Neural Self-training through Spaced Repetition

Hadi Amiri
HMS, Harvard University
hadi@hms.harvard.edu
Motivation

- Labeled data
  - Training models
  - Analyzing data
  - Making decisions, etc.

- Labeled data is expensive & time-consuming to obtain

- Semi-supervised learning
  - Self-training
  - Pre-training

Wu et al., 2016; Hirschberg and Manning, 2015
Self-training

Labeled data → Classification Model Development → Classifier → Final Classification Model

 Classifier

Pseudo labeled data

Unlabeled data

Unlabeled Data Sampling
Self-training

- Labeled data
- Classification Model Development
- Pseudo labeled data
- Unlabeled data
- Classifier
- Unlabeled Data Sampling
- Final Classification Model

Conventional methods rely on prediction confidence of classifier for sampling.

Zhu, 2006; Zhu and Goldberg, 2009
Self-training

Labeled data

Classification Model Development

Classifier

Unlabeled data

Pseudo labeled data

Unlabeled Data Sampling

Final Classification Model

- Only unlabeled instances that match well with current model will be selected, and
- Model may fail to best generalize to complete sample space.

Zhu, 2006; Zhu and Goldberg, 2009
Self-training

Key contribution
Learn a data sampling policy!

Zhu, 2006; Zhu and Goldberg, 2009
Pre-training

1. Layers are first pre-trained by learning to reconstruct their inputs.

2. Encoder parameters are fine-tuned against target task.

Hinton and Salakhutdinov, 2006
Pre-training

1. Layers are first pre-trained by learning to reconstruct their inputs.

2. Encoder parameters are fine-tuned against target task.

**Limitation**
The same neural model or parts thereof must be used in both pretraining and fine-tuning steps.

**Key contribution**
Decouple pretraining and fine-tuning steps.

Hinton and Salakhutdinov, 2006
Data Sampling Policy

• Our data sampling policy
  ▫ Is not predetermined
  ▫ Explores the entire data space
  ▫ Samples based on
    • Performance of learner on target task
    • Easiness of unlabeled instances
  ▫ Decouples pretraining and fine-tuning steps
Neural Self-training

Data sampling based on Leitner Queues:
1. Place all unlabeled data in first queue.
2. Move instances among queues based on:
   - Network predictions
   - Pseudo labels.
Neural Self-training – Leitner Queue

- **Queue Movement:**
  - **Prompt to next queue**
    - Instances with similar class predictions and pseudo labels
  - **Demote to previous queue**
    - Those with opposite predictions and labels
Neural Self-training – Leitner Queue

• Errors (misleading instances):
  ▫ Incorrect pseudo labels
  ▫ Incorrect class predictions
Neural Self-training – Leitner Queue

- Greedy data sampling approach
  - Select instances of the queue that most improves learner’s performance on val. data (designated queue)

alleviates effect of misleading instances as they can’t improve model generalizability.
Model Performance

Predict user opinion against movies

<table>
<thead>
<tr>
<th>Method</th>
<th>IMDb Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Classifier</td>
<td>73.02</td>
</tr>
<tr>
<td>Standard ST</td>
<td>74.43</td>
</tr>
<tr>
<td>Pretraining ST</td>
<td>76.36</td>
</tr>
<tr>
<td>Adversarial ST</td>
<td>76.09</td>
</tr>
<tr>
<td>Knowledge Transfer ST</td>
<td>77.11</td>
</tr>
<tr>
<td>Leitner ST</td>
<td>78.27</td>
</tr>
</tbody>
</table>

Predict user intention about leaving a brand

<table>
<thead>
<tr>
<th>Method</th>
<th>Churn Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base Classifier</td>
<td>65.77</td>
</tr>
<tr>
<td>Standard ST</td>
<td>65.77</td>
</tr>
<tr>
<td>Pretraining ST</td>
<td>67.27</td>
</tr>
<tr>
<td>Adversarial ST</td>
<td>67.7</td>
</tr>
<tr>
<td>Knowledge Transfer ST</td>
<td>67.06</td>
</tr>
<tr>
<td>Leitner ST</td>
<td>69.9</td>
</tr>
</tbody>
</table>
Model Performance

- Perhaps due to
  - Failure of standard ST to explore the data space,
  - Classification mistakes reinforcing each other, or
  - Highly imbalanced classes.
Model Introspection

• **Q1**: What’s the Issue with Highest Queues?

• **Q2**: Does Sample Diversity Matter?

• **Q3**: Do We Need Better Sampling Policies?
Q1: Highest Queues

• How well queue instances match training data?
  ▫ Cosine similarity between representations of training instances queue instances.

Instances in highest queues match well with current training data and are less likely to be informative.
Q2: Queue Diversity

- We compute the extent of diversity that each given queue introduces when added to training data.

Leitner self-training enables sampling diverse sets of instances that contributes to training an improved model.
Q3: Sampling Policy

• Instance movement patterns among queues.

Movements in middle queues are spread out over a larger range of queues than lower/higher queues.

Finer-grained queue-level data sampling may create better data samples.

Figure 4: Deviation in instance movements for each queue (in terms of average standard deviation over all training episodes). At every episode, we keep track of instance movements among queues and measure movement variation among instances that ultimately home in on the same queue.
Q3: Sampling Policy

• Instance movement patterns among queues.

\[ q_0 \rightarrow q_0 \rightarrow q_0 \rightarrow q_0 \rightarrow q_0 : \text{always in } q_0 \]
\[ q_0 \rightarrow q_1 \rightarrow q_0 \rightarrow q_0 \rightarrow q_0 : \text{mainly in } q_0 \]
\[ q_0 \rightarrow q_1 \rightarrow q_0 \rightarrow q_1 \rightarrow q_0 : \text{partially in } q_0 \]
\[ q_0 \rightarrow q_1 \rightarrow q_2 \rightarrow q_1 \rightarrow q_0 : \text{partially in } q_0 \& q_1. \]

Movements in middle queues are spread out over a larger range of queues than lower/higher queues.

Finer-grained queue-level data sampling may create better data samples.
Conclusion

• Contributions
  ▫ Formulation of spaced repetition for self-training methods
    • Novel data sampling approach based on Leitner Queues

• Future work
  ▫ Finer-grained data sampling at queue level
  ▫ Extending our approach to other iterative training settings
    • E.g. boosting
  ▫ Using our model in areas with limited labeled data.
Thank you

hadi@hms.harvard.edu
Instances in the highest queues, although easy to learn for classifier, are not informative and do not contribute to training an improved model.
Q2: Queue Diversity

- Diverse queue
  - At every episode, rank instances of each queue based on their prediction confidence and create a diverse queue by combining top r% instances of each queue.

<table>
<thead>
<tr>
<th></th>
<th>Leitner ST</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 3</td>
</tr>
<tr>
<td>IMDb</td>
<td>76.83</td>
</tr>
<tr>
<td>Churn</td>
<td>65.74</td>
</tr>
</tbody>
</table>

Table 5: Macro-F1 performance of diverse queues across datasets. Compare these results with those obtained by designated queues in Table 4.
Neural Self-training – Leitner Queue

Scheduler: Instances at $q_i$ are reviewed at every $2^{i-1}$ iteration.

$q_0$ epochs = \{1, 2, 3, 4, 5, \ldots\}
$q_1$ epochs = \{2, 4, 6, 8, 10, \ldots\}
$q_2$ epochs = \{4, 8, 12, 16, 20, \ldots\}
\ldots