LEARNING FROM FREQUENCY DOMAIN DATA

Handling Spatial Temporal IoT Data

Reminders and Announcements

- Your 2-page project proposal is due today (by email to me, as a PDF attachment).
- HW3 will be out tonight on the usual homework class page. Submission will be on canvas.
- For your presentations, presentation guidance was posted on Piazza
  - Search for Piazza note @45
  - Note #1: Please do engage in discussing your paper selection with me.
  - Note #2: The final paper selection for your talk is due one week before the talk.
  - Note #3: I will give you example papers roughly 2 weeks before your talk, but if you choose your talk exclusively from my example papers, you are telling me you have not invested in looking for more literature on the topic.
- Project elevator talks are on March 12th.

Reminders and Announcements

Elevator Talks

- Plan for a 5-minute talk that answers the four key Heilmeier questions about your project (see https://en.wikipedia.org/wiki/George_H._Heilmeier), namely:
  - What are you trying to do? Articulate your objectives using no jargon.
  - How is it done today, and what are the limits of current practice?
  - What's new in your approach and why do you think it will be successful?
  - Who cares? If you're successful, what difference will it make?
- You are encouraged to use visuals (it’s a short talk) but plan them well. The visuals should not be pure “eye candy”. They should serve as a vehicle to convey information more efficiently.
- Add a slide on the current status and timeline.

Today's Topic: Spatial Considerations in Self-Supervised Learning from IoT Data

- Self supervised learning must account for content as well as positional structure
  - Example 1: How to account for positional structure in text?
  - Example 2: How to account for positional structure in images?
  - Example 3: How to account for positional structure in graphs?

Spatial Considerations in Self-Supervised Learning from IoT Data

- Self supervised learning must account for content as well as positional structure
  - Example 1: How to account for positional structure in text?
    - Positional embeddings of tokens (one-dimensional sequences)
  - Example 2: How to account for positional structure in images?
    - Positional embeddings of image patches
  - Example 3: How to account for positional structure in graphs?
    - Message passing schemes exchange information on neighboring node embedding, modifying each node's embedding based on that of its neighbors

Spatial Considerations in Self-Supervised Learning from IoT Data

- Self supervised learning must account for content as well as positional structure
  - Example 1: How to account for positional structure in text?
    - Positional embeddings of tokens (one-dimensional sequences)
  - Example 2: How to account for positional structure in images?
    - Positional embeddings of image patches
  - Example 3: How to account for positional structure in graphs?
    - Message passing schemes exchange information on neighboring node embedding, modifying each node's embedding based on that of its neighbors
  - How to account for positional structure in IoT sensor data?

Positional Structure in IoT Data

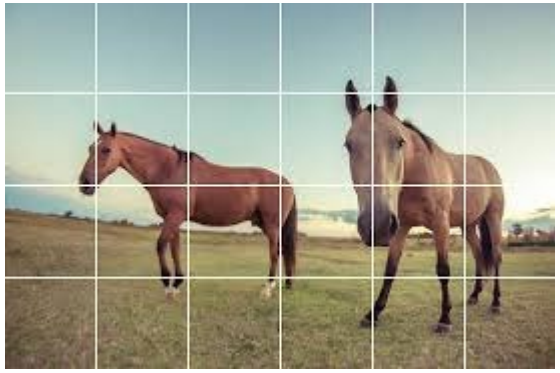
- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***

Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***

Intra-sample Structure: An Image Example

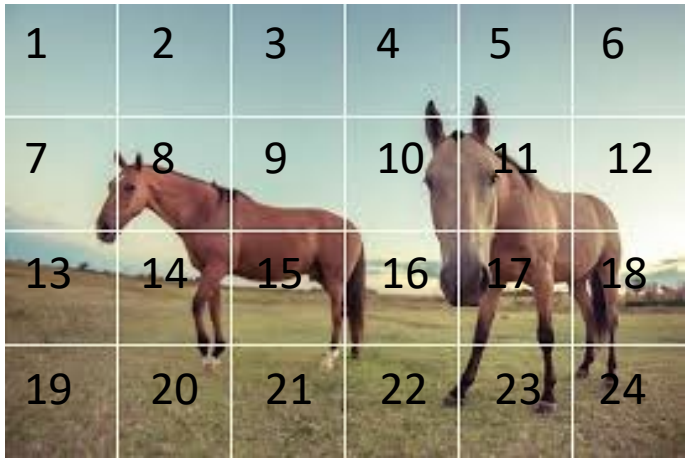
Break image into patches.



Example: ViT
https://github.com/lucidrains/vit-pytorch

Intra-sample Structure: An Image Example

Break image into patches.

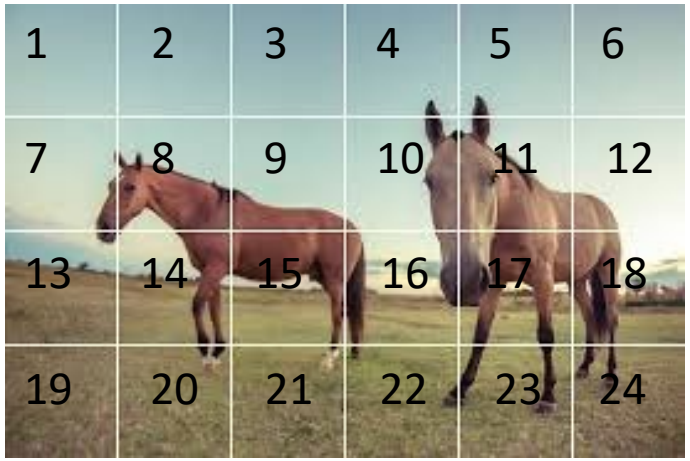


Patches have a positional embedding that encodes their position inside the picture

Example: ViT
https://github.com/lucidrains/vit-pytorch

Intra-sample Structure: An Image Example

Break image into patches.



Example: ViT
https://github.com/lucidrains/vit-pytorch

Patches have a positional embedding that encodes their position inside the picture

Note: Patch ID uniquely determines the position of each patch relative to others.

Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***

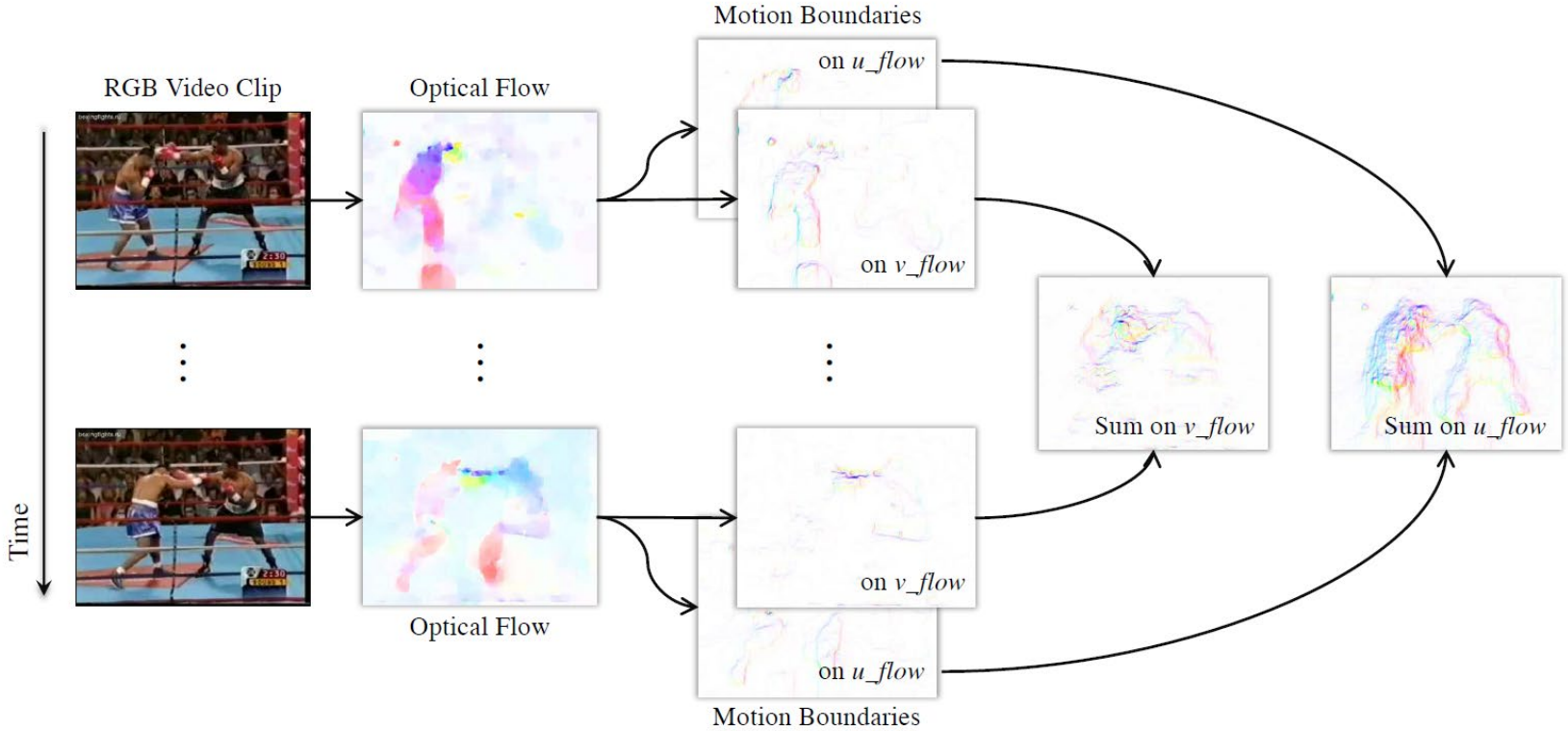
Inter-sample Spatial-Temporal Relations: A Video Example

How to extract spatial and temporal patterns in video content?

Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. "Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4006-4015. 2019.

Inter-sample Spatial-Temporal Relations: A Video Example

Example 1: A handcrafted approach – extract a number of engineered features from frames to best represent motion patterns



Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. "Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics." In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4006-4015. 2019.

Inter-sample Spatial-Temporal Relations: A Video Example

Example 1: A handcrafted approach – extract a number of engineered features from frames to best represent motion patterns

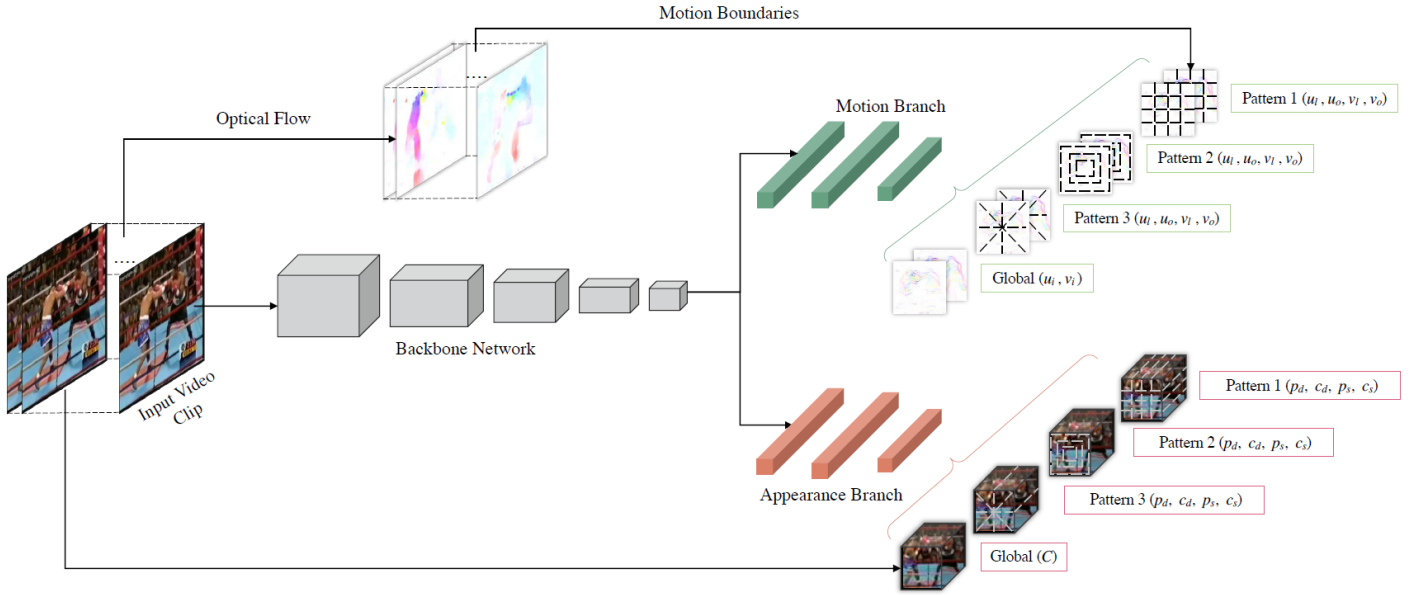


Figure 5. The network architecture of the proposed method. Given a 16-frame video, we regress 14 outputs for the motion branch and 13 outputs for the appearance branch. For each motion pattern, 4 labels are generated by aggregating motion boundaries M_u and M_v: (1) u_l – the largest magnitude location of M_u. (2) u_o – the corresponding orientation of u_l. (3) v_l – the largest magnitude location of M_v. (4) v_o – the corresponding orientation of v_l. For each appearance pattern, 4 labels are predicted: (1) p_d – the position of largest color diversity. (2) c_d – the corresponding dominant color. (3) p_s – the position of smallest color diversity. (4) c_s – the corresponding dominant color.

Handcrafted features are extracted from video and used as “pseudo-labels” that a neural network is trained to predict in scenes.

Inter-sample Spatial-Temporal Relations: A Video Example

Is there a better way? Can we use what we learned in self-supervised learning in this class to avoid the excessive handcrafting?

Inter-sample Spatial-Temporal Relations: A Video Example

Is there a better way? Can we use what we learned in self-supervised learning in this class to avoid the excessive handcrafting?

Example 2: Use contrastive learning

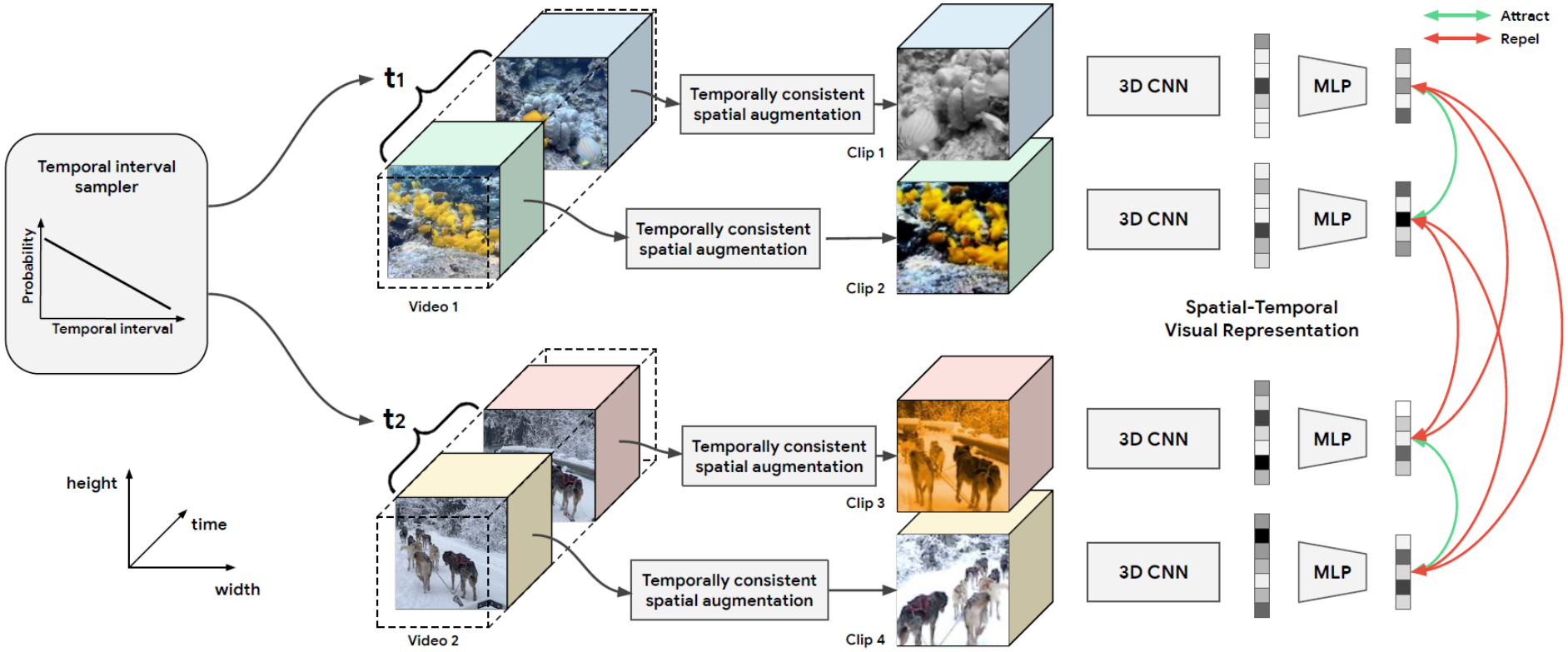
- How would one create a video representation that preserves notions of spatial and temporal similarity, such that temporally-similar and spatially-similar items are closer in the latent space?

*Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. "Spatiotemporal contrastive video representation learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6964-6974. 2021.

Inter-sample Spatial-Temporal Relations: A Video Example*

Example 2: A contrastive learning approach that assumes the following notions of "similarity" (for proximity in the latent space):

- **Temporal similarity:** Nearby frames are similar.
- **Spatial similarity:** A set of content preserving augmentations (cropping, resizing, etc)



*Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. "Spatiotemporal contrastive video representation learning." In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6964-6974. 2021.

Inter-sample Spatial-Temporal Relations: A Video Example*

Temporal "augmentation":

- Pick two nearby frames

Spatial augmentation:

- A combination of image flips, resizing, color jitter, gray-scaling, and Gaussian blur (Algorithm 1)

Algorithm 1: Temporally consistent spatial augmentation

Input: Video clip $V = \{f_1, f_2, \cdots, f_M\}$ with M frames
Crop: Randomly crop a spatial region with size ratio **S** in range of $[0.3, 1]$ and aspect ratio **A** in $[0.5, 2]$
Resize: Resize the cropped region to size of 224×224
Flip: Draw a flag $\mathbf{F}_f$ from $\{0, 1\}$ with 50% on 1
Jitter: Draw a flag $\mathbf{F}_j$ from $\{0, 1\}$ with 80% on 1
Grey: Draw a flag $\mathbf{F}_g$ from $\{0, 1\}$ with 20% on 1
for $k \in \{1, \ldots, M\}$ **do**
$f'_k = \text{Resize}\big(\text{Crop}(f_k, size = \mathbf{S}, aspect = \mathbf{A})\big)$
$f'_k = \text{Flip}(f'_k)$ if $\mathbf{F}_f = 1$
$f'_k = \text{Color\_jitter}(f'_k)$ if $\mathbf{F}_j = 1$
$f'_k = \text{Greyscale}(f'_k)$ if $\mathbf{F}_g = 1$
$f'_k = \text{Gaussian\_blur}(f'_k)$
end for
Output: Augmented video clip $V' = \{f'_1, f'_2, \cdots, f'_M\}$

Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***

Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***
  - Issue 1: Exploitation of different structural positions (or “vantage points”)
  - Issue 2: Exploitation of relations between different structural positions

Issue 1: Exploiting Positional Structure Across Sensor Streams

Motivating example: Human activity recognition from multiple wearables.

Physical activity: walking

Physical activity: walking

Physical activity: walking

Issue 1: Exploiting Positional Structure Across Sensor Streams

Motivating example: Human activity recognition from multiple wearables.

Observation: Not all wearables are equally well-positioned for activity recognition in different contexts

Physical activity: walking

Wrist immobilized by cart.
Use phone not wrist-watch

Physical activity: walking

Wrist is free.
Wrist-watch OK

Physical activity: walking

Issue 1: Exploiting Positional Structure Across Sensor Streams

Motivating example: Human activity recognition from multiple wearables.

Observation: Not all wearables are equally well-positioned for activity recognition in different contexts

- How to automatically choose the best-positioned sensors for the task?

Physical activity: walking

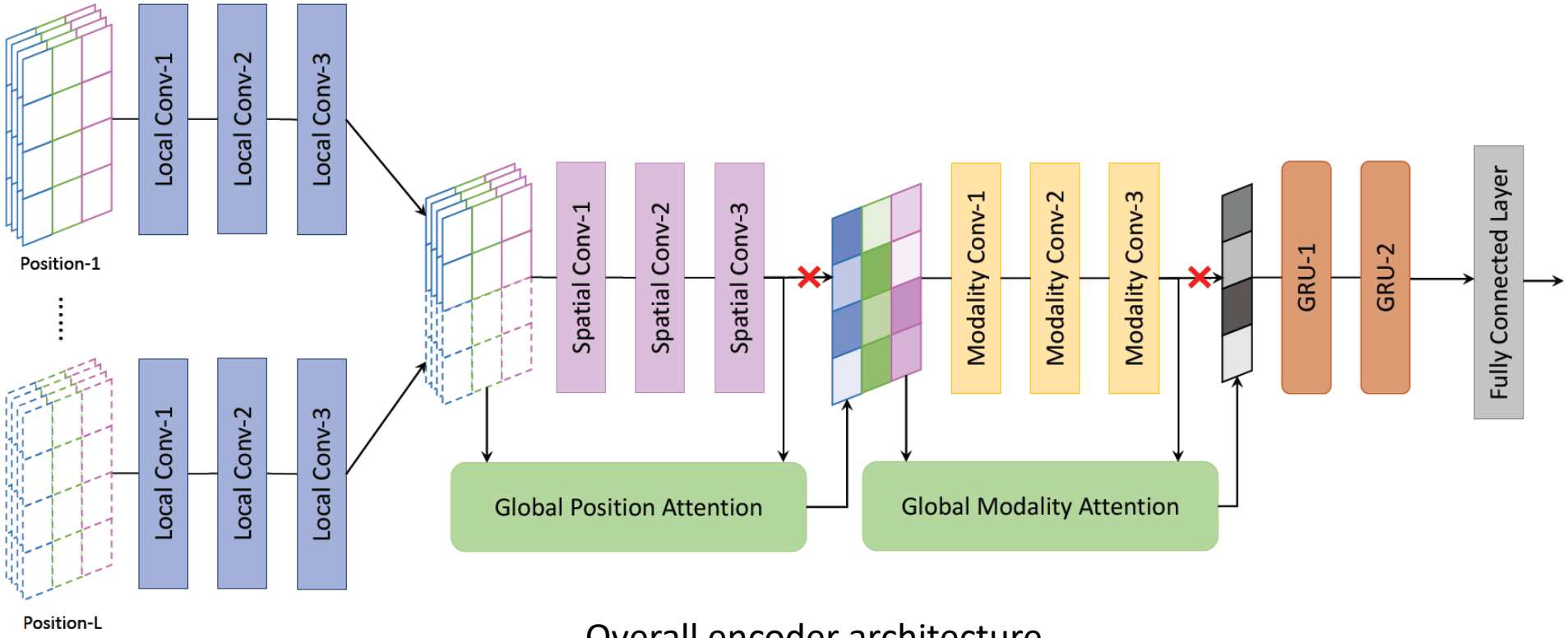
Physical activity: walking

Physical activity: walking

Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. "Giobalfusion: A global attentional deep learning framework for multisensor information fusion." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, no. 1 (2020): 1-27.

Issue 1: Exploiting Positional Structure Across Sensor Streams

Example: An attention layer to determine which sensors responsible for the most useful output features

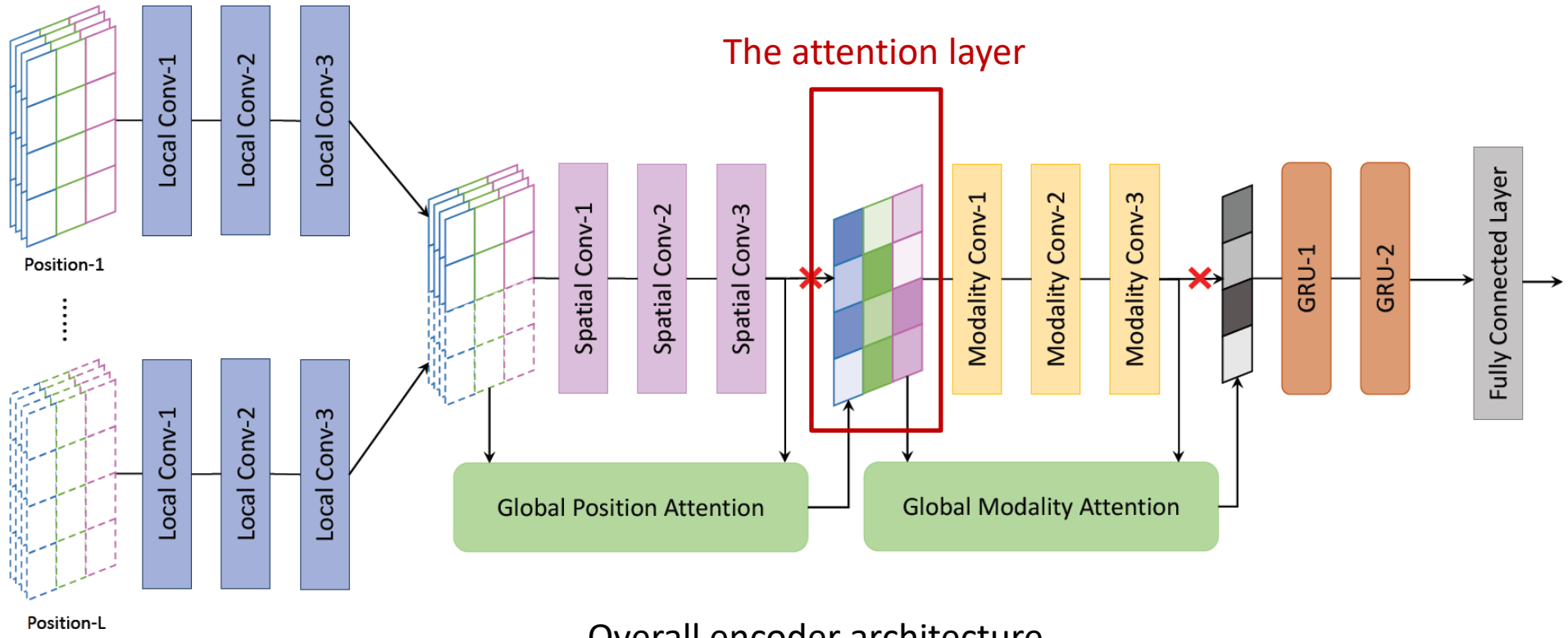


Overall encoder architecture

Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. "Giobalfusion: A global attentional deep learning framework for multisensor information fusion." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, no. 1 (2020): 1-27.

Issue 1: Exploiting Positional Structure Across Sensor Streams

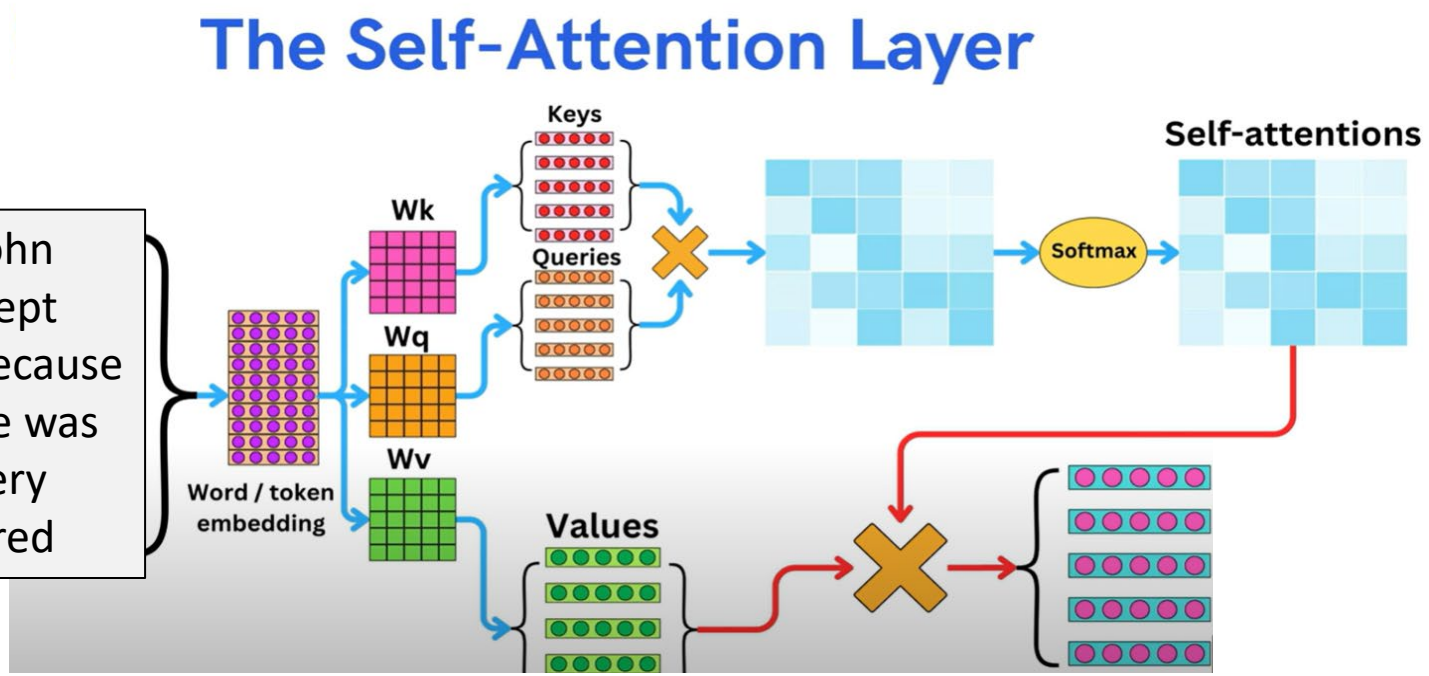
Example: An attention layer to determine which sensors responsible for the most useful output features



Overall encoder architecture

Reminder: The Notion of Attention

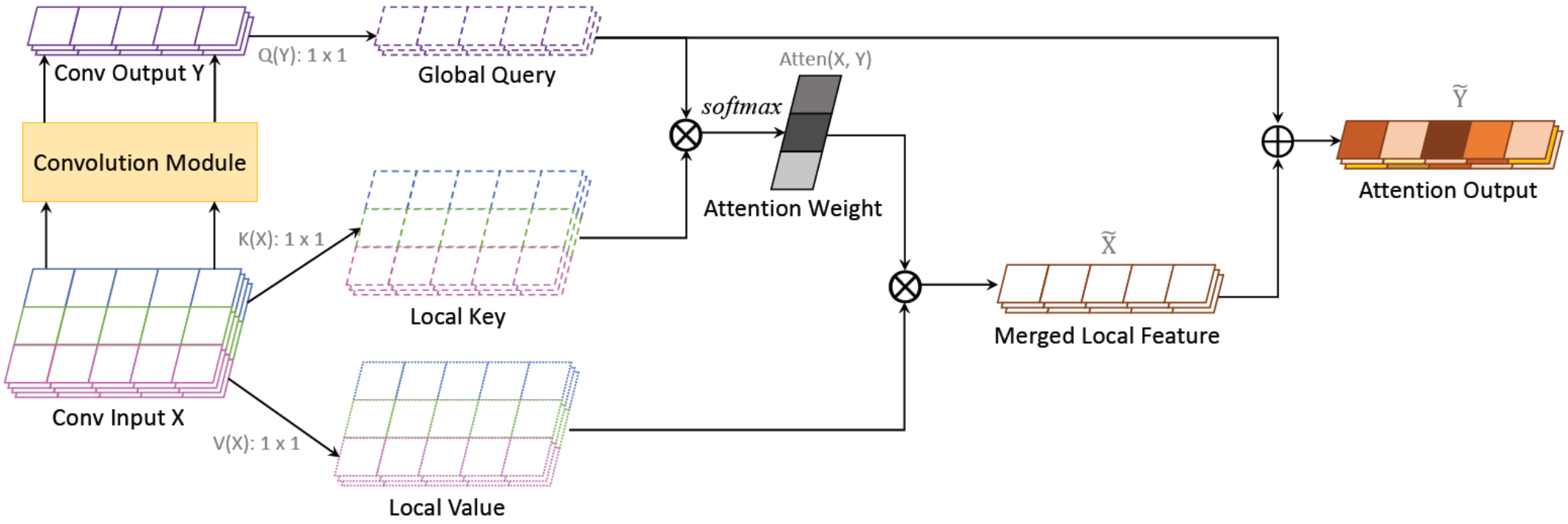
- **Example: Transformers** (Attention is all you need):
  - https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf



Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. "Giobalfusion: A global attentional deep learning framework for multisensor information fusion." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, no. 1 (2020): 1-27.

An Attention Layer to Exploit Positional Structure Across Sensor Streams

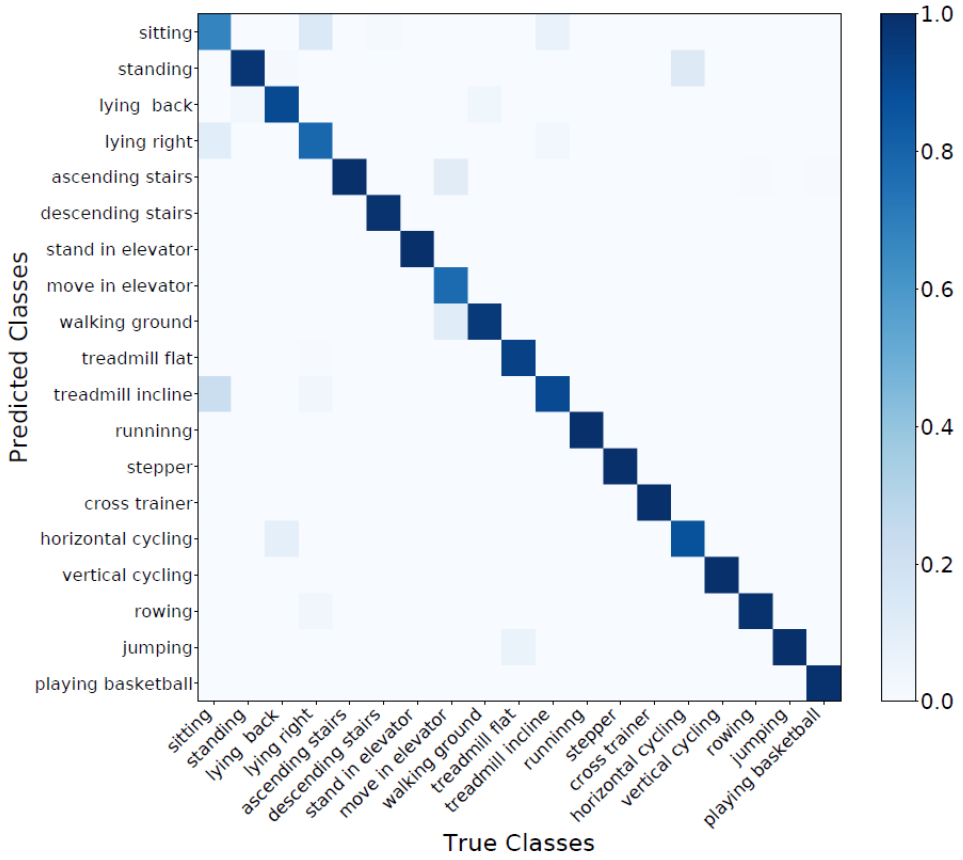
Example: An attention layer to determine which sensors responsible for the most useful output features



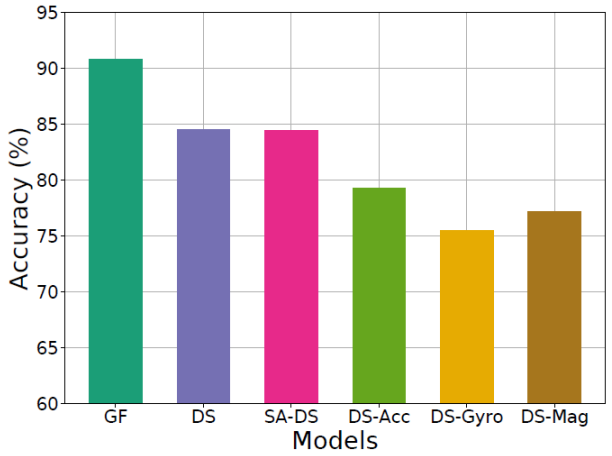
Design of the attention layer

Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. "Giobalfusion: A global attentional deep learning framework for multisensor information fusion." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, no. 1 (2020): 1-27.

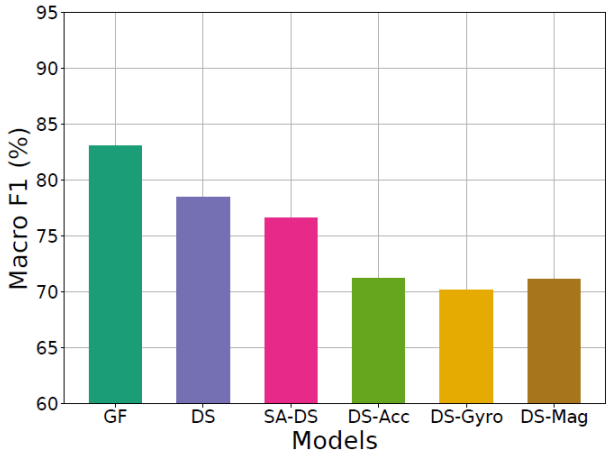
Example Results



Picking the right sensor positions and modalities for the job improves downstream classification accuracy and F1 score.



(a) Accuracy



(b) Macro F1 Score

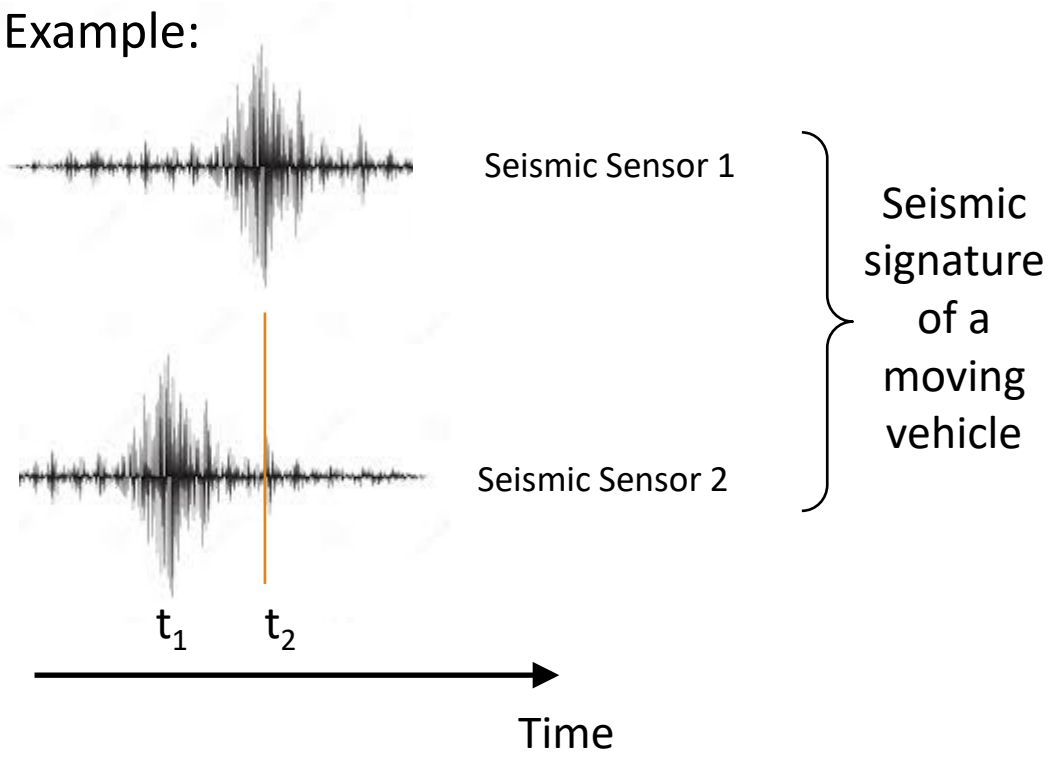
Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***
  - Issue 1: Exploitation of different structural positions (or “vantage points”)
  - Issue 2: Exploitation of relations between different structural positions

Issue 2: Exploiting Structural Relations

Understanding structural relations between sensors can help interpret their observations

Example:

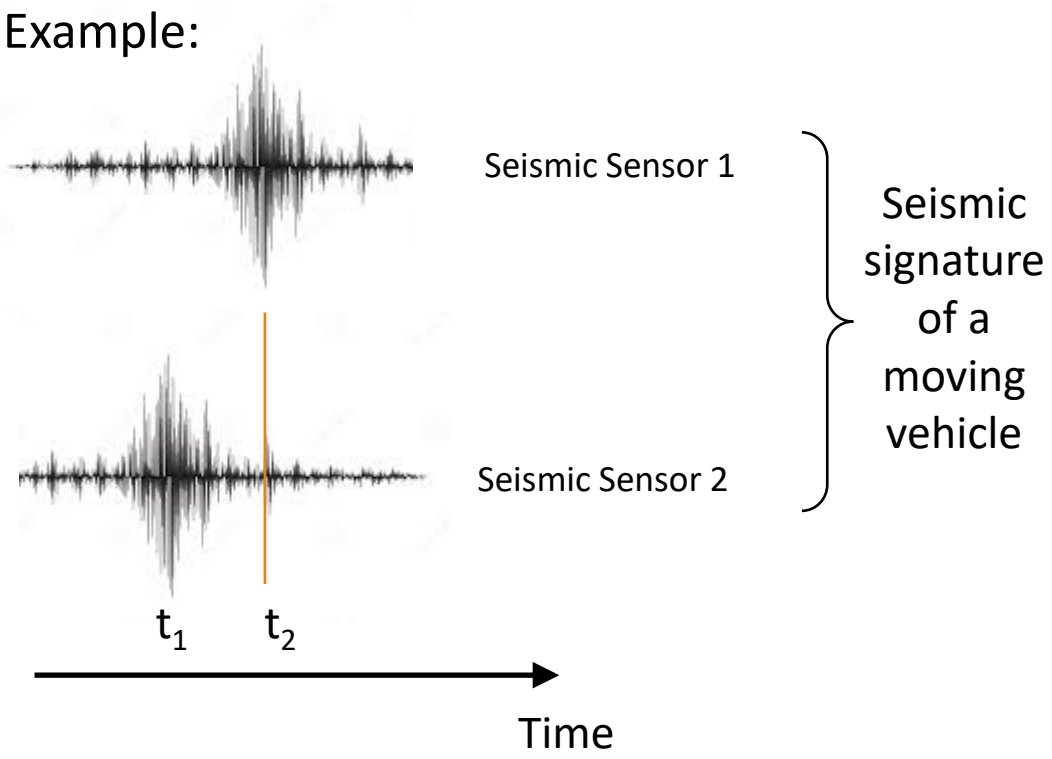


Is this car going north or south?

Issue 2: Exploiting Structural Relations

Understanding structural relations between sensors can help interpret their observations

Example:



Is this car going north or south?

Spatial reasoning requires understanding of relations between sensor positions and observed data.

Issue 2: Exploiting Structural Relations

A test of structural relation understanding: What simple task tests neural network understanding of relations between sensor positions (and sensor data)?

Issue 2: Exploiting Structural Relations

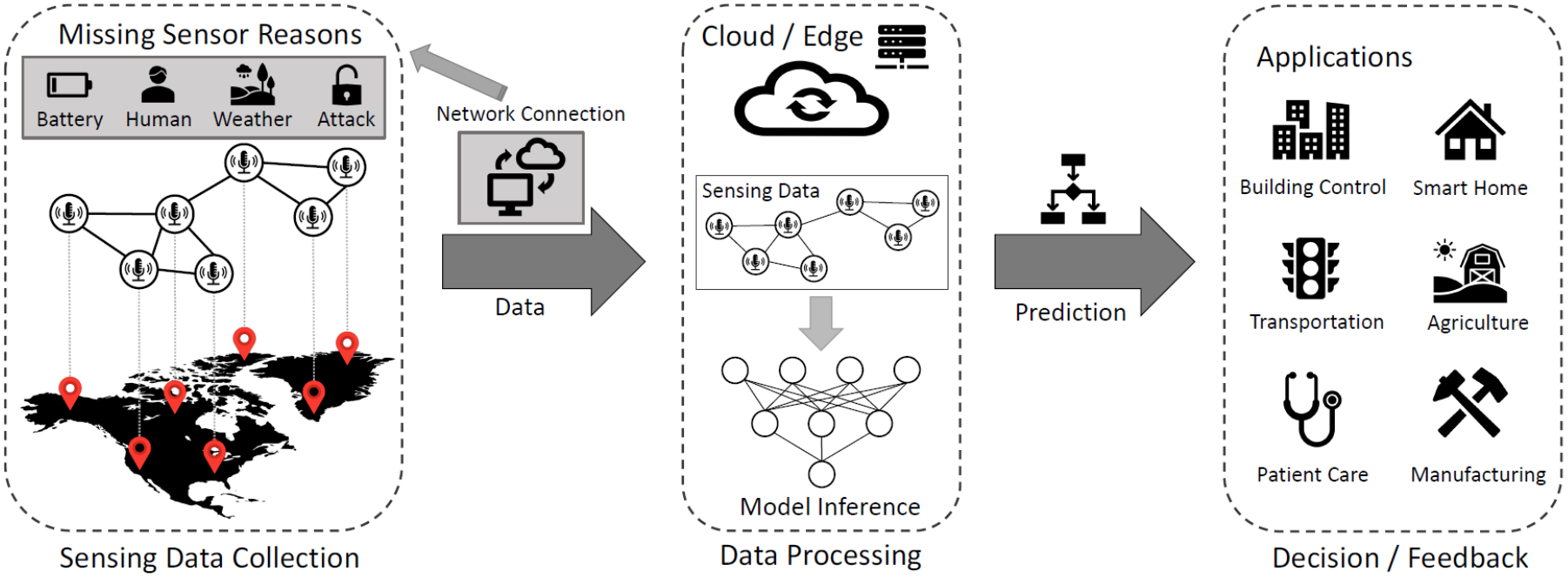
A test of structural relation understanding: What simple task tests neural network understanding of relations between sensor positions (and sensor data)?

- A simple task is to predict sensor values at one position from values at another.

Shengzhong Liu, Shuochao Yao, Yifei Huang, Dongxin Liu, Huajie Shao, Yiran Zhao, Jinyang Li et al. "Handling missing sensors in topology-aware iot applications with gated graph neural network." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, no. 3 (2020): 1-31.

Issue 2: Exploiting Structural Relations

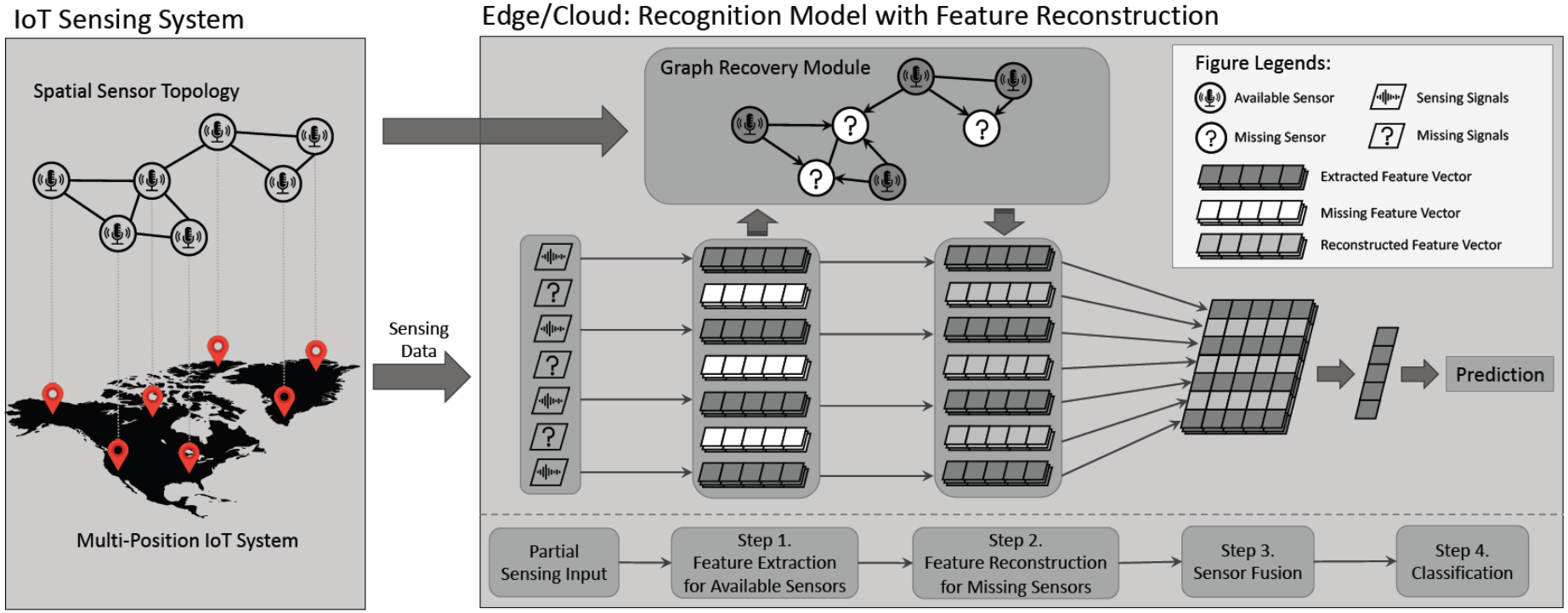
A test of structural relation understanding: Can I guess “missing” sensor waveforms from other sensor waveforms?



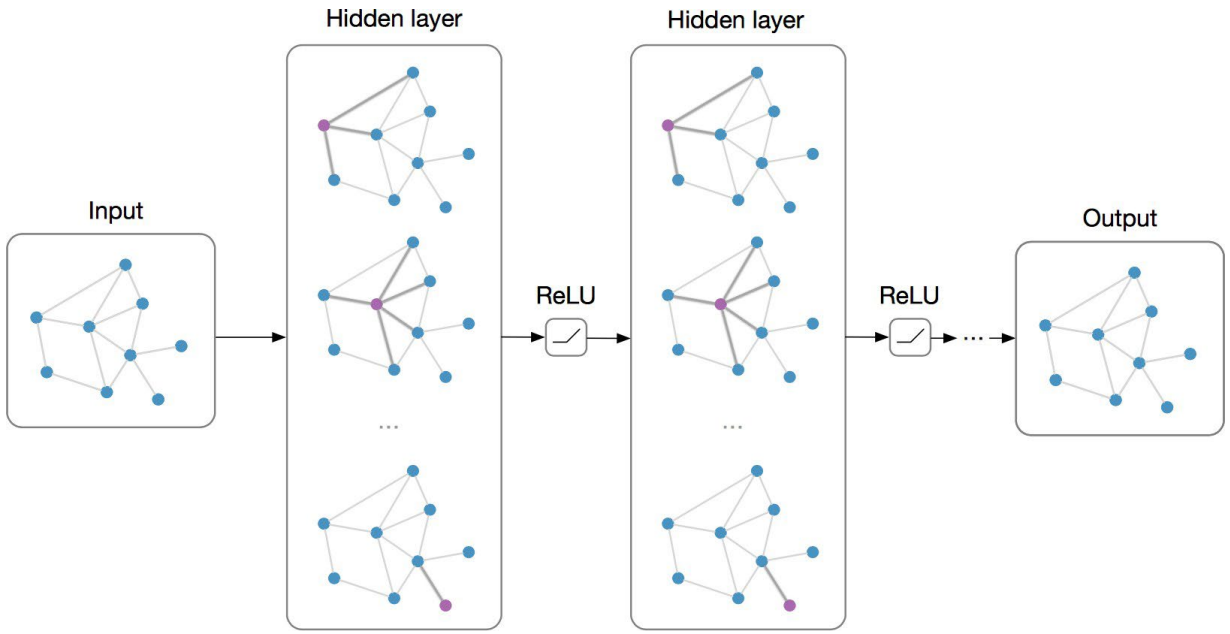
Shengzhong Liu, Shuochao Yao, Yifei Huang, Dongxin Liu, Huajie Shao, Yiran Zhao, Jinyang Li et al. "Handling missing sensors in topology-aware iot applications with gated graph neural network." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, no. 3 (2020): 1-31.

Issue 2: Exploiting Structural Relations

Training the neural network to guess “missing” sensor waveforms from other sensor waveforms. A graph neural network problem.



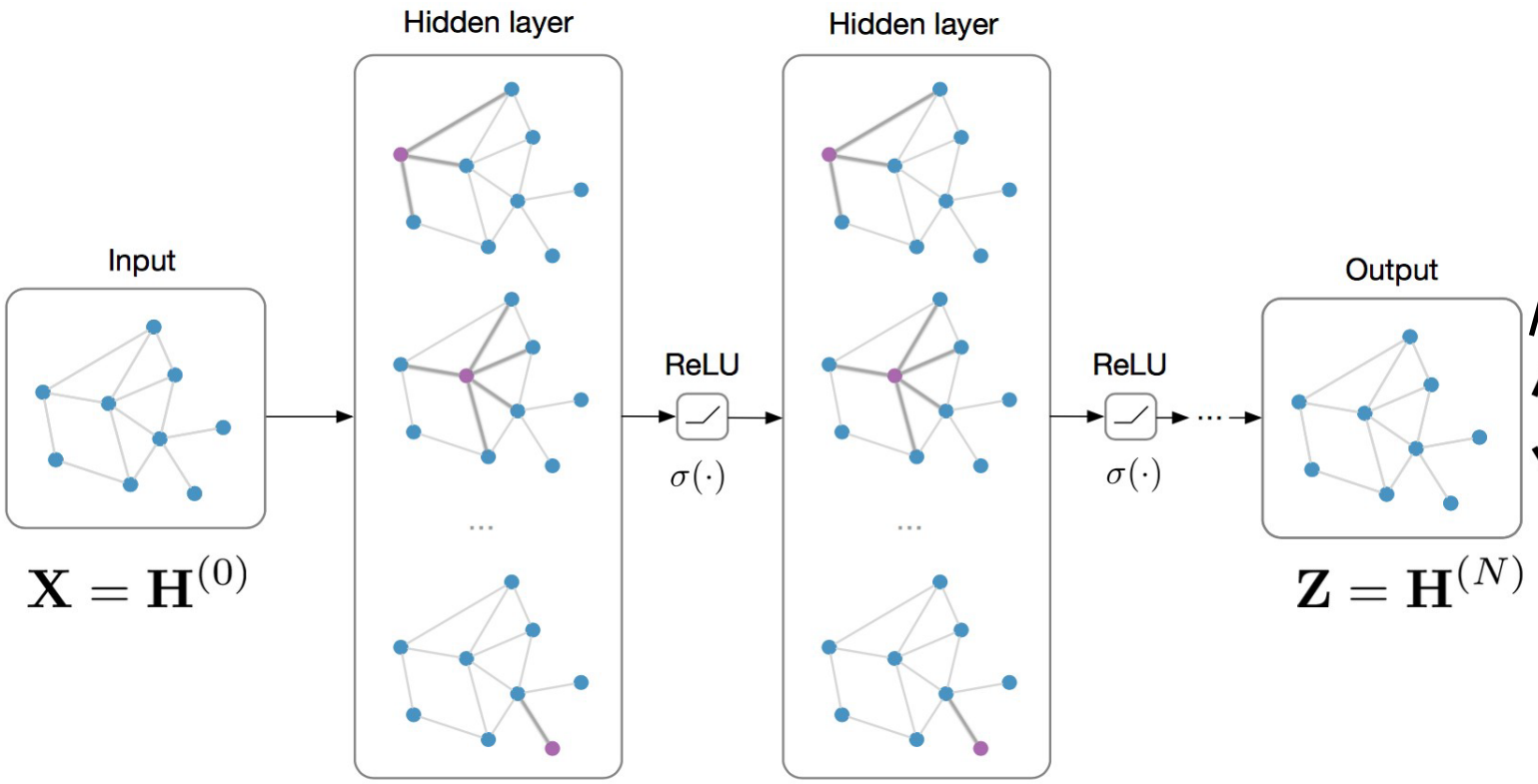
Graph Neural Networks



Main Idea: Pass massages between pairs of nodes and agglomerate
Alternative Interpretation: Pass massages between nodes to refine node (and possibly edge) representations

Graph Neural Networks

Input: Feature matrix $\mathbf{X} \in \mathbb{R}^{N \times E}$, preprocessed adjacency matrix $\hat{\mathbf{A}}$



Reconstruction:

Decode $(\mathbf{z_n})$

Node classification:

$\text{softmax}(\mathbf{z_n})$

e.g. Kipf & Welling (ICLR 2017)

Graph classification:

$\text{softmax}(\sum_n \mathbf{z_n})$

e.g. Duvenaud et al. (NIPS 2015)

Link prediction:

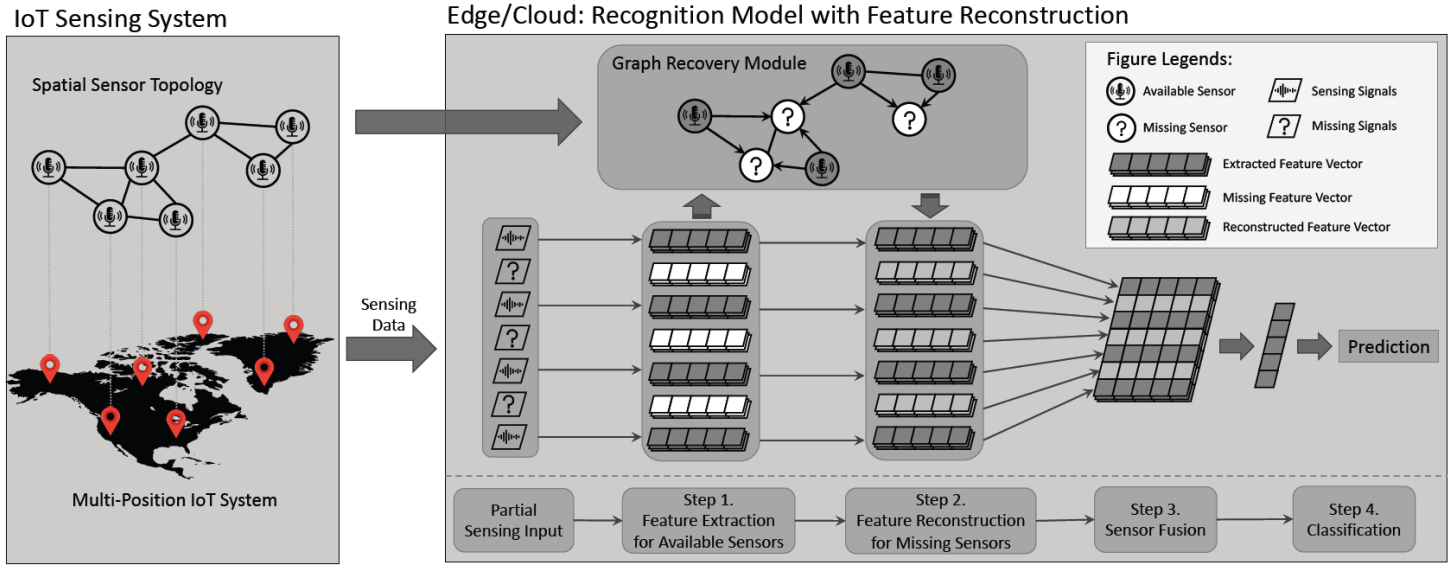
$p(A_{ij}) = \sigma(\mathbf{z_i^T z_j})$

Kipf & Welling (NIPS BDL 2016)

"Graph Auto-Encoders"

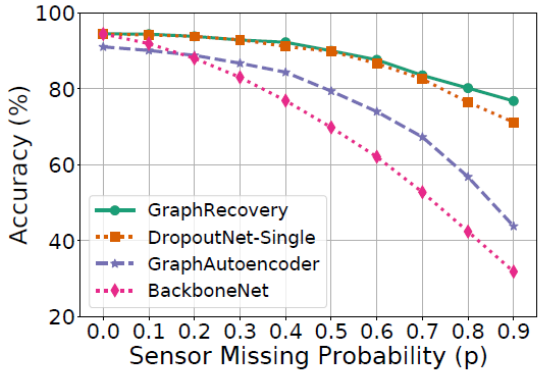
Exploiting Structural Relations

- Use a graph neural network to compute missing node embeddings based on topologically neighboring node embeddings.
- Use a gating mechanism to ensure missing node embeddings do not contaminate neighboring nodes
- Decode sensor streams from reconstructed missing node embeddings

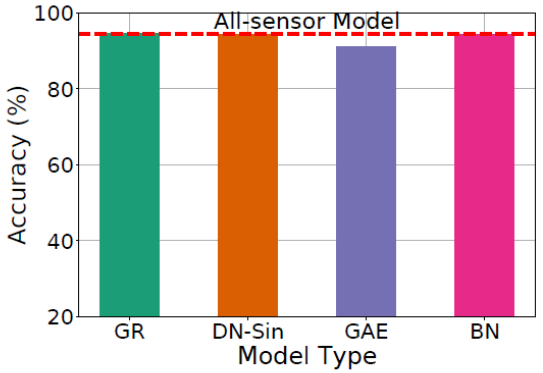


Exploiting Structural Relations

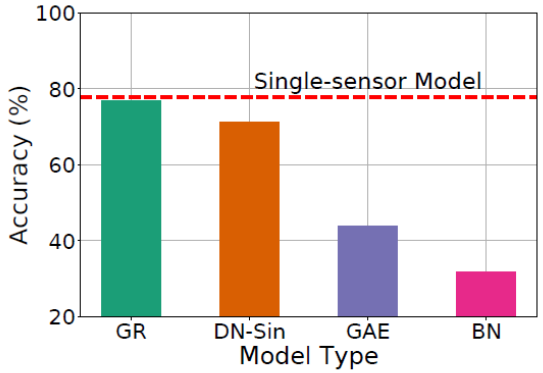
Classification accuracy (in a human activity recognition dataset) from a pool of sensors on different body positions under different sensor "outage" probabilities.



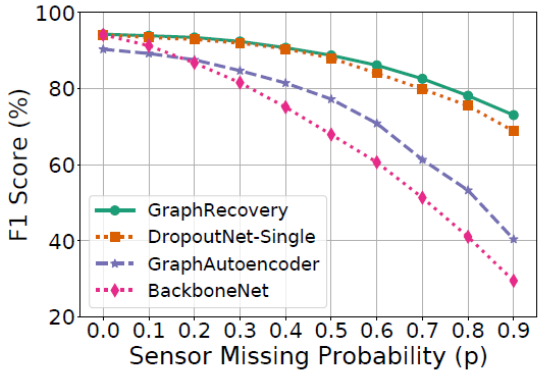
(a) Complete Accuracy Curves.



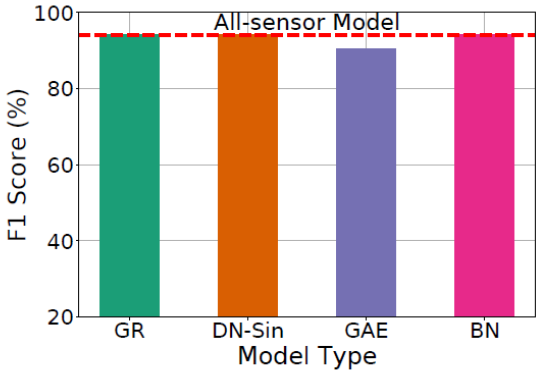
(b) Accuracy when $p = 0$.



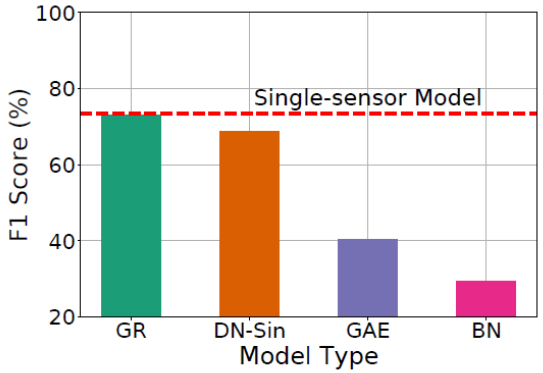
(c) Accuracy when $p = 0.9$.



(d) Complete F1 Score Curves.



(e) F1 score when $p = 0$.



(f) F1 score when $p = 0.9$.

Covered Today: Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***
  - Issue 1: Exploitation of different structural positions (or “vantage points”)
  - Issue 2: Exploitation of relations between different structural positions

Covered Today: Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***
  - Issue 1: Exploitation of different structural positions (or “vantage points”)
  - Issue 2: Exploitation of relations between different structural positions

Limitations (of using an explicit graph structure)?

Covered Today: Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***
  - Issue 1: Exploitation of different structural positions (or "vantage points")
  - Issue 2: Exploitation of relations between different structural positions

Limitations (of using an explicit graph structure)?

The approach learns the particular graph instance (e.g., a particular sensor deployment with a particular sensor layout) but may not generalize to other graphs (e.g., other sensor deployments with a different sensor layout).

Covered Today: Positional Structure in IoT Data

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***
  - Issue 1: Exploitation of different structural positions (or "vantage points")
  - Issue 2: Exploitation of relations between different structural positions

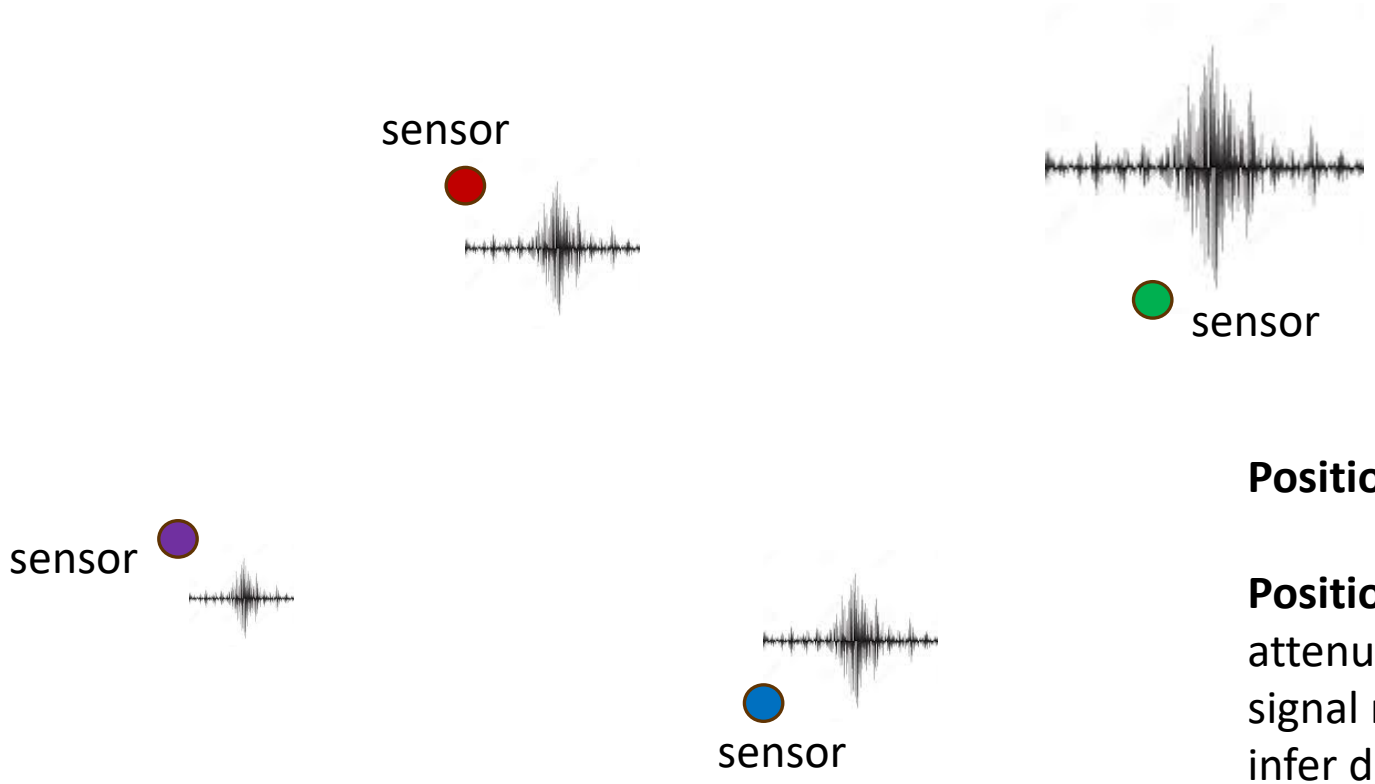
Limitations (of using an explicit graph structure)?

The approach learns the particular graph instance (e.g., a particular sensor deployment with a particular sensor layout) but may not generalize to other graphs (e.g., other sensor deployments with a different sensor layout).

→ Need a global encoding/embedding of sensors' positions that generalizes across deployments, such that learning position/signal relation functions from a number of deployments can directly transfer to a new deployment at a new location and with a new layout.

In the “World before AI”, Generalizable Position Encodings and Position/Signal Relations Have Been Studied at Length!

Example: Target localization

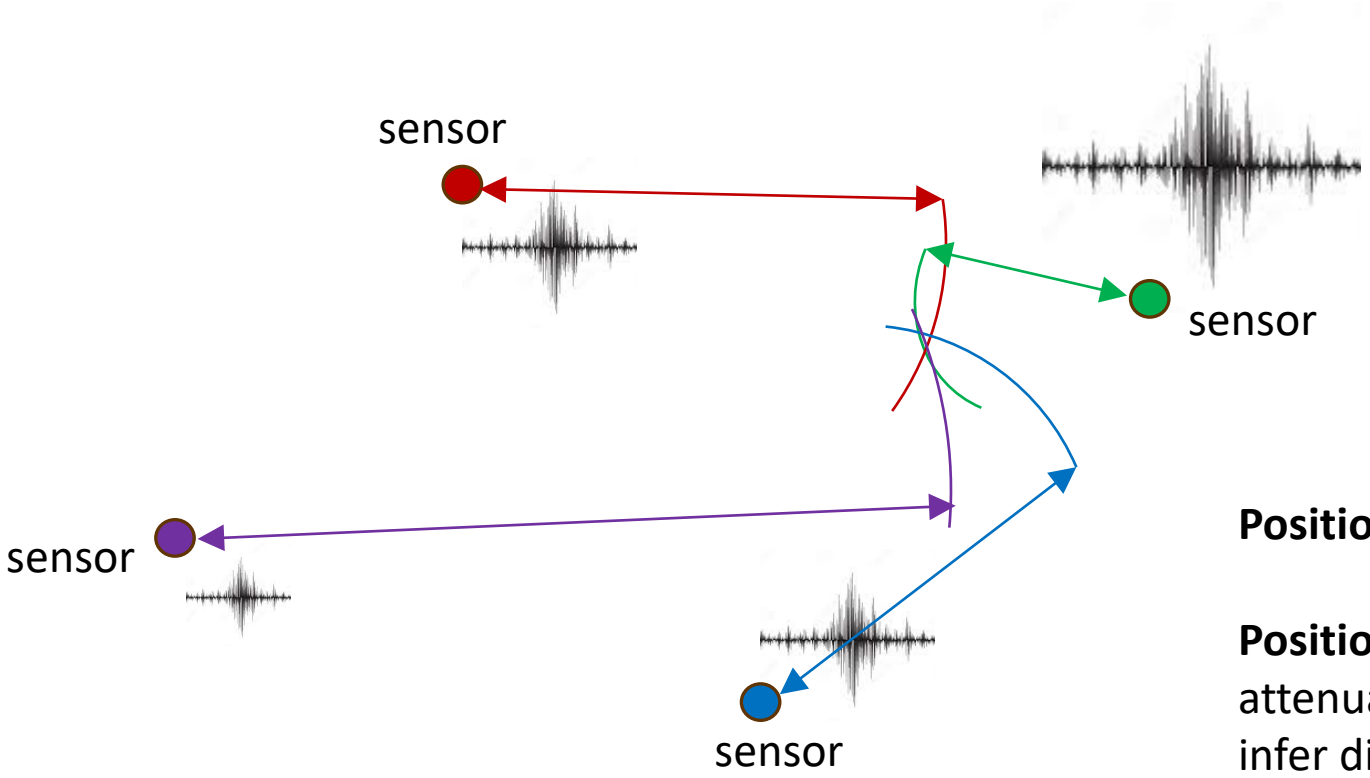


Position encoding: GPS

Position/Signal Relation: Signal attenuates with distance → bigger signal means smaller distance (can infer distance from received signal)

In the “World before AI”, Generalizable Position Encodings and Position/Signal Relations Have Been Studied at Length!

Example: Target localization

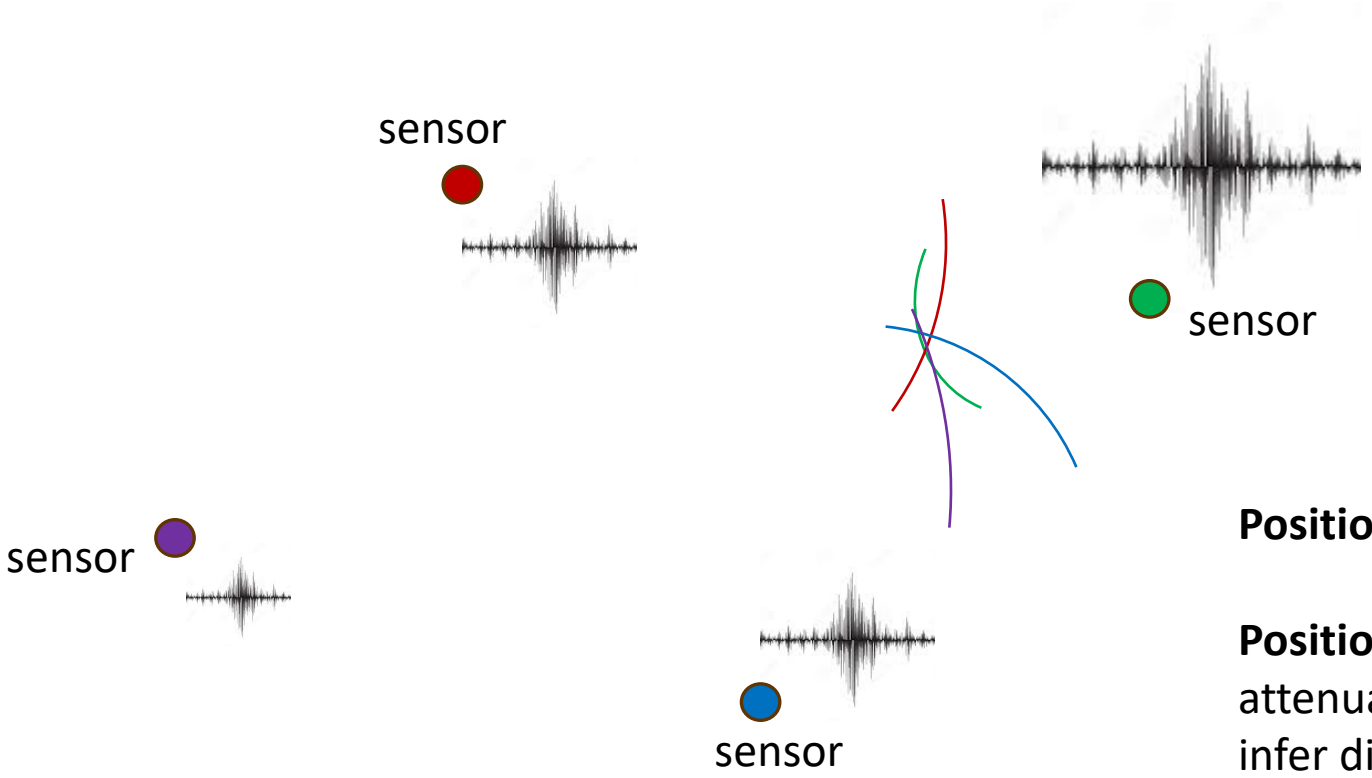


Position encoding: GPS

Position/Signal Relation: Signal attenuates with distance → can infer distance from received signal

In the “World before AI”, Generalizable Position Encodings and Position/Signal Relations Have Been Studied at Length!

Example: Target localization

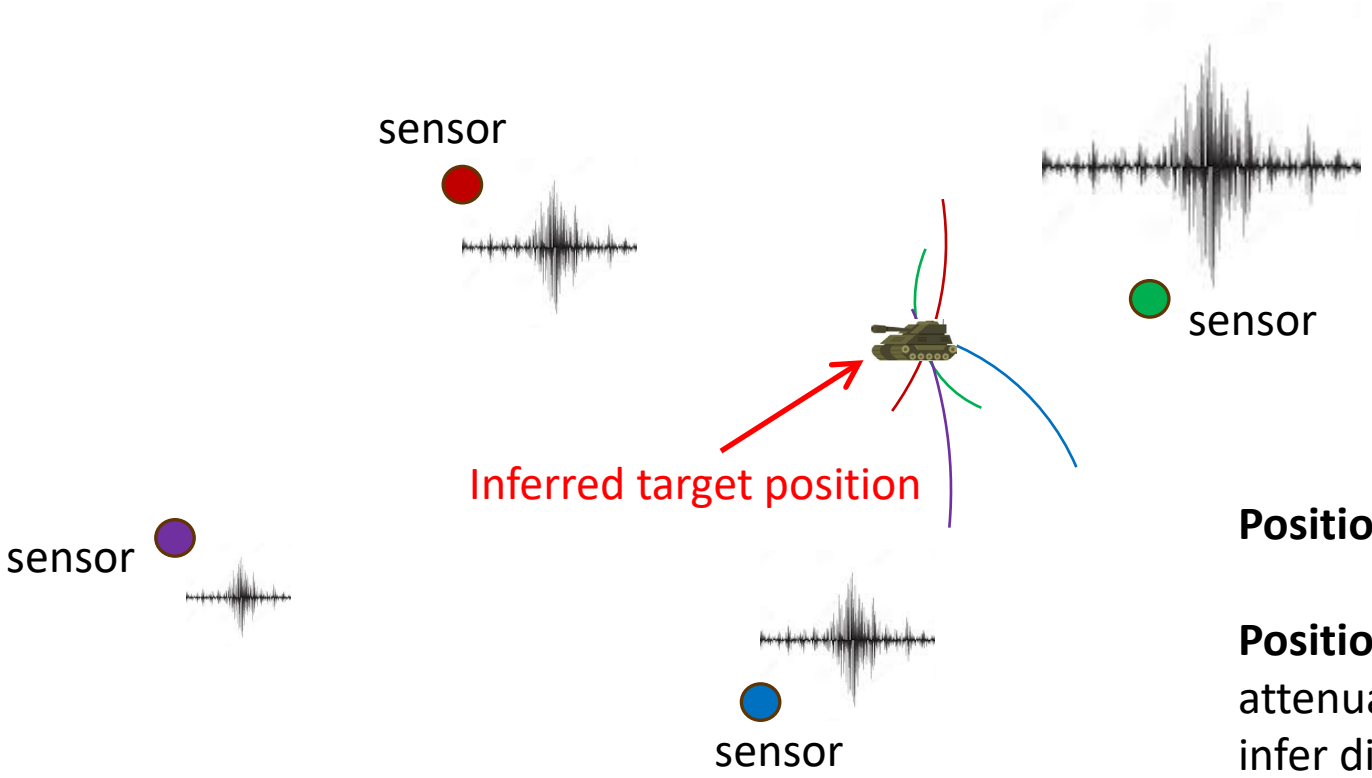


Position encoding: GPS

Position/Signal Relation: Signal attenuates with distance → can infer distance from received signal

In the “World before AI”, Generalizable Position Encodings and Position/Signal Relations Have Been Studied at Length!

Example: Target localization



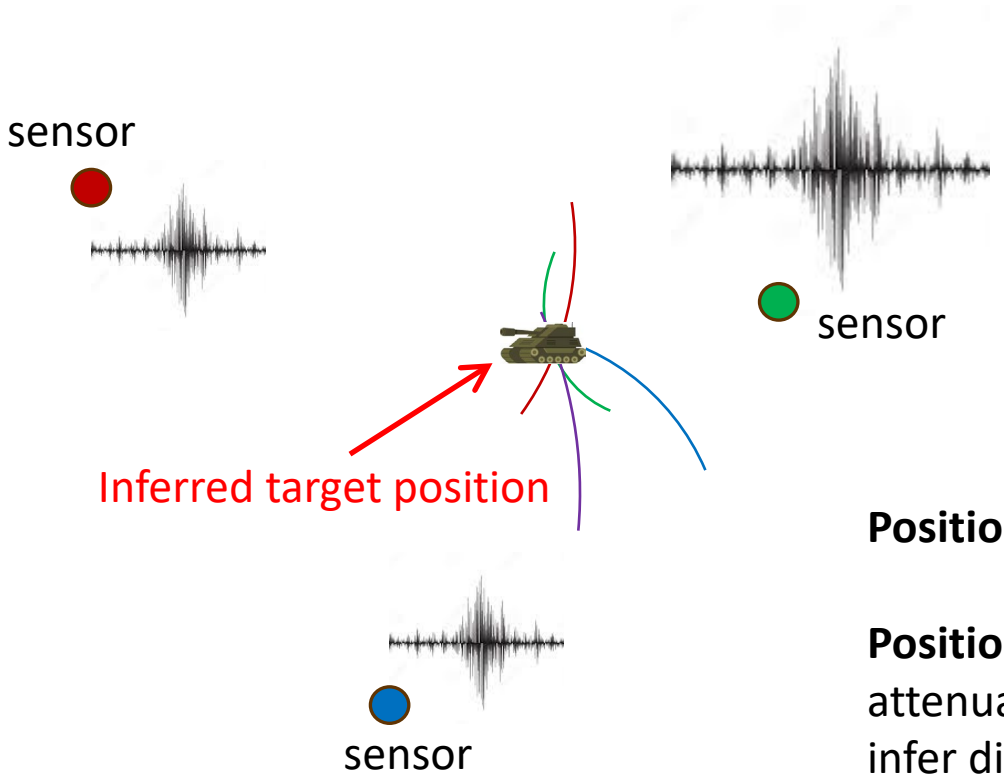
Position encoding: GPS

Position/Signal Relation: Signal attenuates with distance → can infer distance from received signal

In the “World before AI”, Generalizable Position Encodings and Position/Signal Relations Have Been Studied at Length!

Example: Target localization

Note: The localization approach (as a function of sensor positions) generalizes in a “zero shot” manner across different sensor layouts and absolute locations. At present very little work is done on self-supervised learning that achieves the same generalizability across positional embeddings and deployment layouts.



Position encoding: GPS

Position/Signal Relation: Signal attenuates with distance → can infer distance from received signal

Summary

Prior work on learning from spatial-temporal data exploited:

- ***Intra-sample*** spatial structure (e.g., relative layout of objects within an image)
- ***Inter-sample*** structure within the same sensor stream (e.g., spatial-temporal behavior of objects in a video stream)
- Structural relations across ***different sensor streams***
  - Issue 1: Exploitation of different structural positions (or “vantage points”)
  - Issue 2: Exploitation of relations between different structural positions

New solutions are needed for generalized encoding of sensor structural positions and/or geographic locations that allows (zero shot) transfer learning from one deployment layout/structure to another.