A Deep Representation for Invariance and Music Classification

Chiyuan Zhang, Georgios Evangelopoulos, Stephen Voinea, Lorenzo Rosasco, Tomaso Poggio.

Center for Brains, Minds and Machines (CBMM)
Computer Science and Artificial Intelligence Laboratory (CSAIL)
Laboratory for Computational and Statistical Learning (LCSL)
Massachusetts Institute of Technology (MIT)
Istituto Italiano di Tecnologia (IIT)

ICASSP 2014, May 9, 2014, Florence, Italy
A fast learning algorithm for deep belief nets
Cited by 1582 Related articles All 63 versions Save More

[PDF] ePortfolios: Digital Stories of Deep Learning
H Barrett - Work, 2001 - electronicportfolios.com
... Deep Learning (learning for real comprehension) comes from a sequence of –Experience –Reflection –Abstraction –Active testing ... 400 words] – Told in their own voice [record script] – Illustrated (mostly) by still images – Music track to add ... Storytelling as a Theory of Learning ...
Cited by 220 Related articles All 5 versions Save More

[PDF] Learning Features from Music Audio with Deep Belief Networks.
P Hamel, D Eck - ISMIR, 2010 - ismir2010.ismir.net
ABSTRACT Feature extraction is a crucial part of many MIR tasks. In this work, we present a system that can automatically extract relevant features from audio for a given task. The feature extraction system consists of a Deep Belief Network (DBN) on Discrete Fourier ...
Cited by 54 Related articles All 5 versions Save More

[CITATION] A constructivist view of music education: Perspectives for deep learning
S Scott - General Music Today, 2006 - gmt.sagepub.com
Constructivists believe that learning is a social act where students interpret new understandings of their worlds in relation to previous knowledge and experience. When viewed from this perspective, the classroom becomes a community of individuals working ...
Cited by 14 Related articles Save More

H Lee, PT Pham, Y Largman, AY Ng - NIPS, 2009 - papers.nips.cc
Go to "http://www.google.com/profiles" in the first-layer features, which justifies the use of...
What are deep (convolutional) neural networks doing?

Why convolution & pooling?

Why hierarchy / multi-layer?
Related Work

Empirical Investigation

▶ Visualization (M. Zeiler, R. Fergus 2013, ...)
▶ Convolutional vs non-convolutional (...)
▶ Deep vs Shallow architecture (L. Ba, R. Caruana 2013, ...)

Mathematical Justification

▶ Signal recovery from Pooling Representations (J. Bruna, A. Szlam, Y. LeCun 2014)
▶ Deep Scattering Spectrum (J. Andén, S. Mallat 2013)
▶ Invariant Representation Learning (F. Anselmi, J. Leibo, L. Rosasco, J. Mutch, A. Tacchetti, T. Poggio 2013)
▶ ...
What are deep (convolutional) neural networks doing?

Why convolution & pooling?

Why hierarchy / multi-layer?
(Deep) Representation Learning

- What are deep (convolutional) neural networks doing?
- Why convolution & pooling?
- Why hierarchy / multi-layer?

- Learning invariant representation
- Removing task-irrelevant variability
- Hierarchy of different scales / invariance
Outline

- Basic Theory
  - invariant representation
- Neural Realization
  - computational modules / networks based on neuron primitives
- Evaluation
  - music genre classification on GTZAN
Properties of a “good” data representation

- Invariant (to identity-preserving transformations / variability), for representation $R$, signal $x$ and (irrelevant) transformation $G$

$$R(x) = R(g \circ x), \quad \forall x \in \mathcal{X}, g \in G$$

- Discriminative (will not map objects from different classes to the same representation)

$$R(x) \neq R(x') \text{ iff } \nexists g \in G, \text{ s.t. } x' = g \circ x$$

- Stable (Lipschitz continuous)

$$\|R(x) - R(x')\|_R \leq L\|x - x'\|_X, \quad L > 0$$
A model for (compact) group transformation. Example for group transformation: (tempo) scaling, (pitch) shifting / translating.

A group $G$ partitions the signal space $\mathcal{X}$ into equivalent classes (orbits), for any $x \in \mathcal{X}$:

$$[x] = \{ g \circ x : g \in G \}$$

The orbit itself is

- invariant: $[x] = [g \circ x]$, $\forall x \in \mathcal{X}, g \in G$
- discriminative: $[x] \neq [x'] \Leftrightarrow \exists g \in G, s.t. x' = g \circ x$
Basic Theory

The orbit (a set of signals) could be characterized by the probability distribution supported on it. This could be characterized by projections onto unit vectors (Cramer-Wold 1936).
Neural Realization

- \([x] = \{g \circ x : g \in G\}\)
- \(\iff p_x \text{ supported on } [x]\)
- \(\iff p_{\langle t, x \rangle} \text{ for templates } t \text{ sampled from the unit sphere}\)
- \(\langle t, g \circ x \rangle = \langle g^{-1} \circ t, x \rangle \text{ for unitary groups}\)

Algorithm

Fix (random) templates \(t_1, \ldots, t_K\), for an input signal \(x\):
- compute \(\langle g \circ t_k, x \rangle\) for all \(k = 1, \ldots, K\) and \(g \in G\)
- compute (1-D) histogram over the inner-product values for each template \(t_k\)
- concatenate all the histograms
Remarks

- To compute $\langle g \circ t_k, x \rangle$, we only need to observe $x$, instead of all transformed version of $g \circ x$.  
- Learning is implemented by memorizing the “random” templates and their transformed versions $g \circ t_k$, for $g \in G, k = 1, \ldots, K$
- Only basic neuron primitives are used in the feature computation  
  - High-dimensional inner-product (templates are stored as the weights in the synapses of the neurons)  
  - Non-linearity (could be used to implement histogram counting)
- This representation map is Lipschitz continuous
Invariance Module (Simple-Complex Neurons)
Generalization

- Partially Observable Group: pool over a subset of the group, get *partially* invariant representation
  - Limited receptive field size
  - Non-compact group

- Non-group smooth transformations: sample *key transformations* and linearly approximate the orbit locally at each key transformation
Music Genre Classification

- Base representation is spectrogram (370 ms)
- Three layers of invariance module cascades
  - Time warping
  - Local translation in time
  - Pitch shifting
Experiment Setup

GTZAN Dataset
► 1000 audio tracks, each 30 seconds long
► Some tracks contain vocals
► 10 music genres
   – blues, classical, country, disco, hiphop, jazz, metal, pop, reggae and rock

Baseline Features
► Mel-Frequency Cepstral Coefficients (MFCCs)
► Scattering Transform (J. Andén, S. Mallat 2011)
## Classification Results

<table>
<thead>
<tr>
<th>Feature</th>
<th>Error Rates (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>67.0</td>
</tr>
<tr>
<td>Scattering Transform (2nd order)</td>
<td>24.0</td>
</tr>
<tr>
<td>Scattering Transform (3rd order)</td>
<td>22.5</td>
</tr>
<tr>
<td>Scattering Transform (4th order)</td>
<td>21.5</td>
</tr>
<tr>
<td>Log Spectrogram</td>
<td>35.5</td>
</tr>
<tr>
<td>Invariant (Warp)</td>
<td>22.0</td>
</tr>
<tr>
<td>Invariant (Warp + Translation)</td>
<td>16.5</td>
</tr>
<tr>
<td>Invariant (Warp + Translation + Pitch)</td>
<td>18.0</td>
</tr>
</tbody>
</table>
What are the class-preserving transformations for music classification?

What are the (invariant) characteristics of music genres?

- Any transformation that preserves such invariants could be “irrelevant”.

Learning transformations from the data

- Learning needs to see the transformed templates $g \circ t_k$.
- But there is no need to know explicitly what the transformations $G = \{g\}$ are.

Temporal continuity

- Nearby audio segments within the same clip (genre preserved) could be treated as the same identity undergone some unknown smooth transformations.
Summary (Contributions)

- Basic Theory
  - Theoretical framework for invariant representations.

- Neural Realization
  - Implementation of modules and network cascades / hierarchies.

- Evaluation
  - Music genre classification (GTZAN): improved by over scattering (deep) and MFCC (shallow)