Introduction to Distributional Shift

Andrey Malinin

What is distributional shift and why should we care?

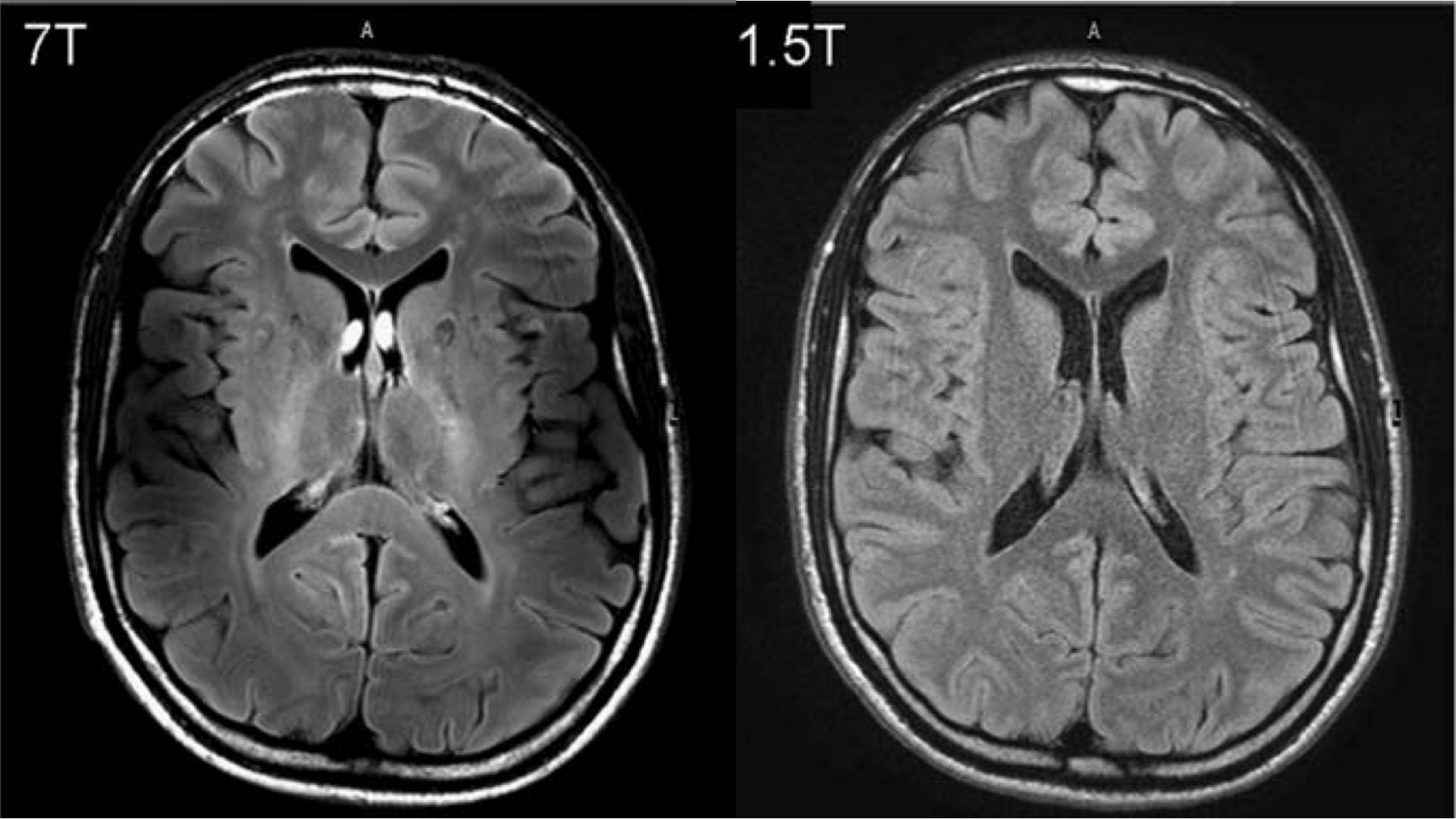
Scenario - Deep learning for high-stakes medical imaging

- You are a good ML specialist and follow all the best practices:
  - Collect all publically available data - yields large dataset
  - Partition it into i.i.d split training, validation and test data
  - Use an implementation of the latest transformer-based model with pre-trained weights
  - Train the model using the Adam optimizer, monitoring validation performance.
- The trained model shows excellent performance on the validation and test sets!
- Excited, you deploy your model on real patient data collected at your **local hospital**…
  - … and the model yields poor performance!
- What happened? Why did the model fail on local patient data, despite good results on test set?

Distributional Shift!

What is distributional shift?

- Distributional Shift is the mismatch between training and evaluation data
  - Models can fail to generalize in when exposed to mismatched data
  - Mismatch can allow learned spurious correlations to detrimentally affect models
- Critical concern for high-risk applications — Medicine, Self-Driving, Finance, etc…
  - Applications where mistakes are costly often are subject to distributional shift
- Distributional Shift is a fundamental challenge in applied Machine Learning
  - Distributional Shift is ubiquitous and highly varied
  - Even humans are also subject to Distributional Shift!


7T
A
1.5T
A

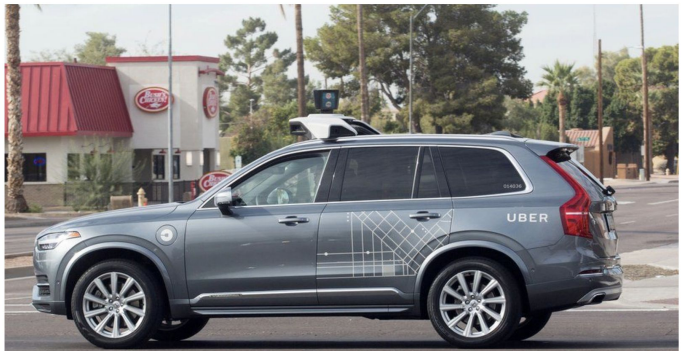
Why should we care?

2 Killed in Driverless Tesla Car Crash, Officials Say

"No one was driving the vehicle" when the car crashed and burst into flames, killing two men, a constable said.

Uber's self-driving operator charged over fatal crash

16 September 2020



NATIONAL

A Tesla driver is charged in a crash involving Autopilot that killed 2 people

January 18, 2022 · 3:00 PM ET

THE ASSOCIATED PRESS

NHTSA probes Tesla Autopilot crash that killed three people

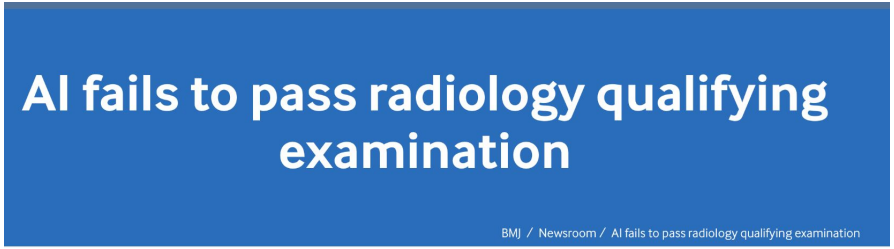
Rebecca Bellan @rebeccabellan / 6:53 PM EDT • May 18, 2022

Comment

Why should we care?

AI May Be More Prone to Errors in Image-Based Diagnoses Than Clinicians

New research indicates that AI may be more prone to making mistakes than humans in image-based medical diagnoses because of the features they use for analysis.



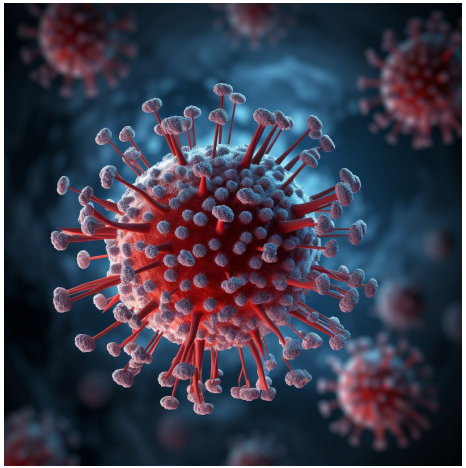
> *"During my brief stint at the innovation arm of the University of Pittsburgh Medical Center, it was not uncommon to see companies pitching AI-powered solutions claiming to provide 99.9% accuracy. In reality, when tested on the internal hospital dataset, they almost always fell short by a large margin." – Sandeep Konam, 2022.*

Shifted and Out-of-Distribution Examples

Types and Degrees of Distributional Shift

- It is extreme to take a binary view of distributional shift
  - Examples are either "in-distribution" of "out-of-distribution"
  - Ex: Previously unseen classes vs known classes
- Broadly, for a discriminative task X -> Y, we can consider two types of shift:
  - Covariate shift: inputs X shift, but targets Y are valid.
  - Task shift: Input X shifts and describes entirely different Y.
- Distributional shift lies on a spectrum between "in-domain" and "out-of-distribution"
  - Degrees of covariate Shift - some changes are less drastic than others
  - Degrees of task shift - some tasks are more closely related than others

Distributional Shift in Medical Imaging: Disease Heterogeneity

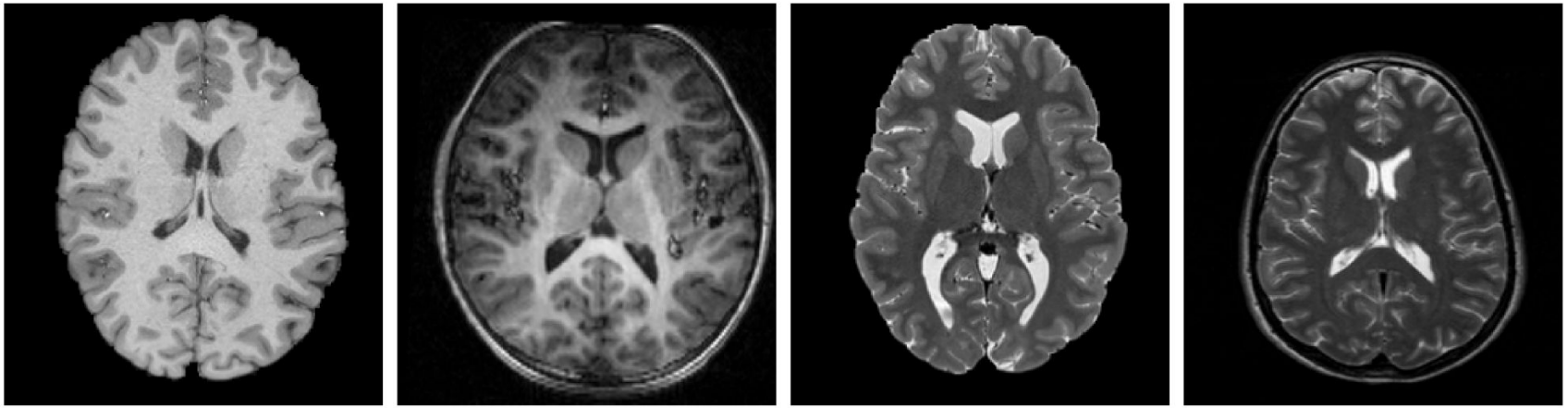


- Symptoms, course and prognosis of disease and prognosis could be highly varied
- Requires a prohibitively large and diverse dataset to sufficiently describe disease

Degrees of Distributional Shift: Covariate Shift

Degrees of Distributional Shift: Covariate Shift

Degrees of Distributional Shift: Covariate Shift

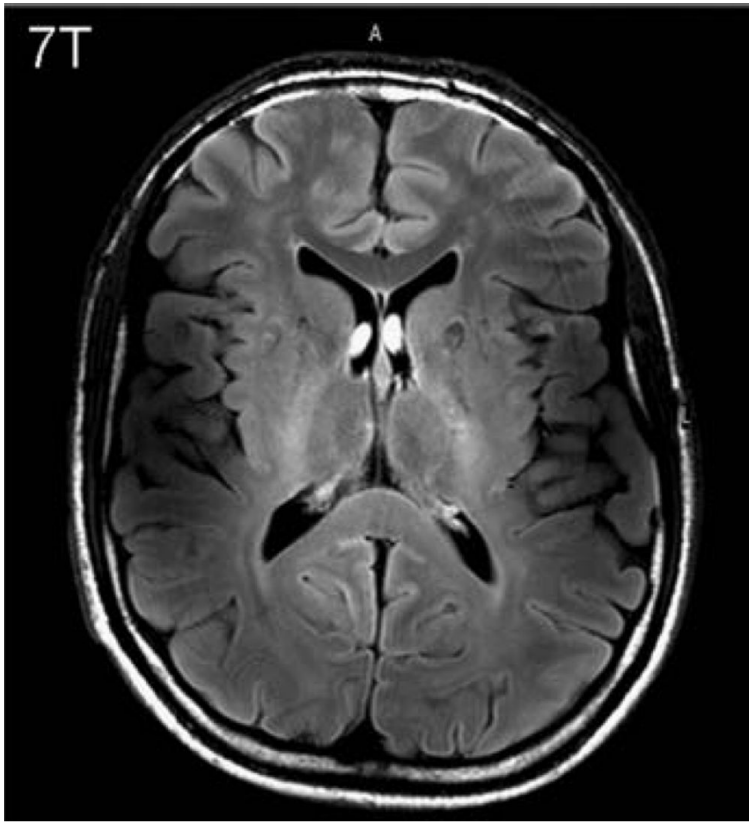


T1-weighted

T2-weighted

Degrees of Distributional Shift: Task Shift

Degrees of Distributional Shift: Task Shift

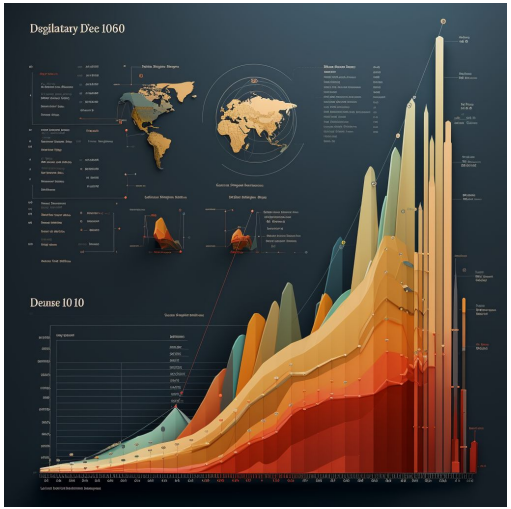


Distributional shift in medical imaging

Distributional Shift in Medical Imaging: Imaging Device

- Diverse range of models from different manufacturers, eg: Philips vs Siemens.
  - Specific devices may differ, due to settings, calibration, wear, deterioration and malfunction
- Operators across different medical centers may have different training and imaging protocol
  - Different operators may use the same device differently

Distributional Shift in Medical Imaging: Demographics



- Genetic, lifestyle and environmental differences may affect how the disease present itself

So what can we do about distributional shift?

What can we do about distributional shift?

- Ideally we would like models which:
  - Are **robust** to distributional shift — good generalization in unfamiliar situations
  - Detect failures of generalisation (errors) via **epistemic** uncertainty estimates
- Covariate shift:
  - Can attempt to **improve generalization** and use epistemic **uncertainty to detect failure** cases.
- Task shift:
  - Can use epistemic uncertainty to detect failure cases and take appropriate action.
  - Challenging to improve generalization (without access to more information).

How to improve robustness to covariate shift?

1. Data Augmentation and Regularisation
    a. Expand range and diversity of data that the model sees
    b. Prevent overfitting to limited data
2. Optimisation towards wider minima
    a. Bias model towards more robust solutions
3. Architectures with good structural priors
    a. Explicitly add appropriate invariances or equivariances
4. Pre-Train + Fine-tuning paradigm
    a. Make use of large amounts of unsupervised data to learn robust representations
5. **Causal Machine Learning**

Causal Machine Learning

- **Key Insight:** mismatch causes errors due to the presence of **spurious correlations**
  - All standard ML models are **correlational** (correlation is not causation!)
  - Correlational model can answer "What does the observed value of X tell us about Y?"
  - Such models can latch onto spurious correlations present in data
- Causal ML aims to model **underlying causal mechanisms** which generated the data
  - Seeks to answer **interventional** and **counterfactual** questions
  - Interventional: "If set X is to a value, keep all else unchanged, how does Y change?"
  - Counterfactual: "If X had been different, what would Y have been?"
- Causal ML promises to yield predictions unaffected by spurious correlations
  - Conceptual Challenge: requires additional assumptions beyond data
  - Practical Challenge: how to combine causal inference with Deep Learning?

What is reasonable to expect?

- There is no free lunch - there are no universal models
  - It is unreasonable to expect a model to be robust to all possible forms of shift
  - It is difficult for a model to generalize to a completely new task (no universal models)
  - Causal ML in early stages and can have practical limitations.
- It should be possible to be robust to a task-specific bounded set of shifts
  - Task variation (low data regime)
  - Shifts due to equivariant transformations
  - Spurious correlations
- The model **will** make an error at some point
  - Long-tail event
  - New form of shift not in the bounded set

How to deal with task shift?

1. Zero-Shot Learning
    a. Same tasks are similar enough that the model may still generalize to a sufficient degree
2. Detect out of task example
    a. Use uncertainty estimates, anomaly scores, etc…
3. Refuse to make prediction
    a. "Example is out of scope, please provide another."
4. Collect, Label and learn / adapt
    a. Collect / extend your training set of newly labelled anomalous examples
5. Other, application-specific actions

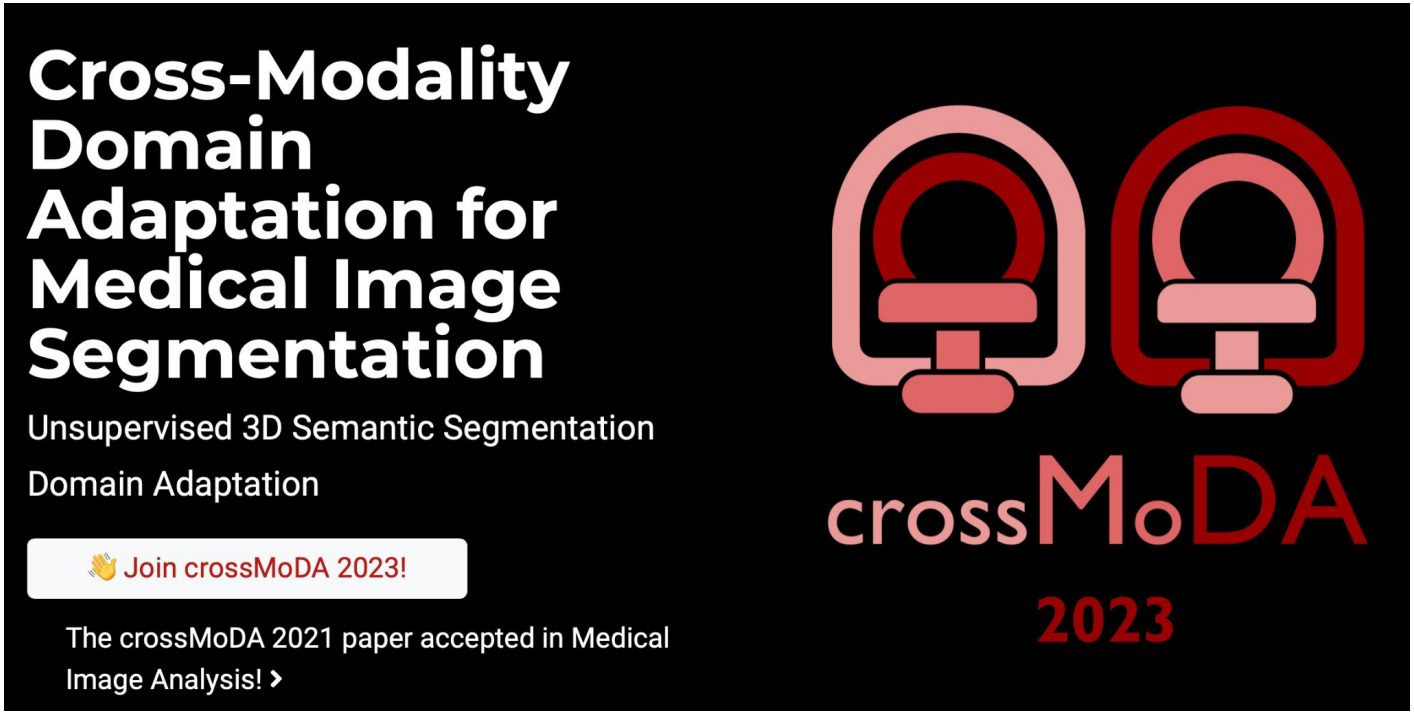
What is reasonable to expect?

- There is no free lunch -
  - There are no perfect anomaly detectors
  - In some tasks, anomaly detectors may have (almost) no margin for error.
- Adaptation can be challenging
  - Mixing tasks / datasets can be non-trivial - in-domain performance can be adversely impacted
  - Labelling may be prohibitively expensive
- It may be unclear how best to provide uncertainty / anomaly information to human user
  - For structured tasks, at which level to predict (ie - per patient, per-voxel, etc…)

What do we have left to work on?

- Generalisation under distributional shift is poorly understood
  - What are the limits of generalisation (no free lunch) ?
  - What level of robustness can we expect?
  - Is there a task-specific bounded set of shifts a model can be robust to?
- Good uncertainty estimation is challenging
  - Bayesian methods, ensembles and non-parametrics are expensive
  - Most approaches developed for classification and sometimes regression
- Most work focuses on image (and text) classification
  - Other predictive tasks and data modalities receive less attention (eg: ASR, NMT, Motion Prediction)
- Few datasets (Wilds, Shifts) contain examples of real distributional shift
  - Most ML datasets contains clean, matched training and test datasets.
  - Popular to contrast small-scale image datasets or add synthetic examples of shift

Related MICCAI Events



https://crossmoda-challenge.ml

Related MICCAI Events

https://sites.google.com/view/causemic

Related MICCAI Events

https://unsuremiccai.github.io/

References

1.