Multimodal AI for Real-World Signals and the Role of Language

Dimitris Spathis
Sr Researcher
Nokia Bell Labs
Cambridge, UK

dispathis.com

AAAI 2024 Health Intelligence workshop • February 2024
Why don’t models learn like humans or animals?

How do babies learn to interact with the world in a few months?

How do teenagers learn to drive with only a few hours of training?
Multimodal data → structurally different signals

- Face masks (breathing pattern, airborne pathogens, inflammation markers)
- Smart textiles (skin temperature, metabolites)
- Electronic epidermal tattoos (stress biomarkers, for example cortisol)
- Smartwatches (activity, sleep, resting heart rate, blood O₂ levels)
- Smart lenses (intraocular pressure)
- On-teeth sensors (drugs, for example antivirals and antibiotics)
- Smart patches (electrocardiogram)
- Microneedle patches (metabolites for example glucose, lactate, inflammation markers for example C-reactive protein), drugs
- Wristbands (electrolytes, metabolites, skin temperature)
- Smart rings (blood pressure)
The future of multimodal capabilities: on device AI + sensing

All-purpose devices with reliance on cloud services
Smart cloud, “dumb” device

Enhanced sensing and on-device capabilities
Smart device and cloud
$0.05/label

easy

DOG
0.05/label
easy
DOG
hard
?
$0.05/\text{label} \quad \rightarrow \quad >80x \quad \rightarrow \quad $4.00+/\text{label}

*assuming a sleep technician charging $50/h and 90-120 sleep stage transitions per 8 hours of sleep

**easy**

**IMAGE RECOGNITION**

**hard**

**SENSOR-BASED SLEEP TRACKING**

**DOG**

?
Fitness is measured with expensive & high-risk treadmill tests
Low fitness: strong predictor of all-cause mortality

<table>
<thead>
<tr>
<th>Variable</th>
<th>HR (95% CI)</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Comorbidity</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smoking</td>
<td>1.41 (1.36-1.46)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>CAD</td>
<td>1.29 (1.24-1.35)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Diabetes</td>
<td>1.40 (1.34-1.46)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Hypertension</td>
<td>1.21 (1.16-1.25)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>ESRD</td>
<td>2.78 (2.53-3.05)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td><strong>Group comparison</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low vs Elite</td>
<td>5.04 (4.10-6.20)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Low vs High</td>
<td>3.90 (3.67-4.14)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Low vs Above Average</td>
<td>2.75 (2.61-2.89)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Low vs Below Average</td>
<td>1.95 (1.86-2.04)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Below Average vs Elite</td>
<td>2.59 (2.10-3.19)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Below Average vs High</td>
<td>2.00 (1.88-2.14)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Below Average vs Above Average</td>
<td>1.41 (1.34-1.49)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Above Average vs Elite</td>
<td>1.84 (1.49-2.26)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Above Average vs High</td>
<td>1.42 (1.33-1.52)</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>High vs Elite</td>
<td>1.29 (1.05-1.60)</td>
<td>.02</td>
</tr>
</tbody>
</table>

Mandsager et al, 2018, JAMA Network Open
Converting raw wearable sensor data into cardio-fitness estimates

Spathis et al, 2022 (Nature Digital Medicine)
Free-living wearable data + anthropometrics = best result

**Anthropometrics**
- Age/Sex/Weight/BMI/Height

**Resting Heart Rate**
- RHR (Sensor-derived)

**Anthropometrics + RHR**
- Age/Sex/Weight/BMI/Height/RHR

**Sensors (68 var.) + RHR + Anthro**
- Acceleration/HR/HRV/MVPA
- Age/Sex/Weight/BMI/Height/RHR

Spathis et al, 2022 (Nature Digital Medicine)
Predicting substantial fitness change almost a decade later (AUC=0.74)

Spathis et al, 2022 (Nature Digital Medicine)
Signal annotation is not straightforward and sometimes infeasible

Annotation is supposed to be the golden standard in collecting ground-truth but rater (dis)agreement introduces further confusion to the models.

For some tasks such as sensor-based activity recognition, an additional video recording is required for annotation, which cannot scale to real-world settings.
Self-supervised learning uses **existing data** as prediction targets.

- Predict any part of the input from any other part.
- Predict the **future** from the **past**.
- Predict the **future** from the **recent past**.
- Predict the **past** from the **present**.
- Predict the **top** from the **bottom**.
- Predict the occluded from the visible.
- Pretend there is a part of the input you don’t know and predict that.
The first models were based on heuristic (pretext) tasks.
We can even create pretext tasks **across different modalities**
Heart rate responses to activity are evident in wearable data
Self-supervised learning enables transferrable models
And it comes in different variants
Current approaches focus on pre-processing & unimodal data. We have to create both positive and negative pairs (not straightforward to pick). There are no considerations for learning both within and across different modalities.
Pre-processing not straightforward: which augmentation first?

(a) MotionSense: Linear Evaluation

<table>
<thead>
<tr>
<th>Transformation</th>
<th>noised</th>
<th>scaled</th>
<th>rotated</th>
<th>negated</th>
<th>time_flipped</th>
<th>permuted</th>
<th>channel_shuffled</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>noised</td>
<td>76.4</td>
<td>70.6</td>
<td>81.7</td>
<td>75.5</td>
<td>78.6</td>
<td>77.4</td>
<td>78.5</td>
<td>78.9</td>
</tr>
<tr>
<td>scaled</td>
<td>82.5</td>
<td>89.5</td>
<td>84.8</td>
<td>80.9</td>
<td>82.9</td>
<td>84.0</td>
<td>84.9</td>
<td>85.0</td>
</tr>
<tr>
<td>rotated</td>
<td>81.0</td>
<td>83.4</td>
<td>81.3</td>
<td>69.6</td>
<td>81.6</td>
<td>82.5</td>
<td>83.2</td>
<td>79.7</td>
</tr>
<tr>
<td>negated</td>
<td>76.8</td>
<td>82.4</td>
<td>82.9</td>
<td>78.5</td>
<td>75.5</td>
<td>74.8</td>
<td>84.5</td>
<td>39.4</td>
</tr>
<tr>
<td>time_flipped</td>
<td>77.1</td>
<td>82.6</td>
<td>84.5</td>
<td>75.8</td>
<td>77.3</td>
<td>77.4</td>
<td>81.6</td>
<td>71.8</td>
</tr>
<tr>
<td>permuted</td>
<td>80.7</td>
<td>84.9</td>
<td>82.2</td>
<td>78.9</td>
<td>80.6</td>
<td>77.7</td>
<td>84.8</td>
<td>84.6</td>
</tr>
<tr>
<td>time_warped</td>
<td>82.7</td>
<td>84.9</td>
<td>83.8</td>
<td>85.7</td>
<td>81.1</td>
<td>84.5</td>
<td>83.2</td>
<td>70.1</td>
</tr>
<tr>
<td>channel_shuffled</td>
<td>73.8</td>
<td>82.8</td>
<td>80.4</td>
<td>67.2</td>
<td>70.1</td>
<td>87.2</td>
<td>76.4</td>
<td>74.1</td>
</tr>
</tbody>
</table>

(b) MotionSense: Fine-tuned Evaluation

<table>
<thead>
<tr>
<th>Transformation</th>
<th>noised</th>
<th>scaled</th>
<th>rotated</th>
<th>negated</th>
<th>time_flipped</th>
<th>permuted</th>
<th>time_warped</th>
<th>channel_shuffled</th>
<th>average</th>
</tr>
</thead>
<tbody>
<tr>
<td>noised</td>
<td>83.1</td>
<td>85.9</td>
<td>91.0</td>
<td>84.0</td>
<td>84.4</td>
<td>84.3</td>
<td>86.6</td>
<td>73.4</td>
<td>84.9</td>
</tr>
<tr>
<td>scaled</td>
<td>85.9</td>
<td>88.1</td>
<td>93.0</td>
<td>89.2</td>
<td>88.6</td>
<td>87.2</td>
<td>88.3</td>
<td>83.2</td>
<td>88.7</td>
</tr>
<tr>
<td>rotated</td>
<td>91.2</td>
<td>90.2</td>
<td>94.2</td>
<td>86.0</td>
<td>91.2</td>
<td>91.4</td>
<td>90.8</td>
<td>90.9</td>
<td>90.8</td>
</tr>
<tr>
<td>negated</td>
<td>84.8</td>
<td>90.7</td>
<td>92.0</td>
<td>84.0</td>
<td>85.0</td>
<td>86.0</td>
<td>90.2</td>
<td>77.6</td>
<td>86.3</td>
</tr>
<tr>
<td>time_flipped</td>
<td>82.5</td>
<td>87.5</td>
<td>92.2</td>
<td>84.9</td>
<td>86.4</td>
<td>84.4</td>
<td>88.0</td>
<td>85.4</td>
<td>86.4</td>
</tr>
<tr>
<td>permuted</td>
<td>85.4</td>
<td>89.3</td>
<td>92.2</td>
<td>87.4</td>
<td>86.4</td>
<td>83.5</td>
<td>88.9</td>
<td>82.4</td>
<td>86.7</td>
</tr>
<tr>
<td>time_warped</td>
<td>87.2</td>
<td>86.7</td>
<td>91.3</td>
<td>89.5</td>
<td>87.8</td>
<td>88.1</td>
<td>87.4</td>
<td>79.5</td>
<td>87.2</td>
</tr>
<tr>
<td>channel_shuffled</td>
<td>77.5</td>
<td>86.1</td>
<td>91.0</td>
<td>81.1</td>
<td>79.5</td>
<td>83.2</td>
<td>81.2</td>
<td>81.9</td>
<td>82.7</td>
</tr>
</tbody>
</table>

Tang, Pozuelo, Spathis et al, 2020, ML4MH NeurIPS
Masked Autoencoders offer a simpler architecture based on masking.

https://multimae.epfl.ch/
Masking has been wildly **successful** in training (Chat)GPT

Unsupervised Pre-training

Expensive training on massive datasets

Dataset: 300 billion tokens of text
Objective: Predict the next word

Example:

```
a robot must ?
```
Core idea: contrasting masked latent embeddings

Unlabelled Multimodal Sensor Data \[ E^{1..M} \] \[ \rightarrow \] Sensor-specific Encoders \[ Q \] \[ \rightarrow \] Intermediate Embeddings \[ Q' \] \[ \rightarrow \] Masked Embeddings Aggregator \[ A \] \[ \rightarrow \] Global embeddings \[ Z \]

Embedding patches \[ \rightarrow \] Masked embedding \[ \rightarrow \] Global embedding

Deldari et al., WSDM 2024 & MLMHD @ ICML 2023
1. Use a separate encoder for each modality/sensor
2. Merge all embeddings to a joint representation
3. Mask the representations in the latent space
4. Train the network to contrast the two views
5. After pre-training is done, we fine-tune the model with labels.
Results
Spatial masking + fine-tuning outperforms other methods

CroSSL: we test our method in two modes

<table>
<thead>
<tr>
<th>Training</th>
<th>Technique</th>
<th>Method</th>
<th>SleepEDF</th>
<th>Dataset PAMAP2</th>
<th>Dataset WESAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>End-2-End</td>
<td>Supervised</td>
<td>DeepConvLSTM</td>
<td>0.717 (.03)</td>
<td>0.879 (.12)</td>
<td>0.884 (.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0.601 (.02)</td>
<td>0.718 (.18)</td>
<td>0.791 (.04)</td>
</tr>
<tr>
<td>SSL (Fixed enc.)</td>
<td>COCOA</td>
<td></td>
<td>0.628 (.02)</td>
<td>0.839 (.11)</td>
<td>0.669 (.01)</td>
</tr>
<tr>
<td></td>
<td>CroSSL (random)</td>
<td></td>
<td>0.628 (.00)</td>
<td>0.802 (.15)</td>
<td>0.642 (.02)</td>
</tr>
<tr>
<td></td>
<td>CroSSL (spatial)</td>
<td></td>
<td>0.722 (.02)</td>
<td>0.822 (.13)</td>
<td>0.667 (.02)</td>
</tr>
<tr>
<td>SSL (Fine-tuned enc.)</td>
<td>COCOA</td>
<td></td>
<td>0.678 (.01)</td>
<td>0.882 (.11)</td>
<td>0.913 (.03)</td>
</tr>
<tr>
<td></td>
<td>CroSSL (random)</td>
<td></td>
<td>0.726 (.00)</td>
<td>0.871 (.11)</td>
<td>0.894 (.02)</td>
</tr>
<tr>
<td></td>
<td>CroSSL (spatial)</td>
<td></td>
<td>0.741 (.00)</td>
<td>0.892 (.10)</td>
<td>0.939 (.03)</td>
</tr>
</tbody>
</table>

Masking
- Random
- Spatial

Transfer learning
- Fixed (frozen)
- Fine-tuned (re-training)
CroSSL is robust to missing modalities in prediction time

<table>
<thead>
<tr>
<th>Missing data at:</th>
<th>Technique</th>
<th>Dataset</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fine-tuning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inference</td>
<td>Masking</td>
<td>Method</td>
<td>SleepEDF</td>
<td>PAMAP2</td>
<td>WESAD</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------</td>
<td>------------------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Supervised</td>
<td>0.717 (.03)</td>
<td>0.879 (.12)</td>
<td>0.884 (.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fixed SSL</td>
<td>0.628 (.00)</td>
<td>0.709 (.18)</td>
<td>0.629 (.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fine-tuned SSL</td>
<td>0.776 (.00)</td>
<td>0.825 (.13)</td>
<td>0.890 (.01)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fixed SSL</td>
<td>0.722 (.02)</td>
<td>0.822 (.14)</td>
<td>0.715 (.06)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fine-tuned SSL</td>
<td>0.741 (.00)</td>
<td>0.892 (.11)</td>
<td>0.925 (.03)</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Supervised</td>
<td>0.703 (.03)</td>
<td>0.897 (.11)</td>
<td>0.894 (.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fixed SSL</td>
<td>0.602 (.03)</td>
<td>0.742 (.18)</td>
<td>0.622 (.03)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fine-tuned SSL</td>
<td>0.738 (.03)</td>
<td>0.859 (.13)</td>
<td>0.899 (.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fixed SSL</td>
<td>0.694 (.01)</td>
<td>0.805 (.16)</td>
<td>0.655 (.02)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fine-tuned SSL</td>
<td>0.739 (.02)</td>
<td>0.899 (.09)</td>
<td>0.923 (.03)</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Supervised</td>
<td>0.202 (.17)</td>
<td>0.469 (.36)</td>
<td>0.304 (.37)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fixed SSL</td>
<td>0.206 (.35)</td>
<td>0.331 (.19)</td>
<td>0.186 (.16)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fine-tuned SSL</td>
<td>0.200 (0)</td>
<td>0.440 (.28)</td>
<td>0.139 (18)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fixed SSL</td>
<td>0.667 (.13)</td>
<td>0.646 (.21)</td>
<td>0.278 (.14)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fine-tuned SSL</td>
<td>0.581 (.24)</td>
<td>0.495 (.35)</td>
<td>0.234 (.17)</td>
</tr>
</tbody>
</table>

Spatial masking is more robust in missing modalities on inference time

Fixed/base models outperform in data-scarce fine-tuning

while supervised models are heavily impacted by missing data

Deldari et al., WSDM 2024 & MLMHD @ ICML 2023
Spatial > Random, masking more effective in larger datasets

High latent masking ratios do not result in high performance, unlike in vision/MAE papers. Performance drop is more visible in random masking.

Fine-tuned CroSSL outperforms the fixed variant in most cases.

Deldari et al., WSDM 2024 & MLMHD @ ICML 2023
Fine-tuned models are label-efficient, fixed ones need warmup

Fine-tuning is as good as supervised models that have access to labeled data, but it is particularly effective in the low-data regime (1-10% of labels)

Deldari et al., WSDM 2024 & MLMHD @ ICML 2023
Takeaways

1. Achieves state of art performance in multimodal signal ML tasks
2. Handles missing data/sensors in an elegant manner
3. Is data & label-efficient with performance on par or better to supervised models
4. Requires no data pre-processing such as negative pair mining or hiding inputs

Deldari et al., WSDM 2024 & MLMHD @ ICML 2023
Problem solved?
Self-supervision needs large **unlabeled** data: where to find them?

<table>
<thead>
<tr>
<th></th>
<th>PPG</th>
<th>ECG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of participants</td>
<td>141,207</td>
<td>106,643</td>
</tr>
<tr>
<td>Number of segments</td>
<td>19,854,101</td>
<td>3,743,679</td>
</tr>
<tr>
<td>Average number of calendar days per participant</td>
<td>92.54</td>
<td>23.27</td>
</tr>
<tr>
<td>Total dataset time span (days)</td>
<td>890</td>
<td>1,240</td>
</tr>
</tbody>
</table>

Apple Heart and Movement Study

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Subjects</th>
<th>#Samples</th>
<th>#Classes</th>
<th>Environment</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>UK-Biobank</td>
<td>~100K</td>
<td>6 B</td>
<td>Unlabelled</td>
<td>Free-living</td>
<td>Doherty et al. (2017)</td>
</tr>
<tr>
<td>Rowlands</td>
<td>55</td>
<td>36K</td>
<td>13</td>
<td>Lab</td>
<td>Eshiger et al. (2011)</td>
</tr>
<tr>
<td>WISDM</td>
<td>46</td>
<td>28K</td>
<td>18</td>
<td>Semi free-living</td>
<td>Weiss et al. (2019)</td>
</tr>
<tr>
<td>REALWORLD</td>
<td>14</td>
<td>12K</td>
<td>8</td>
<td>Lab</td>
<td>Sztyler and Stuckenschmidt (2016)</td>
</tr>
<tr>
<td>Opportunity</td>
<td>4</td>
<td>3.9K</td>
<td>4</td>
<td>Semi free-living</td>
<td>Roggen et al. (2010)</td>
</tr>
<tr>
<td>PAMAP2</td>
<td>8</td>
<td>2.9K</td>
<td>8</td>
<td>Lab</td>
<td>Reiss and Stricker (2012)</td>
</tr>
<tr>
<td>ADL</td>
<td>7</td>
<td>0.6K</td>
<td>5</td>
<td>Lab</td>
<td>Bruno et al. (2013)</td>
</tr>
</tbody>
</table>

Large unlabeled data of that kind is hard to collect

Not publicly available on the web, unlike images or text

Number of potential modalities hampers progress because it requires aligned/paired data

Available pre-trained models are limited in size and generalization capabilities

UK Biobank (wristband) compared to benchmark HAR data

https://openreview.net/forum?id=pC3WJHf51j & Yuan et al., ArXiv 2023
But do Large Language Models understand numbers?
LLM tokenizers are not designed for numbers

Consecutive digit chunking

<table>
<thead>
<tr>
<th>Input</th>
<th>→ Token IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>480, 481, 482</td>
<td>22148, 11, 4764, 16, 11, 4764, 17</td>
</tr>
</tbody>
</table>

Floats

<table>
<thead>
<tr>
<th>Input</th>
<th>→ Token IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.14159</td>
<td>18, 13, 1415, 19707</td>
</tr>
</tbody>
</table>

Case sensitive, trailing whitespaces, arbitrary integer grouping, inconsistent long integer chunking, model-specific behaviours, ...

Spathis & Kawsar, GenAI4PC @ Ubicomp 2023
A case study with activity timeseries data and the GPT tokenizer

Spathis & Kawsar, GenAI4PC @ Ubicomp 2023
Bridging the modality gap with adapters & prompt-tuning

https://blog.research.google/2023/08/multimodal-medical-ai.html
Prompting with numbers in addition to text

**Activity Recognition**

**Prompt:**
Classify the accelerometer data in meters per second squared as either walking or running.

**Response:**
Walking.

**Atrial Fibrillation Classification**

**Prompt:**
Classify the given Interbeat Interval sequence in ms as either Atrial Fibrillation or Normal Sinus.

**Response:**
Atrial Fibrillation.

**Normal Sinus.**

**Stress**

**Prompt:**
Steps [Steps], resting heart rate: [RHR] beats/min, sleep duration: [SleepMinutes] mins, non-rem heart rate: [NREMHR] beats/min, mood last day [Mood] out of 5. What will my stress level be?

**Response:**
Stress: 5 out of 5.

Stress: 1 out of 5.

Liu et al., ArXiv 2023
From prompt engineering to few-shot learning to prompt-tuning.
Instead of prompt-tuning, first (auto)encode the numerical data.
Everything-to-everything **multimodal models**

Interleaved Modalities
(Image + Motion Sensor)

(Given the motion signals)
Write a social media caption for this view.

Pedaling along the San Francisco Bay, taking in breathtaking views of the Bay Bridge! The salty sea breeze invigorates me as I make my way to my next destination.
Pros

- Computationally efficient
- LLM is fixed/frozen
- Allows connecting to other high-performing models (e.g. a sota ECG encoder)
- Breaking down the system to encoder + adapter + LLM enables faster iteration and testing

Cons

- Modularization introduces complexity, gradients don’t propagate all the way
- Adapter <-> LLM communication is no longer interpretable (compared to natural language prompts)

Spathis & Kawsar, GenAI4PC @ Ubicomp 2023
Where are we now and what is missing?

Treating LLMs as generic pre-trained models seems to be working (!)

We still have to "ground" them through

- Verbose hand-engineered prompts
- Extensive aggregation/downsampling
- Careful dataset curation

- Adapters don't need elaborate textual prompts
- Multimodal integration through joint embedding spaces
- Improved digit-level tokenizers
- Longer context windows that fit high-dimensional data
**Latent Masking for Multimodal Self-supervised Learning in Health Timeseries**

Shohreh Deldari 1,2, Dimitris Spathis 3, Mohammad Malekzadeh 3, Fahim Kawsar 3, Flora Salim 3, Akhil Mathur 3

**Abstract**

Limited availability of labeled data for machine learning on biomedical time-series hampers progress in the field. Self-supervised learning (SSL) is a promising approach to learning data representations without labels. However, current SSL methods require expensive computations for negative pairs and are designed for single modalities, limiting their versatility. To overcome these limitations, we introduce CroSSL (Cross-modal SSL). CroSSL introduces two novel concepts: masking intermediate embeddings from modality-specific encoders and aggregating them into a global embedding using a cross-modal aggregator. This enables the handling of missing modalities applications in healthcare, including human activity recognition (HAR) and sleep tracking through brain activity monitoring (Kemp et al., 2000; Tang et al., 2021). However, the reliance on labeled data for training deep neural networks (DNNs) has hindered their scalability (Yuan et al., 2022). Collecting, annotating, and maintaining large labeled datasets can be expensive, time-consuming, and impractical, leading to a growing interest in self-supervised learning (SSL) that learns from unlabeled data (Sayed et al., 2019). SSL defines an artificial task, known as a pretext task, where the supervisory signal is automatically generated from unlabelled data, enabling the training of an encoder model to learn a latent representation of the input data (Yuan et al., 2022). SSL has shown promise in various applications, such as HAR (Tang et al., 2021), by leveraging large amounts of unlabeled data.

Deldari et al, WSDM’24 & ML4MHD @ ICML’23
arxiv.org/abs/2307.16847

**The first step is the hardest: Pitfalls of Representing and Tokenizing Temporal Data for Large Language Models**

Dimitris Spathis 3, Fahim Kawsar 3

**ABSTRACT**

Large Language Models (LLMs) have demonstrated remarkable generalization across diverse tasks, leading individuals to increasingly use them as personal assistants and universal computing engines. Nevertheless, a notable obstacle emerges when feeding numerical/temporal data into these models, such as data sourced from wearables or electronic health records. LLMs employ tokenizers in their input that break down text into smaller units. However, tokenizers are not designed to represent numerical values and might struggle to understand repetitive patterns and context, treating consecutive values as separate tokens and disregarding their temporal relationships. Here, we discuss recent works that employ LLMs for human-centric tasks such as mobile health sensing and present a case study showing that popular LLMs tokenize temporal data incorrectly. To address that, we highlight potential solutions such as prompt tuning with lightweight embedding layers as well as multimodal adapters, that can help bridge this ‘modality gap’. While the capability of language models to generalize to other modalities with minimal or no fine-tuning is exciting, this paper underscores the unintentional fragmentation of continuous sequences into disjuncted tokens. Consequently, the temporal relationships that underpin such data may be lost in translation, potentially undermining the very essence of the information being processed.

In this context, this paper advances into the nuances and obstacles that emerge when LLMs are confronted with the task of representing and tokenizing temporal data. We focus on the interplay between numerical and textual information, uncovering the potential pitfalls that can hamper the effective utilization of LLMs in scenarios where temporal context is important. Last, we discuss potential solutions from the rapidly growing area of parameter-efficient transfer learning and multimodal adapters that could enable better integration of non-textual data into LLMs.

**2 TOKENIZATION IN LANGUAGE MODELS**

Tokenization is a fundamental process underpinning the operation of LLMs. It involves the division of input and output texts into smaller, manageable units known as tokens. These tokens serve

Spathis & Kawsar, GenAI UbiComp’23
arxiv.org/abs/2309.06236
Dimitris Spathis
Sr Researcher
Nokia Bell Labs
Cambridge, UK

dispathis.com

Q&A

AAAI 2024 Health Intelligence workshop • February 2024