Self-paced Curriculum Learning

Lu Jiang ¹, Deyu Meng², Qian Zhao², Shiguang Shan³, Alexander G. Hauptmann ¹

¹School of Computer Science, Carnegie Mellon University
²School of Mathematics and Statistics, Xi'an Jiaotong University
³Institute of Computing Technology, Chinese Academy of Sciences
People

• CMU Informedia Team

Deyu Meng  Qian Zhao  Shiguang Shan  Alexander Hauptmann
Outline

- Motivation
- Background Knowledge
- Self-paced Curriculum Learning
- Experiments
- Conclusions
Motivations

- Noisy
- Highly unbalanced data
- Training non-convex models?

Many algorithms have been proposed. One solution is biologically inspired: what we will do if we are asked to learn something from the big data.
Curriculum Learning and Self-paced Learning

• Curriculum Learning (Bengio et al. 2009) or self-paced learning (Kumar et al. 2010) represents a recently proposed learning paradigm that is inspired by the learning process of humans and animals.

• The samples are not learned randomly but organized in a meaningful order which illustrates from easy to gradually more complex examples.

• Curriculum: a sequence of gradually learned samples.

Curriculum Learning and Self-paced Learning

• Samples are organized in a meaning order (curriculum).
• Learning is conducted iteratively.
• Models are becoming increasingly complex.
Curriculum Learning and Self-paced Learning

- Samples are organized in a meaning order (curriculum).
- Learning is conducted iteratively.
- Models are becoming increasingly complex.

*The above of real examples in the TRECVID SIN dataset (http://trecvid.nist.gov/).
Easy and Complex samples in Google Image Search

Samples of “Dog” to learn earlier.

Samples of “Dog” to learn later.

In Big data, we see a lot more examples like this.
Outline

- Motivation
- Background Knowledge
- Self-paced Curriculum Learning
- Experiments
- Conclusions
Curriculum Learning

- **Curriculum Learning (CL):** assign learning priorities to training samples, according to prior knowledge or heuristics about specific problems.
- Teaching a robot: leverage human curriculum.

---


Curriculum Learning

- **Curriculum Learning (CL)**: assign learning priorities to training samples, according to prior knowledge or heuristics about specific problems.

- Teaching a robot: leverage human curriculum.

- Parsing in Natural Language Processing (NLP):
  - From shorter sentences to longer sentence.

Spitkovsky, V. I.; Alshawi, H.; and Jurafsky, D. 2009. Baby steps: How less is more in unsupervised dependency parsing. In NIPS
Self-paced Learning

- **Self-paced Learning (SPL):** the curriculum is determined by the learned models.
- Solving a joint optimization problem of the learning objective with the curriculum (a sequence of gradually added samples).
  - **From** → smaller loss to the already learned model.
  - **to** → larger loss to the already learned model.


Curriculum Learning versus Self-paced Learning

Curriculum Learning (CL)

• Pros
  – Flexible to incorporate prior knowledge/heuristics.

• Cons
  – Curriculum is determined beforehand which may not be consistent with dynamically learned models.

Self-paced Learning (SPL)

• Pros
  – Learn consistent models.
  – Concise optimization problem.

• Cons
  – Cannot use prior knowledge.
  – Random starting values (can be sensitive to the performance).

Difficult to judge which one is better in practice.
Curriculum Learning versus Self-paced Learning

Curriculum Learning (CL) - instructor-driven

Self-paced Learning (SPL) - student-driven

Difficult to judge which one is better in practice.
Self-paced Curriculum Learning

Curriculum Learning (CL)  +  Self-paced Learning (SPL)

instructor-driven  +  student-driven

Self-paced Curriculum Learning (SPCL)

instructor-student-collaborative

Unified in a single framework: SPCL
Outline

- Motivation
- Background Knowledge
- Self-paced Curriculum Learning
- Experiments
- Conclusions
Self-paced Curriculum Learning

• Formulated as an optimization problem (based on SPL). Consider a binary classification problem:

$$\arg\min_{\mathbf{w}, \mathbf{v} \in [0,1]^n} \sum_{i=1}^{n} v_i L(y_i, g(x_i, \mathbf{w})) + f(\mathbf{v}, \lambda)$$

subject to $\mathbf{v} \in \Psi$

$\mathbf{w} \Rightarrow$ parameters in the off-the-shell model
$L(y_i, g(x_i, w)) \Rightarrow$ loss for the $i^{th}$ sample
$\mathbf{v} = [v_1, \ldots, v_n] \Rightarrow$ weight vector for all samples
$f(\mathbf{v}, \lambda) \Rightarrow$ regularizer determines the learning scheme
$\lambda \Rightarrow$ model age
$\Psi \Rightarrow$ feasible region that encodes the prior knowledge

Off-the-shell model (SVM, deep neural networks etc.)
Weight vectors in self-paced learning
Prior knowledge in curriculum learning
Self-paced Curriculum Learning

How to solve (alternative search):

- Fixing $v$ and optimize model parameters $w$.
- Fixing $w$ and optimize weight variables $v$.
- Increase the model age to train a more complex model.

Self-paced Curriculum Learning

How to solve (alternative search):

- Fixing \( v \) and optimize model parameters \( w \).
- Fixing \( w \) and optimize weight variables \( v \).
- Increase the model age to train a more complex model.

Recalculating the loss and select more examples.
Self-paced Curriculum Learning

How to solve (alternative search):

- Fixing $v$ and optimize model parameters $w$.
- Fixing $w$ and optimize weight variables $v$.
- Increase the model age $\lambda$ to train a more complex model.

Increase the model age to include more examples

smaller loss  bigger loss
Self-paced Curriculum Learning

• Formulated as an optimization problem (based on SPL):  
\[
\arg \min_{\mathbf{w}, \mathbf{v}} \sum_{i=1}^{n} v_i L(y_i, g(x_i, \mathbf{w})) + f(\mathbf{v}, \lambda)
\]

subject to \( \mathbf{v} \in \Psi \)
Self-paced Curriculum Learning

• Formulated as an optimization problem (based on SPL):

\[
\text{arg} \min_{\mathbf{w}, \mathbf{v}} \sum_{i=1}^{n} v_i L(y_i, g(x_i, \mathbf{w})) + \ell(\mathbf{v}, \lambda)
\]

(subject to \( \mathbf{v} \in \Psi \))

• **Novelty**: when optimizing \( \mathbf{v} \) with the fixed \( \mathbf{w} \):
  – Encode heuristics/prior knowledge in the feasible region \( \Psi \):
    • E.g. \( v_1 \) learned before \( v_3 \), \( v_2 \) before \( v_3 \) \( v_1 \geq v_2 \geq v_3 \)
  – Represent the regularizer to present different learning scheme.
    Apply different regularizer to different problems:
    • Start from easy to complex examples?
    • From easy and diverse to complex examples?
    • Even from complex to easy (for very smart learner/student for example)?
Self-paced Curriculum Learning

- Formulated as an optimization problem (based on SPL):
  \[
  \arg \min_{w,v} \sum_{i=1}^{n} v_i L(y_i, g(x_i, w)) + f(v, \lambda)
  \]
  subject to \( v \in \Psi \)

- When optimizing \( v \) with the fixed \( w \):
  - Encode heuristics/prior knowledge in the feasible region:
    - E.g. \( v_1 \) learned before \( v_3 \), \( v_2 \) before \( v_3 \)
  - Represent the regularizer to present different learning scheme. Apply different regularizers to different problems:
    - Start from easy to complex examples?
    - From diverse to complex examples?
    - Even from complex to easy (for very smart learner/student for example)?
Outline

- Motivation
- Background Knowledge
- Self-paced Curriculum Learning
- Experiments
- Conclusions
Experiments

• Matrix factorization:

\[
\begin{bmatrix}
    n \\
    \hline
    d \\
    \hline
    \end{bmatrix}
    \times
    \begin{bmatrix}
    h \\
    \hline
    d \\
    \hline
    \end{bmatrix}
    = \begin{bmatrix}
    n \\
    \hline
    U \\
    \hline
    V^T \\
    \end{bmatrix}
\]

• Content-based video retrieval:
Experiments

Table 2: Performance comparison of SPCL and baseline methods for matrix factorization.

<table>
<thead>
<tr>
<th></th>
<th>$L_2$-norm MF</th>
<th>$L_1$-norm MF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baseline</td>
<td>SPL</td>
</tr>
<tr>
<td>RMSE</td>
<td>9.3908</td>
<td>0.2585</td>
</tr>
<tr>
<td>MAE</td>
<td>6.8597</td>
<td>0.0947</td>
</tr>
</tbody>
</table>

RMSE (Root Mean Square Error)
Lower -> better

Table 3: Performance comparison of SPCL and baseline methods for zero-example event reranking.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CL</th>
<th>SPL</th>
<th>SPCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>MED13Test</td>
<td>10.1</td>
<td>10.8</td>
<td><strong>12.9</strong></td>
</tr>
<tr>
<td>MED14Test</td>
<td>7.3</td>
<td>8.6</td>
<td><strong>9.2</strong></td>
</tr>
</tbody>
</table>

MAP (Mean Average Precision)
Higher -> better

Incorporating prior knowledge into statistical learning tends to be instrumental.
Outline

- Motivation
- Background Knowledge
- Self-paced Curriculum Learning
- Experiments
- Conclusions
Conclusions

Take home messages:

– Proposed a novel learning framework that unifies the existing curriculum learning and self-paced learning paradigms.

– SPCL is general and has pluggable components:

  • Off-the-shell model ➔ Student
  • Regularizers ➔ Learning schemes
  • Feasible region ➔ Prior knowledge

– Observed benefits for the non-convex problems and the problems with noisy and unbalanced data.
THANK YOU.

Q&A?
Self-paced Curriculum Learning

- Self-paced curriculum (SPCL) Learning unified curriculum learning (CL) and self-paced learning (SPL) into a universal framework.

Table 1: Comparison of different learning approaches.

<table>
<thead>
<tr>
<th>Comparable to human learning</th>
<th>CL</th>
<th>SPL</th>
<th>Proposed SPCL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curriculum design</td>
<td>Instructor-driven</td>
<td>Student-driven</td>
<td>Instructor-student collaborative</td>
</tr>
<tr>
<td>Learning schemes</td>
<td>Prior knowledge</td>
<td>Learning objective</td>
<td>Learning objective + prior knowledge</td>
</tr>
<tr>
<td>Iterative training</td>
<td>Multiple</td>
<td>Single</td>
<td>Multiple</td>
</tr>
<tr>
<td></td>
<td>Heuristic approach</td>
<td>Gradient-based</td>
<td>Gradient-based</td>
</tr>
</tbody>
</table>