Deep Learning in Medical Imaging: What’s Needed for Training Data?
Francesca Galassi

To cite this version:
Francesca Galassi. Deep Learning in Medical Imaging: What’s Needed for Training Data?. École thématique. France. 2023. hal-04233355

HAL Id: hal-04233355
https://inria.hal.science/hal-04233355
Submitted on 9 Oct 2023
Deep Learning in Medical Imaging: What's Needed for Training Data?

Galassi Francesca, MCF, Empenn lab
Inria Rennes
francesca.galassi@irisa.fr
Deep Learning in Medical Imaging

Applications in Empenn

Galassi Francesca, Empenn lab


Good vs. bad

"Data, data everywhere and not a set to process."
Neil Lawrence
Data Quality

- **Artifacts, noise, and corruption**
- Considerations for **robustness** and **quality control**

Data Variety

- Training on homogenous data may not generalize well to heterogeneous test data.
- Consider adopting a versatile or narrowly focused approach.
Data Volume

---

*How much data do you need?* - Clinician

*How much data do you have?* - DL Researcher

- Traditional statistical rules don’t always fit the machine learning paradigm

- DL has *mechanisms* to *determine* whether more *data* could help
Data Volume

Source of inspiration: Andrew NG machine learning course
Data Readiness

- Are data ready for a DL process?
- Consider using a framework to evaluate readiness

Data Readiness

Are data **ready** for a DL process?

Consider using a **framework** to evaluate **readiness**

- **A**: Annotated, Powdered, Contextual
- **B**: Structured, Visualisable, Quality controlled
- **C**: Access controlled, Anonymised, Ethical clearance
- **D**: Inaccessible, Unknown quality, Non-anonymised

Dataset shift

“Dataset shift is any situation in which the training and test data distributions disagree due to exogenous factors, e.g. dissimilar cohorts or inconsistent acquisition processes.”

Daniel C. Castro
Predictive Modeling

- Given an image $X$, train a model to predict some annotation $Y$

$$P(Y|X)$$

Assumptions:

- Availability of sufficient training data $(X, Y)$ – data volume
- Consistency in training and testing data distribution – data variety

Challenges in DL for medical imaging: Data Scarcity & Data Mismatch
Image and Annotation – a causal perspective

What is the relationship between image $X$ and annotation $Y$?

$cause \ X \rightarrow \ Y \ effec\ causal$

predict effect from cause

Example: MS Lesion segmentation

$X$ - structural brain MRI scan

$Y$ - manual segmentation masks

Dataset Shift – data mismatch

- **Training** and **test** data distributions may **disagree** due to ...

---

**Population shift**
- \( Z \): unobserved true anatomy
- \( X \): image
- \( Y \): annotation

**Acquisition shift**
- \( Z \): unobserved true anatomy
- \( X \): image
- \( Y \): annotation

**Annotation shift**
- \( Z \): unobserved true anatomy
- \( X \): image
- \( Y \): annotation

---

Domain Adaptation – a case study

- Unsupervised DA – unlabeled target data accessible at training time

\[ \min_{\gamma, f, g} \sum_i L_s( y^s_i, f(g(x^s_i))) + \sum_{ij} Y_{ij} \| g(x^s_i) - g(x^t_j) \|^2 + \beta L_t( y^s_i, f(g(x^t_j))) \]

the coupling matrix

Annotation Shift – a case study

- Considerable inter- and intra-rater variability observed
- High precision and low sensitivity of individual experts

Key Takeaways

- Data quality, volume and variety matter

- Error analysis guides model improvements and data collection
  
  - A data readiness framework can help assessing data suitability for a DL process

- Dataset shifts arise from factors like population, acquisition, and annotation variations
  
  - Causal diagram can help describing the shift in the specific project

- Consider strategies for addressing these challenges, i.e., label noise reduction and domain adaptation