Uniform Sampling Over Episodic Difficulty

S. M. R. Arnold*, G. S. Dhillon*, A. Ravichandran, S. Soatto

October 18, 2021
Episodic Training in Few-Shot Learning

Few-Shot Learning

- Learn a model to solve new tasks with limited labelled data.

Episodic Training in 3 Steps

1. Sample an episode from distribution.
2. Solve the episode with limited data.
3. Update the model to improve generalization of solution.

Plenty of Recent Methods

- Gradient-based: MAML, ANIL, MetaCurvature, KFO, MT-Nets, ...
- Metric-based: ProtoNets, MetaOptNet, FEAT, DeepEMD, ...
Driving Question

- How should we sample episodes for best transfer accuracy?

Contributions

- Analysis of episode difficulty and its distribution.
- Simple method to approximate any episodic sampling distribution.
- Main result: uniform sampling over episode difficulty improves episodic training.
Few-Shot Classification Episodes

Baseline Episode Sampling

1. Sample $n$ classes from base dataset.
2. Sample $k$ samples / class for support set $\tau_S$.
3. Sample $k'$ samples / class for query set $\tau_Q$.

Solving an Episode with ProtoNets

- Compute class centroids on support set:
  \[
  \phi^c_\theta = \frac{1}{k} \sum_{(x,y) \in \tau_S, y = c} \phi_\theta(x)
  \]

- Classify query set with nearest centroid:
  \[
  l_\theta(y \mid x, \tau_S) = \frac{\exp (-d(\phi_\theta(x), \phi^y_\theta))}{\sum_{y' \in C_r} \exp (-d(\phi_\theta(x), \phi^{y'}_\theta))}
  \]

Distribution of Episode Difficulty

Definition

- The difficulty of an episode $\tau$ is given by:
  \[ \Omega_{l_\theta}(\tau) = -\log l_\theta(\tau) \]
  for model likelihood $l$, support set $S$, and query set $Q$.

Why This Definition?

- Easy to compute, readily available during training.
- Model-agnostic — applies to all methods.
- No discretization artifacts (unlike, say, accuracy).

Empirical Analysis

For many algorithms and models:

**Episode difficulty is approximately normally distributed.**
Implicit Dependence on Modelling Choices

Is episode difficulty the same across different...

... model architectures? ✔

- Average Spearman correlation: 0.59.

... model parameters?

... training algorithms?
Implicit Dependence on Modelling Choices

Is episode difficulty the same across different...

... model architectures? ✔

- Average Spearman correlation: 0.59.

... model parameters? ✔

- Compare across iterations in training run.
- Hard episodes remain hard; easy remain easy.

... training algorithms?
Implicit Dependence on Modelling Choices

Is episode difficulty the same across different...

... model architectures? ✓

- Average Spearman correlation: 0.59.

... model parameters? ✓

- Compare across iterations in training run.
- Hard episodes remain hard; easy remain easy.

... training algorithms? ✓

- Compare across MAML, ANIL, ProtoNet (Euclidean & Cosine).
- Average Spearman correlation: 0.65.
How should we sample episodes?

5 Candidate Distributions

- **Baseline** — Normal distribution over difficulty.
- **Easy** — Sample uniformly over the easier 50% of episodes.
- **Hard** — Sample uniformly over the harder 50% of episodes.
- **Curriculum** — Sample easier episode when training starts, harder episodes towards end.
- **Uniform** — Sample uniformly over the difficulty range.
A Simple Method to Study Episode Sampling

Importance Sampling for Episodic Training

- **Reweight episodes** to approximate target distribution $p(\tau)$:

  \[
  \mathbb{E}_{\tau \sim q(\cdot)} \left[ w(\tau) \log l_\theta(\tau) \right] \quad \text{where} \quad w(\tau) = \frac{p(\tau)}{q(\tau)}
  \]

  and $q(\tau)$ is the distribution induce by Baseline sampling.

Adjusted Mini-Batching with Expected Sample Size

- Get the right number of sample from the target distribution:

  \[
  \mathbb{E}_{\tau \sim p(\cdot)} \left[ \log l_\theta(\tau) \right] \approx \frac{1}{\text{ESS}(B)} \sum_{\tau \in B} w(\tau) \log l_\theta(\tau)
  \]

  where \( \text{ESS}(B) = \frac{(\sum_{\tau \in B} w(\tau))^2}{\sum_{\tau \in B} w(\tau)^2} \)
Sampling Matters for Episodic Training

Experimental Setup

Compare 5 candidate distributions on different:

- **Architectures**: CNN4, ResNet12
- **Algorithms**: MAML, ANIL, ProtoNet - Euclidean & Cosine
- **Datasets**: mini-ImageNet, tiered-ImageNet
- **Setting**: 5-ways 1-shot, 5-ways 5-shots

Total: 24 different few-shot scenarios.

Results

Uniform sampling dominates, Baseline second best.
Sampling Improves Cross-Domain Transfer

Experimental Setup

Compare Baseline v.s. Uniform when:

- training on mini-ImageNet or tiered-ImageNet
- testing on:
  - CUB-200,
  -Describable Textures,
  - FGVC-Aircraft, and
  - VGG Flowers.

Total: 64 Cross-Domain Scenarios.

Results

Uniform sampling dominates.

Thanks Reviewer KvDG!
... And Improves Upon SOTA

<table>
<thead>
<tr>
<th></th>
<th>Mini-ImageNet</th>
<th>Tiered-ImageNet</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-shot (%)</td>
<td>5-shot (%)</td>
</tr>
<tr>
<td>FEAT</td>
<td>66.02±0.20</td>
<td>81.17±0.14</td>
</tr>
<tr>
<td>+ UNIFORM (Online)</td>
<td><strong>66.27±0.20</strong></td>
<td><strong>81.54±0.14</strong></td>
</tr>
</tbody>
</table>

Experimental Setup

- Compare Baseline v.s. Uniform when training with FEAT.

Results

Uniform sampling also improves SOTA algorithms.

Takeaways

1. **Sampling matters** in episodic training.
2. **Episodic difficulty** is (mostly) agnostic to architecture, algorithm, and parameter choice.
3. **Uniform sampling outperforms** other sampling schemes.

Learn More

- PDF, poster, slides: sebarnold.net/projects/eis
- Code: bit.ly/3p7cYz5 or learn2learn.net
- Contact: smr.arnold@gmail.com