with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-10}$, a batch size of 128 and a weight decay of 10^{-4}, an initial learning rate of $\lambda = 10^{-4}$ which is halved every m epochs depending on the experiment. Each model is trained for 30 epochs for smaller datasets and 80 epochs for larger datasets. For convenience, from here on the CNN model and the proposed method are referred to as IMU-CNN and IMU-Transformer.

- **The effect of window size:**

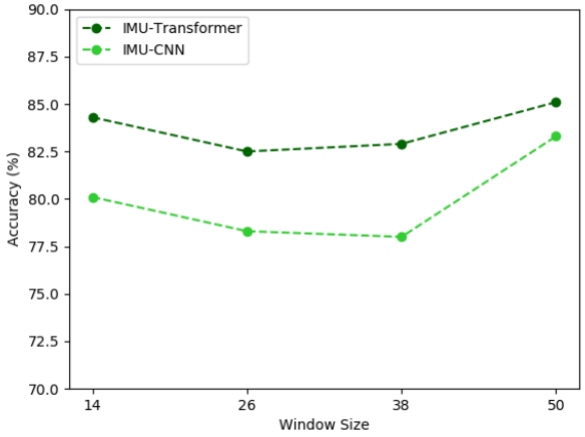


Figure 2: Accuracy on the SHAR dataset with different window sizes

- **Accuracy on different datasets:**

Experiment	Window Size	IMU-CNN Accuracy	IMU-Transformer Accuracy
SLR	50	96.5%	97.4%
HAR	50	86.2%	89.6%
SHAR	50	83.3%	85.1%
Overall	50	88.7%	90.7%

Figure 3: Compare accuracy on different datasets

6. **Limitations:**

   The paper does not specifically mention limitations, but there may be some issues that need further clarification, such as whether the model works well with noisy inertial sensor data or under inappropriate conditions.

7. **Future researches:**

   Future research directions could involve improving the Transformer model to increase accuracy in recognizing human activity from inertial sensor data, or potentially applying this model to various types of data and different contexts.

Tăng cường nhận dạng hoạt động con người dựa trên cảm biến quán tính với Transformers

AI VIETNAM - RESEARCH TEAM

Ngày 24 tháng 9 năm 2023

Ngày công bố:	02/04/2021
Tác giả:	Yoli Shavit, Itzik Klein
Nguồn:	2021 IEEE
Mã nguồn:	https://github.com/yolish/har-with-imu-transformer
Từ khóa:	Nhận dạng hoạt động con người, nhận dạng vị trí điện thoại thông minh, cảm biến quán tính, mạng nơ ron tích chập, phân tích chuỗi, pedestrian dead reckoning, Transformers
Người tóm tắt:	Nguyễn Đình Quý

1. **Mục đích:**

   - Bài báo "Tăng cường nhận dạng hoạt động con người dựa trên cảm biến quán tính với Transformers" tập trung vào việc cải thiện nhận dạng hoạt động con người (activity recognition) dựa trên dữ liệu từ cảm biến quán tính (inertial sensor data). Mục tiêu chính của nghiên cứu này là giới thiệu một mô hình nhận dạng hoạt động dựa trên cấu trúc Transformers, cung cấp một khuôn khổ học tập (learning framework) tổng quát và hiệu quả.
   - Trong bối cảnh các phương pháp học máy hiện tại cho nhận dạng hoạt động từ dữ liệu quán tính thường sử dụng các kiến trúc mạng nơ ron tích chập (CNN) hoặc mạng trí nhớ ngắn hạn định hướng dài hạn (LSTM), Transformers đã được chứng minh là vượt trội hơn những kiến trúc này cho các tác vụ phân tích chuỗi.
   - Bài báo này đã thử nghiệm và đánh giá mô hình trên nhiều tập dữ liệu (multiple datasets), với hơn 27 giờ ghi lại (recordings) dữ liệu quán tính (inertial data) thu thập từ 91 người dùng, đại diện cho các kịch bản hoạt động (activity scenarios) khác nhau của người dùng với các độ khó khác nhau. Kết quả cho thấy phương pháp đề xuất đạt được độ chính xác tốt hơn và mang tính tổng quát hóa tốt hơn trên tất cả các tập dữ liệu đã được kiểm tra (examined datasets) và các kịch bản khác nhau.

2. **Đóng góp:**

   - Bài báo đã giới thiệu một khuôn mẫu (framework) cho việc phân loại dữ liệu quán tính sử dụng Transformers. Phương pháp được đề xuất lần đầu tiên giới thiệu một kiến trúc dựa trên Transformers như một khuôn mẫu tổng quát cho việc nhận dạng hoạt động trong các tác vụ HAR (Nhận dạng hoạt động con người - Human Activity Recognition) và SLR (Nhận dạng vị trí điện thoại thông minh - Smartphone Location Recognition).
   - Bài báo đã thiết kế một kiến trúc mạng Transformer để xử lý các đo lường quán tính (inertial measurements), cùng với việc cung cấp mã nguồn mở của kiến trúc.

- Bài báo đánh giá khuôn mẫu được đề xuất cho ba tác vụ phân loại: HAR, SLR và sự kết hợp của cả hai - SHAR (Nhận dạng điện thoại thông minh và hoạt động con người - Smartphone and Human Activity Recognition), cho thấy sự cải thiện đồng nhất về độ chính xác và khả năng tổng quát hóa tốt hơn trên các tập dữ liệu.

3. **Dữ liệu đầu vào:**

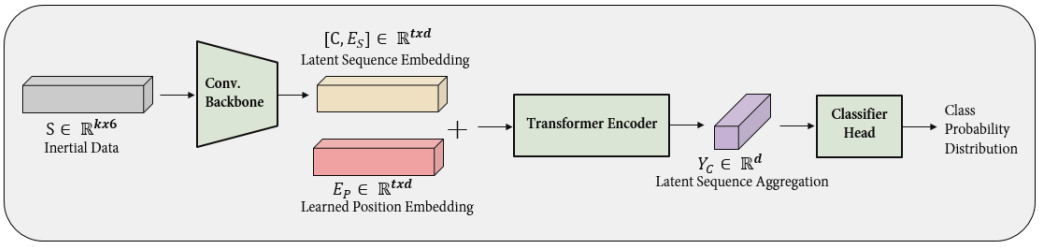
   Bao gồm 3 bộ dữ liệu:

   - **Bộ dữ liệu SLR:** bao gồm năm vị trí điện thoại (smartphone locations) khác nhau. Bộ dữ liệu này được tạo ra bằng cách kết hợp sáu bộ dữ liệu SLR khác nhau. Vị trí điện thoại thông minh ít nhất ở một trong năm vị trí: Nhắn tin (Texting), Túi (Pocket), Đu đưa (Swing), Nói chuyện (Talking) và Trên người (Body), trong khi người dùng đang đi bộ.
   - **Bộ dữ liệu HAR:** Bao gồm sáu hành động khác nhau của con người: Đi bộ (Waling), Chạy bộ (Jogging), Ngồi (Sitting), Đứng (Standing), Đi xuống cầu thang (Stairs down) và Đi lên cầu thang (Stairs up).
   - **Bộ dữ liệu SHAR:** Bao gồm 21 lớp, chứa dữ liệu với nhãn lớp SLR và HAR kết hợp với nhau, được gọi là SHAR. Ví dụ, lớp 'Walking Pocket' đề cập đến một trường hợp của một người đi bộ (HAR) với chiếc điện thoại thông minh đặt trong túi của họ (SLR).

   Tổng cộng, ba bộ dữ liệu SLR, HAR và SHAR chứa 27,76 giờ các bản ghi do 91 người tạo ra. Mỗi bộ dữ liệu chứa nhiều tệp (files) khác nhau. Mỗi tệp có tên của người dùng đã tạo bản ghi và mô tả loại của nó (ví dụ: user1 walking texting), và có thời lượng khác nhau. Khi tạo ra bộ dữ liệu hợp nhất, tất cả các tệp từ tất cả người dùng đã được hợp nhất thành một tệp duy nhất.

4. **Phương pháp luận:**

   - **Định nghĩa vấn đề:** Bài báo trình bày một mô hình nhận dạng hoạt động dựa trên cấu trúc Transformers, cung cấp một khuôn mẫu cải tiến và tổng quát cho việc học các tác vụ nhận dạng hoạt động.
   - **Định dạng dữ liệu đầu vào:** Bộ phận đo lường quán tính (IMU) đo lực (specific force) $\mathbf{f} \in \mathbb{R}^3$ và vectơ vận tốc góc (angular velocity vector) $\mathbf{w} \in \mathbb{R}^3$ theo thời gian. Hai đầu ra này được kết hợp với nhau sao cho mỗi mẫu (sample) $\mathbf{S} \in \mathbb{R}^{k \times 6}$ đại diện cho một chuỗi k phép đo - áp dụng cửa sổ (window) có kích thước k.
   - **Kiến trúc mạng:**



Hình 1: Kiến trúc mạng đề xuất

Áp dụng một loạt 4 lớp tích chập 1 chiều (1D convolutions) với hàm GELU phi tuyến lên mỗi mẫu $\mathbf{S} \in \mathbb{R}^{k \times 6}$ để tạo ra 1 chuỗi nhúng $\mathbf{E_S}$ (đặc trưng ẩn – latent features). Class token $\mathbf{C} \in \mathbb{R}^d$ được thêm vào trước chuỗi nhúng. Mỗi $\mathbf{E_{Pi}} \in \mathbb{R}^d$ cho mỗi vị trí $\mathbf{P_i}$ trong chuỗi được học và thêm vào biểu diễn của chuỗi ẩn. Từ đó, input đầu vào cho khối Transformer Encoder là: $\mathbf{Z_0 = [C, Es] + E_S} \in \mathbb{R}^{t \times d}$ với $t = k + 1$.

Một cấu trúc Encoder tiêu chuẩn được áp dụng, gồm L tầng, mỗi tầng bao gồm cơ chế tự chú ý đa đầu (MHA) và (MLP). Trong khối MLP, có hai lớp kết nối đầy đủ (FC) với số chiều ẩn là $2 * d$ và hàm phi tuyến GELU. Với ba chuỗi có độ dài t và số chiều d, cụ thể là một câu hỏi (query) $\mathbf{Q} \in \mathbb{R}^{t \times d}$, một khóa (key) $\mathbf{K} \in \mathbb{R}^{t \times d}$ và một giá trị (value) $\mathbf{V} \in \mathbb{R}^{t \times d}$, mỗi đầu h sẽ tính toán tổng hợp trọng số của $\mathbf{V}$ theo $\mathbf{Q}$:

$$\mathbf{h(Q,\ K,\ V)} = \text{softmax}\left(\frac{\mathbf{Q_h K_h^T}}{\sqrt{d_k}}\right)\mathbf{V_h} \in \mathbb{R}^{t \times d'}, \text{ trong đó:}$$

$$\mathbf{Q_h} = \mathbf{Q W_h^Q} \in \mathbb{R}^{t \times d'}$$

$$\mathbf{K_h} = \mathbf{K W_h^K} \in \mathbb{R}^{t \times d'}$$

$$\mathbf{V_h} = \mathbf{V W_h^V} \in \mathbb{R}^{t \times d'}$$

Biểu diễn kết quả được cập nhật là tổng hợp có trong số của chuỗi ở mỗi vị trí, dựa trên tầm quan trọng tương đối với các vị trí khác.

Với mỗi lớp $l, l = 1..L$ trong khối Encoder, với LN là LayerNorm, ta lại có:

$$\mathbf{Z'_l} = sMHA(LN(\mathbf{Z_{l\text{-}1}})) + \mathbf{Z_{l\text{-}1}} \in \mathbb{R}^{t \times d}$$

$$\mathbf{Z_l} = MLP(LN(\mathbf{Z'})) + \mathbf{Z'} \in \mathbb{R}^{t \times d}$$

Đầu ra của khối Encoder tại vị trí của class token thể hiện tổng hợp thông tin chuỗi đầu vào:

$$\mathbf{Y_C} = \mathbf{Z_L[0]} \in \mathbb{R}^d$$

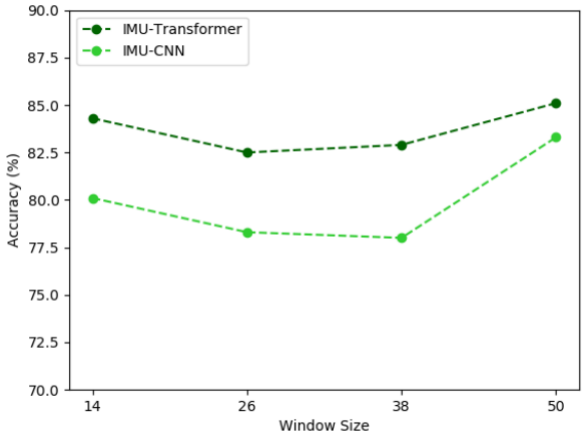
$\mathbf{Y_C}$ là đầu vào của một đầu phân loại, bao gồm các lớp LN và FC cùng hàm phi tuyến GELU và Dropout làm giảm số chiều xuống $\frac{d}{4}$. Một lớp FC thứ hai ánh xạ $\frac{d}{4}$ thành số lượng lớp. Log SoftMax được áp dụng lên vector đầu ra cùng với Negative Log Likelihood (NLL) loss để học một tác vụ phân loại nhiều nhãn.

5. **Kết quả:**

- **Thiết lập thử nghiệm:**

  Bài báo so sánh phương pháp được đề xuất với một mô hình CNN đã được chứng minh có hiệu suất tốt trên các tác vụ SLR khác nhau. Mỗi tập dữ liệu được chia thành tập huấn luyện và tập kiểm tra, với 85% các mẫu được chọn cho tập huấn luyện. Cả hai mô hình (mô hình CNN và mô hình được đề xuất) sử dụng thuật toán tối ưu Adam, với $\beta_1 = 0.9$, $\beta_2 = 0.999$ và $\epsilon = 10^{-10}$, batch size là 128 và weight decay là 10^{-4}, learning rate ban đầu là $\lambda = 10^{-4}$ và sẽ giảm đi một nửa sau mỗi m epoch tùy thuộc vào thử nghiệm. Mỗi mô hình được huấn luyện 30 epoch cho các tập dữ liệu nhỏ và 80 epoch cho các tập dữ liệu lớn hơn. Để thuận tiện, từ đây mô hình CNN và phương pháp được đề xuất được gọi là IMU-CNN và IMU-Transformer.

- **Ảnh hưởng của window size:**



Hình 2: Độ chính xác trên tập SHAR với các window size khác nhau

- **Độ chính xác trên các tập dữ liệu khác nhau:**

Experiment	Window Size	IMU-CNN Accuracy	IMU-Transformer Accuracy
SLR	50	96.5%	97.4%
HAR	50	86.2%	89.6%
SHAR	50	83.3%	85.1%
Overall	50	88.7%	90.7%

Hình 3: So sánh độ chính xác trên các tập dữ liệu

- **Tổng kết kết quả:**

  Kết quả thử nghiệm cho thấy mô hình đề xuất đạt được độ chính xác cao hơn và tổng quát hóa tốt hơn trên tất cả các tập dữ liệu và kịch bản đề ra. Cụ thể, mô hình IMU-Transformer đã cải thiện trung bình 2% về độ chính xác so với mô hình IMU-CNN trên các tập dữ liệu SLR, HAR và SHAR.

  Trong một thử nghiệm khác với các hoạt động liên quan đến cầu thang, mô hình IMU-Transformer đã cải thiện độ chính xác từ 86.6% (với mô hình IMU-CNN) lên 92.3%, giảm tỷ lệ phân loại sai của nhãn Upstairs Uparm từ 28% xuống còn 7.5%.

6. **Hạn chế:**

   Bài báo không đề cập cụ thể đến các hạn chế nhưng có thể có một số vấn đề cần làm rõ thêm như mô hình có hoạt động tốt với dữ liệu cảm biến quán tính nhiễu hoặc trong các điều kiện không phù hợp hay không.

7. **Các nghiên cứu trong tương lai:**

   Các hướng nghiên cứu trong tương lai có thể cải tiến mô hình Transformer để tăng độ chính xác trong việc nhận dạng hoạt động con người từ dữ liệu cảm biến quán tính hoặc có thể theo hướng áp dụng mô hình này vào các loại dữ liệu và ngữ cảnh khác nhau.