Large-scale Labeled Datasets to Fuel Earth Science Deep Learning Applications

Manil Maskey and Rahul Ramachandran
NASA/MSFC

J.J. Miller
University of Alabama in Huntsville

DEEP LEARNING

- A subfield of machine learning
- Algorithms inspired by function of the brain (ANN)
- Scales with amount of DATA (training)
- Powerful tool without the need for feature engineering
- Suitable for Earth Science applications

RECENT DEEP LEARNING SUCCESS

- Facebook
  - Translates about 2 billion user posts per day in more than 40 languages
  - Photo search and photo organization
- Microsoft
  - Speech-recognition products: Bing voice search, X-Box voice commands
  - Search rankings, photo search, translation systems
- Google
  - Almost all services
- Medical Science
  - Diagnosis Language translation
- Playing strategy games
- Self driving cars

WHAT IS NEEDED?

- One thing in common
  - Large number of data points needed to learn large number of parameters in the model that machines have to learn
- Barrier for using deep learning
- ~~Data~~ **Training Data** is the New Oil
- Manually creating labeled training data is bottleneck

EXAMPLES

	VGGNET	DeepVideo	GNMT
Task	Classify image	Classify video	Translate
Input Data	Image	Video	English Text
Output	1000 Classes	47 Classes	French Text
# of Parameters	~140 million	~100 million	~380 million
Labeled Data Size	1.2 million images	1.1 million videos	6 million sentence pairs 340 million words

DEEP LEARNING FOR EARTH SCIENCE APPLICATIONS AT MSFC

- Hurricane intensity (wind speed) estimation
- Severe storm (hailstorm) detection .. Forecast?
- Transverse bands detection
- Dust climatology
- Phenomena identification
- Ephemeral water detection

LABELED TRAINING DATA

Application	Training Data Size ~	Methodology
Hurricane intensity (wind speed) estimation	49,000	Combining imagery with storm database
Severe storm (hailstorm) detection	93,000	Storm reports
Transverse bands detection	9,000	Manual
Dust climatology	8,000	Manual
Ephemeral water detection	650,000	Combining shapefiles and time series analysis

STRATEGIES?

- Data Augmentation
- Transfer Learning
- Permutation Invariance
- Data Programming

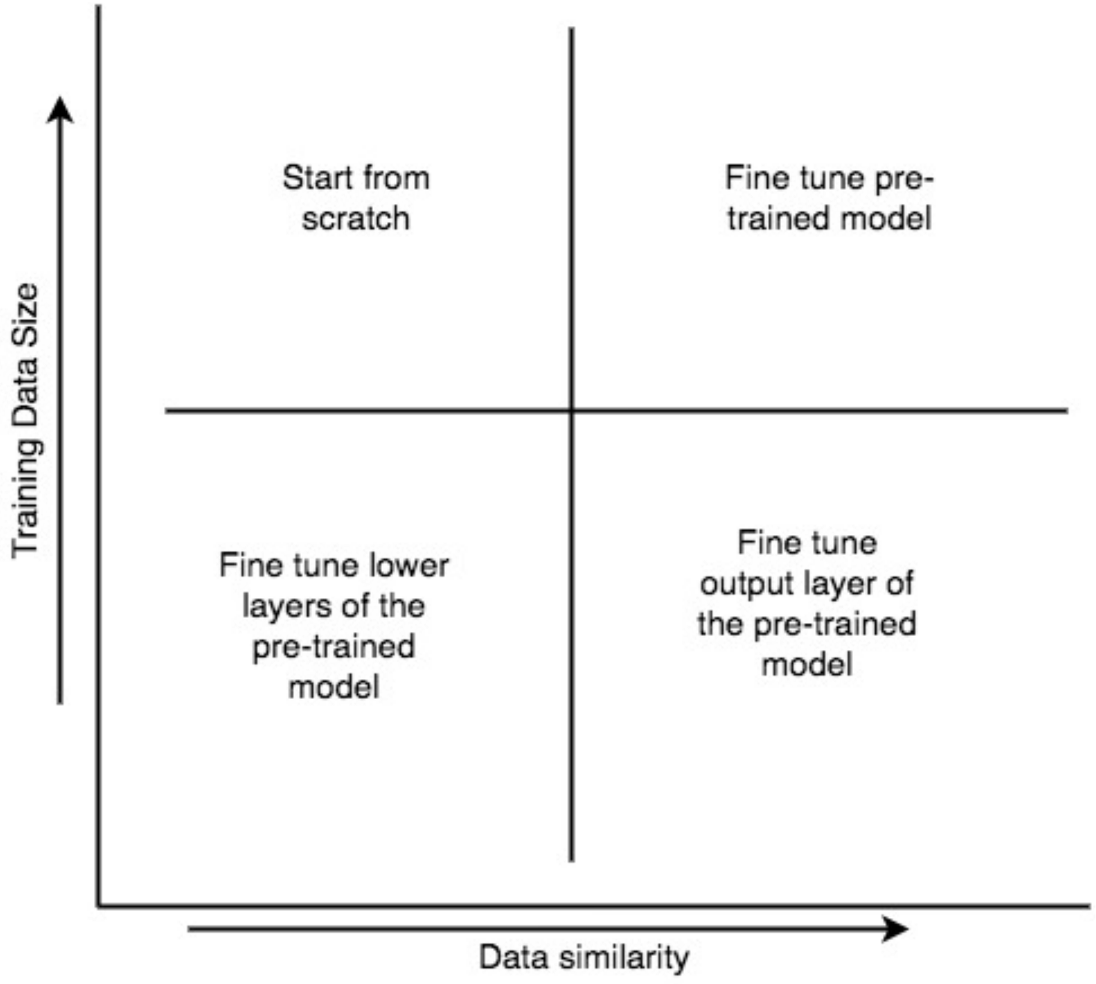
DATA AUGMENTATION

- For computer vision tasks
- Mirroring
- Random cropping
- Color shifting
- PCA

TRANSFER LEARNING

- Network gains knowledge from training data
- Compiled as “weights” of the network
- Weights can be extracted and then transferred to another network
- Instead of training network from scratch, “transfer” the learned features
- Pre-trained model
  - Created by someone else to solve similar problem
- Ways to fine tune the model
  - Feature extraction
  - Architecture
  - Train some – freeze some

USING PRE-TRAINED MODELS



PERMUTATION INVARIANCE

- Example:

$$f(x_1,x_2,x_3) = f(x_2,x_1,x_3) = f(x_3,x_1,x_2) = \ldots$$

- Represent data that does not have spatial relationship

DATA PROGRAMMING

- Programmatic creation of training dataset

- User
  - Provides unlabeled data
  - Writes labeling functions (LFs) – weak supervision
    - expresses supervision strategies
  - Chooses a discriminative model

WEAK SUPERVISION

- Domain rules/heuristics
- Existing ground-truth data that is not exact fit (distant supervision)
- Weak classifiers ("boosting")
- Non-expert annotations ("crowdsourcing")

EXAMPLE

- Information Extraction from Earth Science Literature
- Unstructured text
- Extract information: dataset usage, hypothesis validation, etc.
- No large labeled training dataset
- Various ontologies, vocabularies, and glossaries?
- Custom heuristics?
- Regular expressions
- Rule-of-thumb
- Negative label generation

STUDYING DUST EVENTS

Sample text:

"Meteorological conditions during dust storms were analyzed using aerosol."

```python
def labelingFunction1(input):
    concept = (input.phenomenon,input.property)
    return 1 if concept in DOMAIN_KB else 0


def labelingFunction2(input):
    found = re.search(r'.*analyzed.*',input.text.between)
    return 1 if found else 0
```

Sample Labeling Functions to extract mentions of dust events and physical properties

- labelingFunction1: Leverage existing Earth Science knowledgebase (e.g., SWEET)
- labelingFunction2: Domain heuristics

SNORKEL

- Data programming framework
  - Training data creation and management
- Creates a noisy training set – by applying LFs to data
- Learns a model of the noise (learns accuracy of LFs)
- Trains a noise-aware discriminative model

LABELED TRAINING DATA

Application	Training Data Size ~	Methodology	Strategy
Hurricane intensity (wind speed) estimation	49,000	Combining imagery with storm database	Data Augmentation
Severe storm (hailstorm) detection	163,000	Storm reports	None
Transverse bands detection	9,000	Manual	Data Augmentation and Transfer Learning
Dust climatology	8,000	Manual	Data Augmentation and Transfer Learning
Ephemeral water detection	650,000	Combining shapefiles and timeseries analysis	None

PUBLISHING DATASET

- Should Earth science training dataset be published as traditional datasets?
- Catalog – NASA CMR?

Available Public Datasets on AWS

Geospatial and Environmental Datasets

Learn more about working with geospatial data on AWS at Earth on AWS.

- Landsat on AWS: An ongoing collection of satellite imagery of all land on Earth produced by the Landsat 8 satellite.
- Sentinel-2 on AWS: An ongoing collection of satellite imagery of all land on Earth produced by the Sentinel-2 satellite.
- GOES on AWS: GOES provides continuous weather imagery and monitoring of meteorological and space environment data across North America.
- SpaceNet on AWS: A corpus of commercial satellite imagery and labeled training data to foster innovation in the development of computer vision algorithms.

TAKEAWAYS

- Deep learning is ideal for “supervised” learning
- Algorithms can be fine tuned for customized applications
- **Large labeled datasets** fuel impressive classification accuracy
- **Challenge**:
  - Creating/Identifying/Accumulating large labeled datasets
- **Addressing Limited Labeled Data**
  - **Many approaches – depends on application**

CONTACT

Manil Maskey
manil.maskey@nasa.gov