Curriculum Learning

Yoshua Bengio, U. Montreal
Jérôme Louradour, A2iA
Ronan Collobert, Jason Weston, NEC

ICML, June 16th, 2009, Montreal

Acknowledgment: Myriam Côté
Curriculum Learning

Guided learning helps training humans and animals

Start from simpler examples / easier tasks  (Piaget 1952, Skinner 1958)
The Dogma in question

It is best to learn from a training set of examples sampled from the same distribution as the test set. Really?
Question

Can machine learning algorithms benefit from a curriculum strategy?

Cognition journal:  
(Elman 1993) vs (Rohde & Plaut 1999),  
(Krueger & Dayan 2009)
Convex vs Non-Convex Criteria

- **Convex criteria**: the order of presentation of examples should not matter to the convergence point, but could influence *convergence speed*

- **Non-convex criteria**: the order and selection of examples could yield to a *better local minimum*
Deep Architectures

- Theoretical arguments: deep architectures can be exponentially more compact than shallow ones representing the same function
- Cognitive and neuroscience arguments
- Many local minima
  - **Guiding the optimization** by unsupervised pre-training yields much better local minima o/w not reachable
- Good candidate for testing curriculum ideas
Deep Training Trajectories

(Erhan et al. AISTATS 09)

Random initialization

Unsupervised guidance
Starting from Easy Examples

1. Easiest
   - Lower level abstractions

2. Most difficult examples
   - Higher level abstractions

• Most difficult examples
• Higher level abstractions
Continuation Methods

Target objective

Final solution

Track local minima

Easy to find minimum

Heavily smoothed objective = surrogate criterion
Curriculum Learning as Continuation

1. Easiest
   - Lower level abstractions

2. Initial peak on easier/simpler ones

3. Most difficult examples
   - Higher level abstractions

- Sequence of training distributions
- Initially peaking on easier/simpler ones
- Gradually give more weight to more difficult ones until reach target distribution
How to order examples?

- The right order is not known

- 3 series of experiments:
  1. Toy experiments with simple order
     - Larger margin first
     - Less noisy inputs first
  2. Simpler shapes first, more varied ones later
  3. Smaller vocabulary first
Larger Margin First: Faster Convergence

![Graph showing Easiness based on margin](image)
Cleaner First: Faster Convergence

![Graph showing easiness based on number of noisy inputs](image)

- Red line: no curriculum
- Purple line: curriculum

Y-axis: Average test error
X-axis: Input dimension
Shape Recognition

First: easier, basic shapes

Second = target: more varied geometric shapes
Shape Recognition Experiment

- 3-hidden layers deep net known to involve local minima (unsupervised pre-training finds much better solutions)
- 10,000 training / 5,000 validation / 5,000 test examples

Procedure:
1. Train for k epochs on the easier shapes
2. Switch to target training set (more variations)
Shape Recognition Results
Language Modeling Experiment

- **Objective**: compute the score of the next word given the previous ones (ranking criterion)

- **Architecture of the deep neural network** (Bengio et al. 2001, Collobert & Weston 2008)
Language Modeling Results

- Gradually increase the vocabulary size (dips)
- Train on Wikipedia with sentences containing only words in vocabulary
Conclusion

Yes, machine learning algorithms can benefit from a curriculum strategy.
Why?

- Faster convergence to a minimum
- Wasting less time with noisy or harder to predict examples
- Convergence to better local minima

Curriculum = particular continuation method

- Finds better local minima of a non-convex training criterion
- Like a regularizer, with main effect on test set
Perspectives

- How could we define better curriculum strategies?
- We should try to understand general principles that make some curricula work better than others
- Emphasizing harder examples and riding on the frontier
THANK YOU!

- Questions?
- Comments?
Training Criterion: Ranking Words

\[ C_s = \sum_{w \in D} \frac{1}{|D|} \quad C_{s,w} = \sum_{w \in D} \frac{1}{|D|} \max(0, 1 - f(s) + f(s^w)) \]

with
- \( S \) a word sequence
- \( C_s \) score of the next word given the previous one
- \( w \) a word of the vocabulary
- \( D \) the considered word vocabulary
Curriculum = Continuation Method?

- Examples $z$ from $P(z)$ are weighted by $0 \leq W_\lambda(z) \leq 1$
- Sequence of distributions $Q_\lambda(z) \propto W_\lambda(z) P(z)$ called a curriculum if:
  - the entropy of these distributions increases (larger domain)
    \[ H(Q_\lambda) < H(Q_{\lambda+\epsilon}) \quad \forall \epsilon > 0 \]
  - $W_\lambda(z)$ monotonically increasing in $\lambda$:
    \[ W_{\lambda+\epsilon}(z) \geq W_\lambda(z) \quad \forall z, \forall \epsilon > 0 \]