Self-paced Curriculum Learning

Lu Jiang [1], Deyu Meng[2], Qian Zhao [2], Shiguang Shan [3], Alexander G. Hauptmann [1]

[1]School of Computer Science, Carnegie Mellon University

[2]School of Mathematics and Statistics, Xi'an Jiaotong University

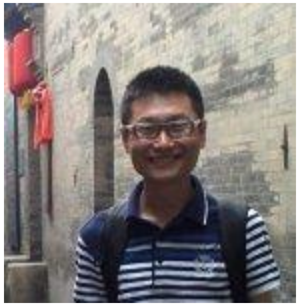
[3]Institute of Computing Technology, Chinese Academy of Sciences

AAAI 2015

People

- CMU Informedia Team

Deyu Meng



Qian Zhao

Shiguang Shan

Alexander Hauptmann

Outline

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

Carnegie Mellon University

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

Carnegie Mellon University

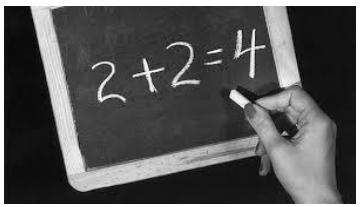
- Curriculum Learning (Bengio et al. 2009) or self-paced learning (Kumar et al 2010) represents a recently proposed learning paradigm that is inspired by the learning process of humans and animals.
- The samples are not learned randomly but **organized in a meaningful order** which illustrates from **easy** to gradually more **complex** examples.
- Curriculum: a sequence of gradually learned samples.

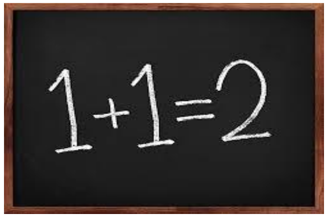
Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML, 2009.*

M. P. Kumar, B. Packer, and **D. Koller**. Self-paced learning for latent variable models. In NIPS, pages 1189–1197, 2010.

Curriculum Learning and Self-paced Learning

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





Age

$$
\begin{aligned}
\frac{1}{g-kv}\frac{dv}{dt} &= 1 \\
\int_0^T \frac{1}{g-kv}\frac{dv}{dt}\,dt &= \int_0^T dt \\
\int_{v_0}^{v(T)} \frac{1}{g-kv}\,dv &= T \\
-\frac{1}{k}\ln|g-kv|\,\Big|_{v_0}^{v(T)} &= T \\
\ln\left|\frac{g-kv(T)}{g-kv_0}\right| &= -kT \\
\frac{g-kv(T)}{g-kv_0} &= e^{-kT}
\end{aligned}
$$

Curriculum Learning and Self-paced Learning

Carnegie Mellon University

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

"bus" samples to learn earlier

Age

"bus" samples to learn later

*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.

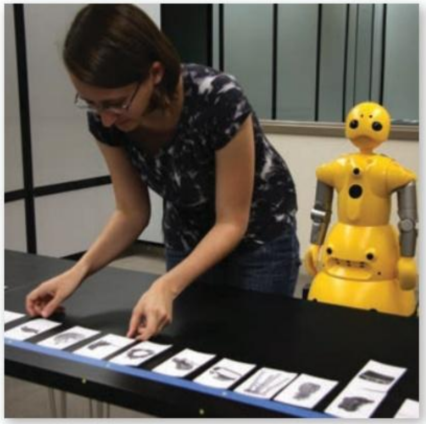
In Big data, we see a lot more examples like this.

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.



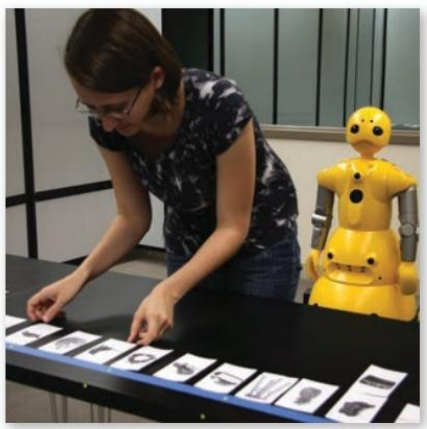
Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML, 2009.*

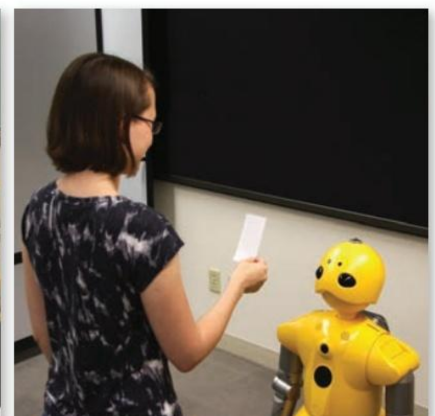
Khan, F.; Zhu, X.; and Mutlu, B. 2011. How do humans teach: On curriculum learning and teaching dimension. In NIPS.

Carnegie Mellon University

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.

M. P. Kumar, B. Packer, and **D. Koller**. Self-paced learning for latent variable models. In NIPS, pages 1189–1197, 2010.
Jiang, L.; Meng, D.; Yu, S.-I.; Lan, Z.; Shan, S.; and Hauptmann, A. G. 2014b. Self-paced learning with diversity. In NIPS.

Curriculum Learning **versus** Self-paced Learning

Carnegie Mellon University

Curriculum Learning (CL)

- Pros
  - Flexible to incorporate prior knowledge/heuristics.
- Cons
  - Curriculum is determined beforehand which may not be consistent with dynamically learned modles.

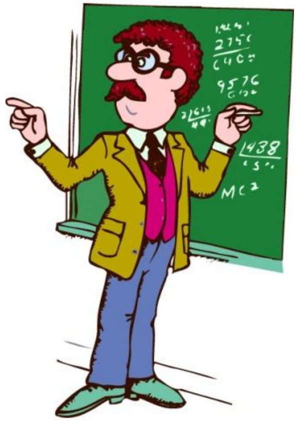
Self-paced Learning (SPL)

- 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.

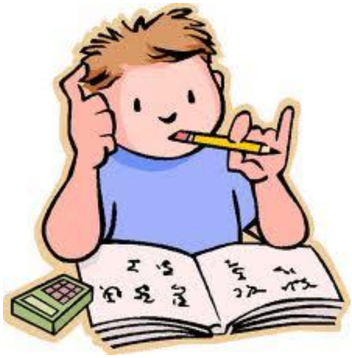
Curriculum Learning **versus** Self-paced Learning

Curriculum Learning (CL)



instructor-driven

Self-paced Learning (SPL)



student-driven

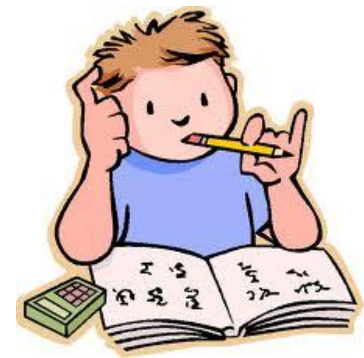
Difficult to judge which one is better in practice.

Self-paced Curriculum Learning

Curriculum Learning (CL)

instructor-driven

Self-paced Learning (SPL)



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:

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

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

$$\left.\begin{array}{r} \mathbf{w} \Rightarrow \text{parameters in the off-the-shell model} \\ L(y_i, g(\mathbf{x}_i, w)) \Rightarrow \text{loss for the } ith \text{ sample} \end{array}\right\}$$

Off-the-shell model (SVM, deep neural networks etc.)

$$\left.\begin{array}{r} \mathbf{v} = [v_1, \ldots, v_n] \Rightarrow \text{weight vector for all samples} \\ f(\mathbf{v}, \lambda) \Rightarrow \text{regularizer determines the learning scheme} \\ \lambda \Rightarrow \text{model age} \end{array}\right\}$$

Weight vectors in self-paced learning

$$\left.\Psi \Rightarrow \text{feasible region that encodes the prior knowledge}\right\}$$

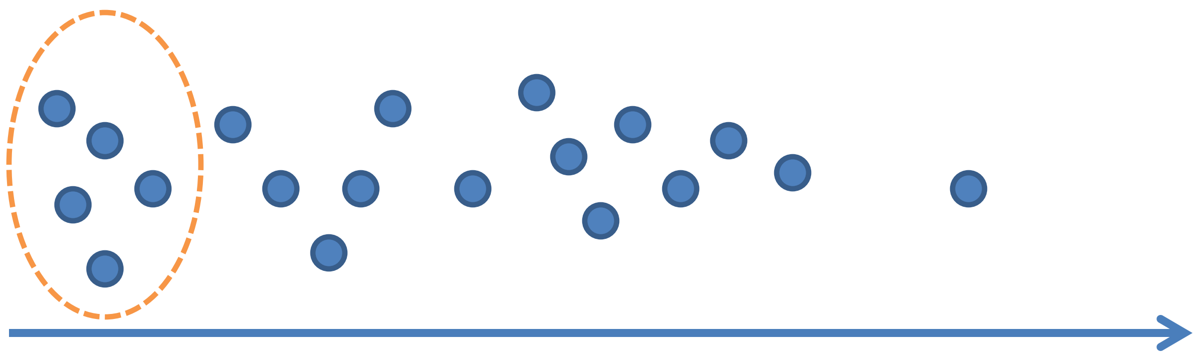
Prior knowledge in curriculum learning

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.

Training a model



smaller loss

bigger loss

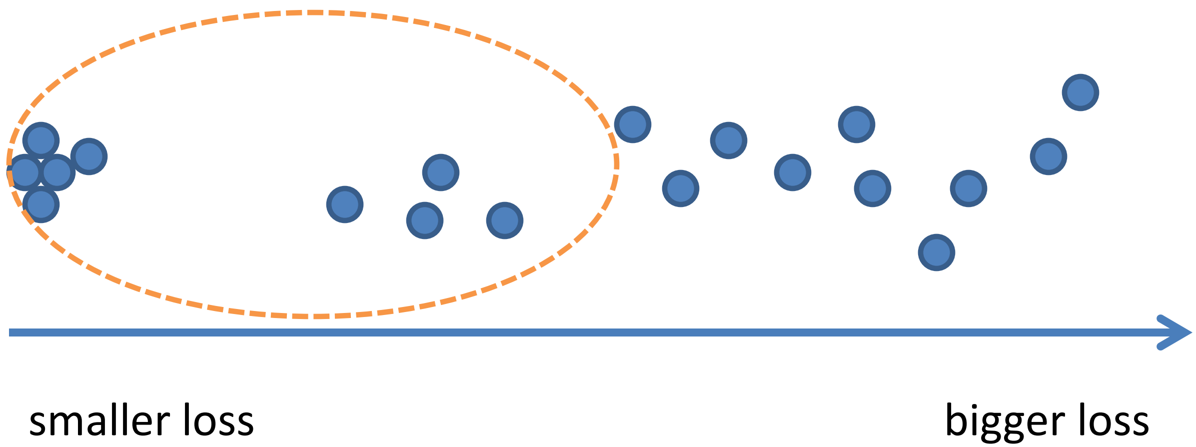
M. P. Kumar, B. Packer, and **D. Koller**. Self-paced learning for latent variable models. In NIPS, pages 1189–1197, 2010.

Self-paced Curriculum Learning

How to solve (alternative search):

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

Recalculating the loss and select more examples.

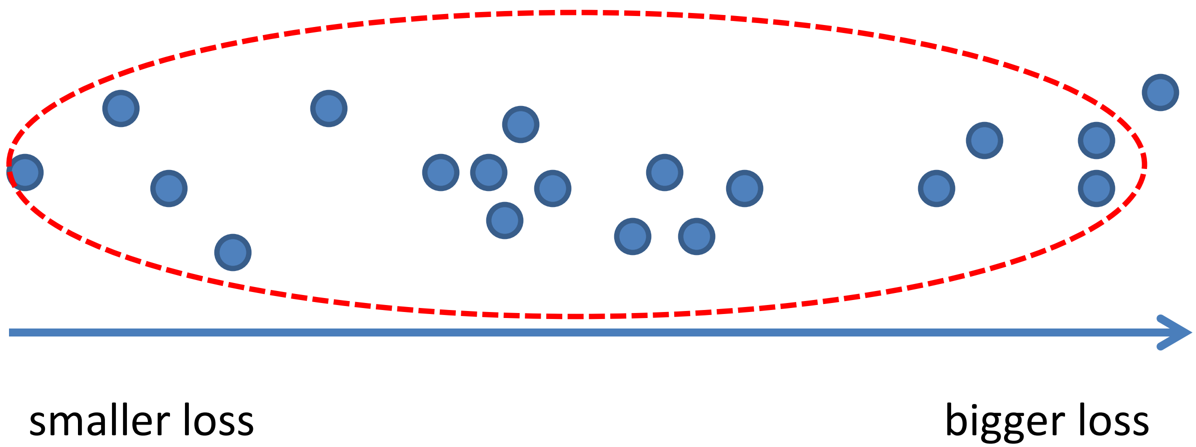


Self-paced Curriculum Learning

How to solve (alternative search):

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

Increase the model age to include more examples



Self-paced Curriculum Learning

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

$$\arg\min_{\mathbf{w},\mathbf{v}} \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i, \mathbf{w})) + f(\mathbf{v}, \lambda)$$

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

Self-paced Curriculum Learning

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

$$\arg\min_{\mathbf{w},\mathbf{v}} \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i, \mathbf{w})) + f(\mathbf{v}, \lambda)$$

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

- **Novelty**: when optimizing $\mathbf{v}$ with the fixed $\mathbf{w}$:
  - Encode heuristics/prior knowledge in the feasible region Ψ :
    - E.g. v1 learned before v3, v2 before v3 $v_1 \geq v_2 \geq v_3$
  - 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):

$$\arg\min_{\mathbf{w},\mathbf{v}} \sum_{i=1}^{n} v_i L(y_i, g(\mathbf{x}_i, \mathbf{w})) + \boxed{f(\mathbf{v}, \lambda)}$$

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

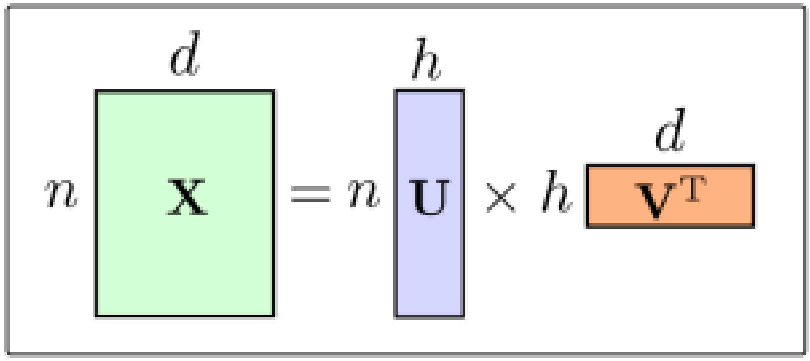
- When optimizing $\mathbf{v}$ with the fixed $\mathbf{w}$:
  - Encode heuristics/prior knowledge in the feasible region :
    - E.g. v1 learned before v3, v2 before v3
  - Represent the regularizer to present different learning scheme. Apply different regularizers to different problems:
    - Start 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:



- Content-based video retrieval:

Experiments

Carnegie Mellon University

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

	L_2-norm MF			L_1-norm MF		
	Baseline	SPL	SPCL	Baseline	SPL	SPCL
RMSE	9.3908	0.2585	**0.0654**	2.8671	0.1117	**0.0798**
MAE	6.8597	0.0947	**0.0497**	1.4729	0.0766	**0.0607**

RMSE (Root Mean Square Error)

Lower -> better

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

Dataset	CL	SPL	SPCL
MED13Test	10.1	10.8	**12.9**
MED14Test	7.3	8.6	**9.2**

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

Carnegie Mellon University

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.

	CL	SPL	Proposed SPCL
Comparable to human learning	Instructor-driven	Student-driven	Instructor-student collaborative
Curriculum design	Prior knowledge	Learning objective	Learning objective + prior knowledge
Learning schemes	Multiple	Single	Multiple
Iterative training	Heuristic approach	Gradient-based	Gradient-based