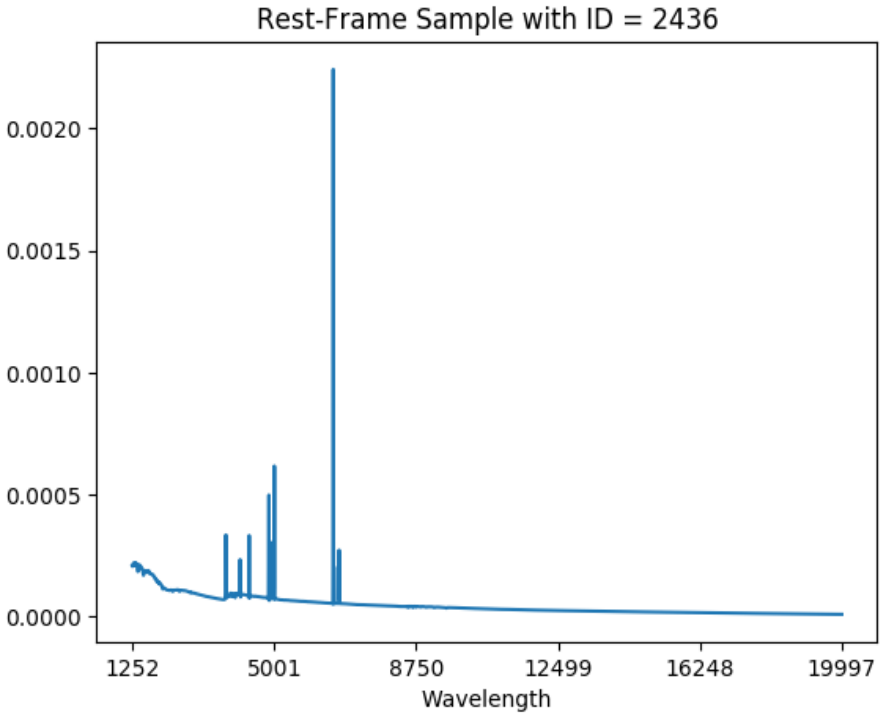
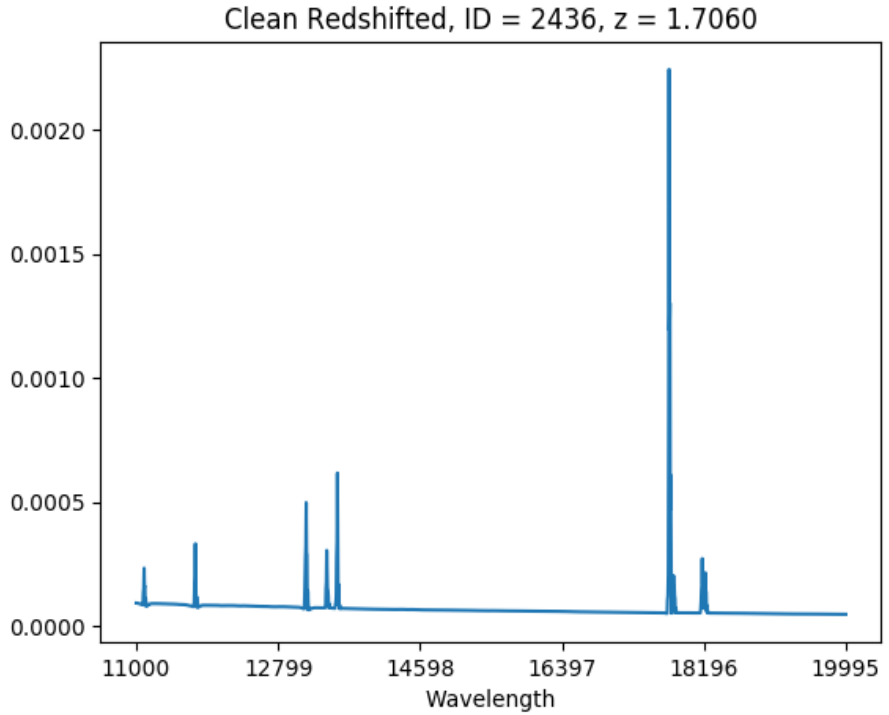
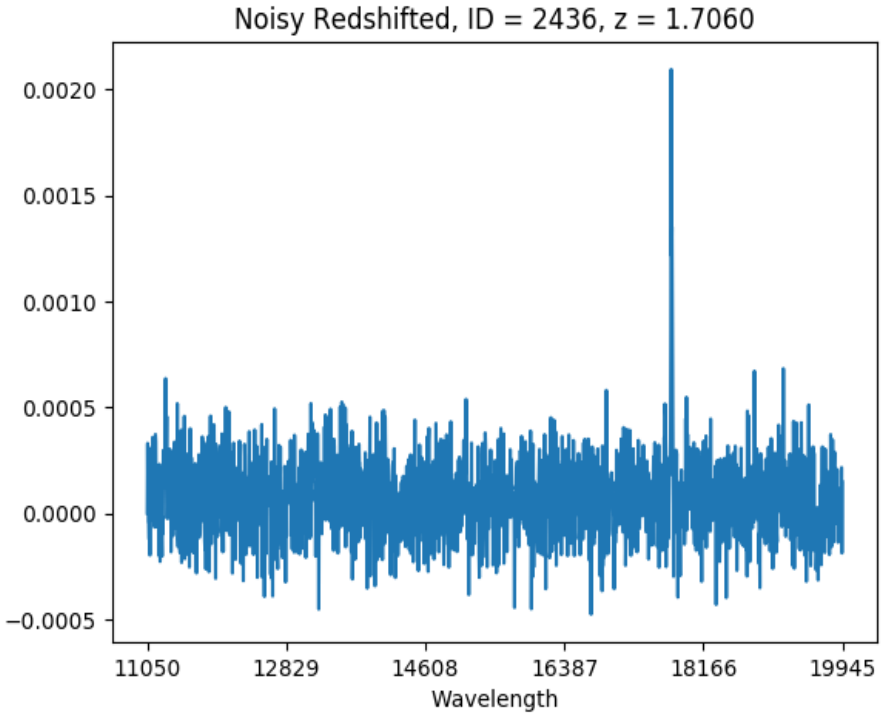
CNN FOR SPECTROSCOPIC REDSHIFT ESTIMATION ON EUCLID DATA

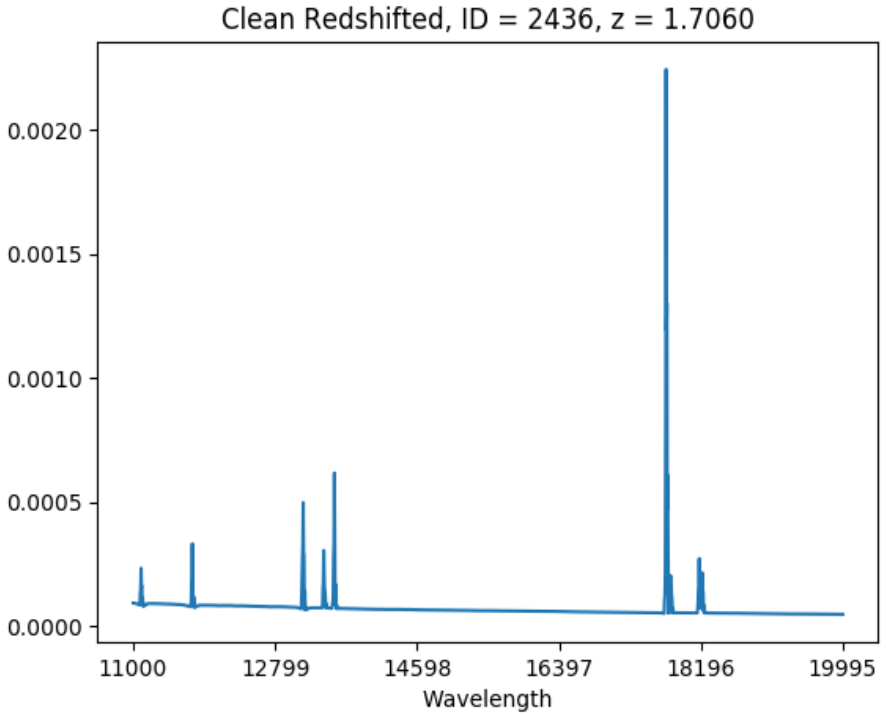
EUCLID SEDs


Rest-Frame Sample with ID = 2436
0.0020
0.0015
0.0010
0.0005
0.0000
1252
5001
8750
12499
16248
19997
Wavelength


Clean Redshifted, ID = 2436, z = 1.7060
0.0020
0.0015
0.0010
0.0005
0.0000
11000
12799
14598
16397
18196
19995
Wavelength

EUCLID SEDs





low SNR of observations

+

Big Data

}

no reliance on human annotation

+

need for reliable automated methodologies

PREDICTIVE MODEL

Redshift Estimation

Real-valued, non-negative number (z)

Regression Analysis

- not very robust algorithms
- not state-of-the-art

Split the examined redshift interval into ordinal classes, based on Euclid's characteristic resolution

Classification Problem

- utilization of state-of-the-art classifiers, that fully harness the power of Big Data
- probabilistic approach using Softmax - quantification of the levels of confidence for each estimation

DATASET & METHODOLOGY

- 10K initial clean rest-frame spectral profiles (SEDs)

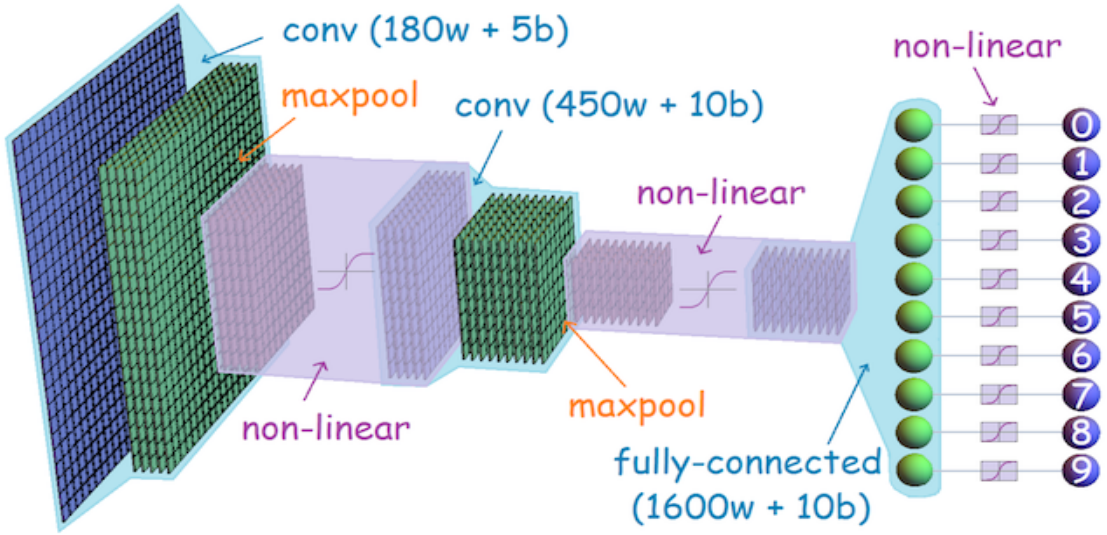
- <u>Generation of randomly redshifted examples</u>
  - $z = [1, 1.8)$, similar to Euclid's specification
  - $\log(1 + z) = \log(\lambda_{observed}) - \log(\lambda_{emit}) \quad \leftrightarrow \quad 1 + z = \frac{\lambda_{observed}}{\lambda_{emit}}$
  - optional addition of white Gaussian noise (idealistic vs. realistic)

- <u>Quantization of the utilized redshift range</u>
  - split into 800 discrete classes, which implies a resolution of 0.001

R. Stivaktakis, G. Tsagkatakis, B. Moraes, F. Abdalla, J-L Starck, P. Tsakalides, "Convolutional Neural Networks for Spectroscopic Redshift Estimation on Euclid Data", IEEE Transactions on Big Data (Special Issue on Big Data From Space, 2020)

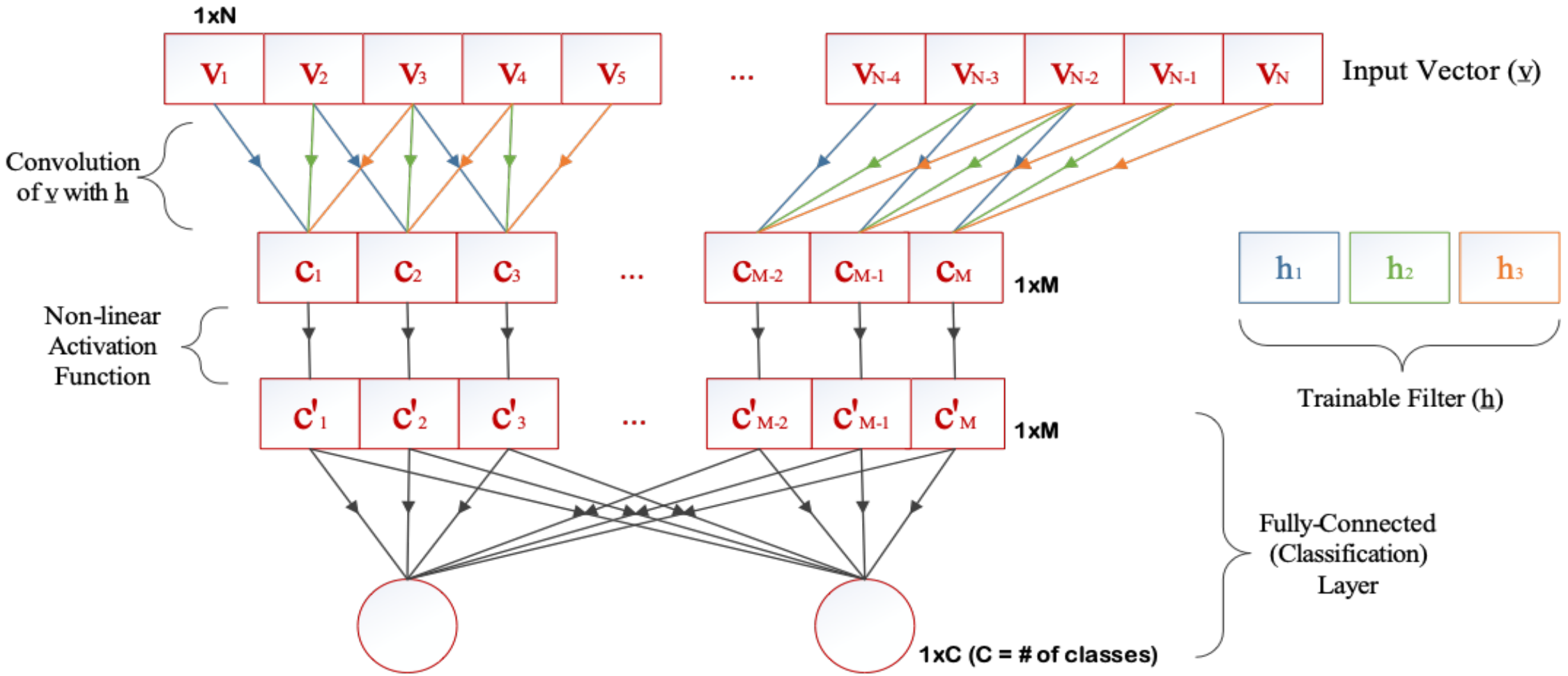
CONVOLUTIONAL NEURAL NETWORKS

- inspired by the concept of “visual receptive fields’
- automated feature extractors
- exhibit spatial correlations of the given input



- less prone to overfitting
  - local-connectivity property & weight-sharing lead to a dramatic decrease in total parameters
  - effective utilization of Big Data results in a high-generalization capacity
  - Dropout & Batch Normalization

1-DIMENSIONAL CNN



EXPERIMENTAL SETUP

Hyper-parameters

- 3 Conv. + ReLU Layers
- 1 Dense + Softmax Layer
- Kernel size = 8, Stride = 1
- # Filters = 16 per Conv. Layer
- Categorical Cross-Entropy Loss

Sample Size

- 400,000 training samples
- 10,000 CV samples
- 10,000 testing samples

Compared Classifiers

- k Nearest Neighbors
- Random Forests
- Support Vector Machines

SOME IDEAS….

- Same input -> different outputs
- Same outputs <- different inputs
- Different model architectures (ML, Bayesian, etc.)
- Noise in labels and/or input
- Uncertainty labels and/or input
- Classification vs regression
- Loss functions
- Few-shot learning / Imbalanced labels
- ….