**Motivation and Contributions**
- Tuning hyperparameters can be critical for generalization, but difficult when hyperparameters are high-dimensional.
- If we can tune as many hyperparameters as weights, we have a new paradigm of learned, flexible regularization.
- Learned data augmentation
- Data distillation
- Tuning millions of regularization hyperparameters
- We scale gradient-based hyperparameter optimization to high dimensions by combining implicit differentiation with efficient & stable inverse-Hessian approximations.

**Hyperparameter Optimization is Nesting Optimization**

\[ \lambda^* := \arg \min \mathcal{L}(\lambda) \]

where \( \mathcal{L}(\lambda) := \mathcal{L}(\lambda, w(\lambda)) \) and \( w(\lambda) := \arg \min_{w} \mathcal{L}(\lambda, w) \)

**Calculating Hypergradients Efficiently**

1. while not converged do
2. for \( k = 1 \ldots N \) do
3. \( w' := \frac{\partial \mathcal{L}}{\partial \mathcal{W}} w \)
4. \( \lambda' := \text{hypergradient}(\mathcal{L}_v, \mathcal{L}_T, \lambda', w') \)
5. return \( \lambda', w' \)

The entire hypergradient computation can be performed efficiently using vector-Jacobian products, given an inverse-Hessian-vector approximation.

**Comparing Inverse-Hessian Approximations**

We train a U-net for held-out data with weights as hyperparameters, which takes data & noise, and outputs augmented data.

**Tuning Regularization Parameters**

<table>
<thead>
<tr>
<th>Method</th>
<th>Validation</th>
<th>Test</th>
<th>Time(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grid Search</td>
<td>97.32</td>
<td>94.58</td>
<td>100k</td>
</tr>
<tr>
<td>Random Search</td>
<td>84.81</td>
<td>81.46</td>
<td>100k</td>
</tr>
<tr>
<td>Bayesian Opt.</td>
<td>72.13</td>
<td>69.29</td>
<td>100k</td>
</tr>
<tr>
<td>STN</td>
<td>70.30</td>
<td>67.68</td>
<td>25k</td>
</tr>
<tr>
<td>Ours</td>
<td>69.22</td>
<td>66.40</td>
<td>18.5k</td>
</tr>
<tr>
<td>Ours, Many</td>
<td>68.18</td>
<td>66.14</td>
<td>18.5k</td>
</tr>
</tbody>
</table>

For LSTMs on PTB, we tune the same 7 (and millions of) hyperparameters faster, with comparable memory, and to a lower perplexity.