Slice-based Learning: A Programming Model for Residual Learning in Critical Data Slices

Vincent S. Chen  
Department of Computer Science  
Stanford University  
vincentsc@cs.stanford.edu

Sen Wu  
Department of Computer Science  
Stanford University  
sewu@stanford.edu

Zhenzhen Weng  
Department of Computer Science  
Stanford University  
zzweng@stanford.edu

Alexander Ratner  
Department of Computer Science  
Stanford University  
ajratner@stanford.edu

Christopher Ré  
Department of Computer Science  
Stanford University  
chrismre@cs.stanford.edu

Abstract

In real-world machine learning applications, data subsets correspond to especially critical outcomes: vulnerable cyclist detections are safety-critical in an autonomous driving task, and “question” sentences might be important to a dialogue agent’s language understanding for product purposes. While machine learning models can achieve quality performance on coarse-grained metrics like F1-score and overall accuracy, they may underperform on these critical subsets—we define these as slices, the key abstraction in our approach. To address slice-level performance, practitioners often train separate “expert” models on slice subsets or use multi-task hard parameter sharing. We propose Slice-based Learning, a new programming model in which the slicing function (SF), a programmer abstraction, is used to specify additional model capacity for each slice. Any model can leverage SFs to learn slice-specific representations, which are combined with an attention mechanism to make slice-aware predictions. We show that our approach, which maintains a parameter-efficient representation, improves over baselines by up to 19.0 F1 on slices and 4.6 F1 overall on datasets spanning language understanding (e.g. SuperGLUE), computer vision tasks, and production-scale industrial systems.

1 Introduction

In real-world applications, some model outcomes are more important than others: for example, a data subset might correspond to safety-critical but rare scenarios in an autonomous driving setting (e.g. detecting cyclists or trolley cars [17]) or critical but lower-frequency healthcare demographics (e.g. handling younger patients with certain cancers). Traditional machine learning systems optimize for overall quality, which may be too coarse-grained; models that achieve high overall performance might produce unacceptable failure rates on slices of the data. In many production settings, the key challenge is to maintain overall model quality while improving slice-specific metrics.

Preprint. Under review.
Figure 1: Slice-based Learning via synthetically generated data: (a) The data distribution contains critical slices ($S_1$, $S_2$) that represent a small proportion of the dataset. (b) A vanilla neural network correctly learns the general, linear decision boundary but fails to learn the perturbed decision boundary of slices. (c) A user uses Slicing Functions (SFs) ($\lambda_1$, $\lambda_2$) to heuristically target critical slices. (d) The model spends additional capacity to learn slice-specific indicators + representations. (e) Upon reweighting slice-specific representations, the slice-aware model learns to classify the fine-grained slices with higher F1 score.

To formalize this challenge, we introduce the notion of slices: application-critical data subsets, specified programmatically by model developers, for which we’d like to improve model performance. This leads to three technical challenges:

- **Coping with Noise** Defining slices precisely can be challenging. While engineers often have a clear intuition of a slice, typically as a result of an error analysis, translating that intuition into a machine-understandable description can be a challenging problem, e.g., “the slice of data that contains a yellow light at dusk.” As a result, any method must be able to cope with imperfect, overlapping definitions of data slices—for example, slices representing areas of lower quality training data, as specified by weak supervision.

- **Stable Improvement of the Model** Given a description of a set of slices, we want to improve the prediction quality on each of the slices without hurting overall model performance. Often, these goals are in tension: in many baseline approaches, steps to improve the slice-specific model performance would degrade the overall model performance, and vice-versa.

- **Scalability** There may be many slices. Indeed, in industrial deployments of slicing-based approaches, hundreds of slices are commonly introduced by engineers [28]. This incompatibility must be judicious with adding additional parameters as the number of slices grow.

To improve fine-grained, i.e. slice-specific, performance, an intuitive solution is to create a separate model for each slice. To produce a single prediction at test time, one often trains a mixture of experts model (MoE) [16]. However, with the growing size of ML models, MoE is often untenable due to runtime performance, as it could require training and deploying hundreds of large models—one for each slice. Another strategy draws from multi-task learning (MTL), in which slice-specific task heads are learned with hard-parameter sharing [6]. This approach is computationally efficient but may not effectively share training data across slices, leading to suboptimal performance. Moreover, in MTL, tasks are distinct, while in slicing, a single base task is refined by slices.

We propose a novel programming model, called Slice-based Learning, in which practitioners provide slicing functions (SFs), a programming abstraction for heuristically targeting data subsets of interest. SFs coarsely map input data to slice indicators that specify data subsets for which the model should allocate additional model capacity, and the ML model being trained learns slice-specific representations via a self-attention mechanism. To improve slice-level performance, we introduce slice-residual-attention modules (SRAMs) that explicitly model residuals between slice-level and the overall task predictions. SRAMs are agnostic to the architecture of any neural network model that they are added to—which we refer to as the backbone model—and we demonstrate our approach on state-of-the-art text and image models. Our model initializes slice-specific representations that are responsible for learning slice-membership indicators and class predictors for examples in a particular slice. Then, slice indicators and prediction confidences are used in an attention-mechanism to reweight and combine each slice-specific representation into a slice-aware featurization of the
data—learning to reweight features based on residuals between slice and base representations—to make a final prediction.

Our work fits into an emerging class of programming models that sit on top of deep learning systems [26, 17]. We are the first to introduce and formalize Slice-based Learning, a key programming abstraction for improving ML models in real-world applications subject to slice-specific performance objectives. Using an independent error analysis for the recent GLUE natural language understanding benchmark tasks [34], by simply encoding the identified error categories as slices in our framework, we show that we can improve the quality of state-of-the-art models by up to 4.6 F1 points, and we observe slice-specific improvements of up to 19.0 points. We also evaluate our system on autonomous vehicle data and show improvements up to 15.6 F1 points on context-dependent slices (i.e. presence of bus, traffic light, etc.) and 2.3 F1 points overall. Anecdotally, when deployed in production systems, Slice-based Learning provides a practical programming model with improvements of up to 40 F1 points in critical test-time slices. On the SuperGlue [33] benchmark, this procedure accounts for a 2.7 improvement in overall score using the same architecture as a state-of-the-art modeling result. In addition to the proposal of SRAMs, we perform an in-depth analysis to explain the mechanisms by which SRAMs improve quality. We validate the efficacy of quality and noise estimation in SRAMs and compare to weak supervision frameworks [26] that estimate the quality of supervision sources to improve overall model accuracy. We show that by using SRAMs, we are able to produce accurate quality estimates, which leads to higher downstream performance on such tasks by an average of 1.1 overall F1 points.

2 Related Work

Our work draws inspiration from three main areas: mixture of experts, multi-task learning, and weak supervision. Jacobs [16] proposed a technique called mixture of experts that divides the data space into different homogeneous regions, learns the regions of data separately, and then combines results with a single gating network [32]. This work is a generalization of popular ensemble methods, which have been shown to improve predictive power by reducing overfitting, avoiding local optima, and combining representations to achieve optimal hypotheses [31]. We were motivated in part by reducing the runtime cost and parameter count for such models.

MTL models provide the flexibility of modular learning—specific task heads, layers, and representations can be changed in an application-specific, ad hoc manner. Furthermore, MTL models benefit from the computational efficiency and regularization afforded by hard parameter sharing [6]. There are often also performance gains seen from adding auxiliary tasks to improve representation learning objectives [7, 29]. While our approach draws high-level inspiration from MTL, we highlight key differences: whereas tasks are disjoint in MTL, slice tasks are formulated micro-tasks that are direct extensions of a base task—they are designed specifically to learn deviations from the base-task representation. In particular, sharing information, as seen in MTL cross-stitch networks [24], requires \( \Omega(n^2) \) weights across \( n \) local tasks; our formulation only requires attention over \( O(n) \) weights, as slice tasks operate on the same base task. For example, practitioners might specify yellow lights and night-time images as important slices; the model learns a series of micro-tasks—based solely on the data specification—to inform how its approach for the base task, object detection, should change in these settings. As a result, slice tasks are not fixed ahead of time by an MTL specification; instead, these micro-task boundaries are learned dynamically from corresponding data subsets. This style of information sharing is adjacent to cross-task knowledge in recent multi-task learning (MTL) models [37, 30], and we were inspired by these methods.

Weak supervision has been viewed as a new way to incorporate data of varying accuracy including domain experts, crowd sourcing, data augmentations, and external knowledge bases [27, 1, 25, 23, 4, 10, 12, 5, 19]. We take inspiration from labeling functions [27] in weak supervision as a programming paradigm, which has seen success in industrial deployments [1]. In weak supervision, a key challenge is to assess the accuracy of a training data point, which is a function of the sources that supervise it. In contrast, this work models this accuracy in a fine-grained manner, based on a learned representation—this leads to higher overall quality.

Both weak supervision and multitask learning can be viewed as orthogonal to slicing: we have observed them used alongside Slice-based Learning in academic projects and industrial deployments [28].
3 Slice-based Programming

We propose Slice-based Learning as a programming model for training neural networks where users specify important data subsets to improve model performance. We describe the core technical challenges that lead to our notion of slice-residual-attention modules (SRAMs).

3.1 Problem statement

To formalize the key challenges of slice-based learning, we introduce some basic terminology. In our base task, we use a standard supervised input, \( x \in \mathcal{X} \), where our goal is to predict \( y \in \mathcal{Y} \) according to a standard loss function. In addition, the user provides a set of \( k \) functions called slicing functions \( \{\lambda_1, \ldots, \lambda_k\} \) in which \( \lambda_i : \mathcal{X} \rightarrow \{0, 1\} \). These SFs are not assumed to be perfectly accurate; for example, SFs may be based on noisy or weak supervision sources in functional form \cite{27}. Our approach makes a key assumption: \textit{SF outputs not available during inference—i.e., after model deployment, they may be too expensive to compute or rely on unavailable metadata in online settings.}

Formally, \( s_i \in \{0, 1\}^k \) denotes the unobserved, indicator random variable of whether a point is in one of \( k \) slices, and each user-specified \( \lambda_i \) is a corresponding, noisy specification. Ultimately, the model’s goal is to improve (or avoid damaging) the overall accuracy on the base task while improving the model on the specified slices.

Our problem can be phrased as follows: given an input tuple \((\mathcal{X}, \mathcal{Y}, \{\lambda_i\}_{i=1}^k)\) consisting of a dataset \((\mathcal{X}, \mathcal{Y})\), and \( k \) different user-defined SFs \( \lambda_i \), our goal is to learn a model \( \hat{f}_{\theta}(\cdot) \)—i.e., estimate model parameters \( \hat{w} \)—that predicts \( P(Y | \{s_i\}_{i=1}^k, \mathcal{X}) \) with high average slice-specific accuracy without substantially degrading overall accuracy.

SFs can come from domain-specific heuristics, distant supervision sources, or other inexpensive, off-the-shelf models, as seen in Figure\( \ref{fig:example1} \). Then, when our model is served at inference, \textit{SFs are not necessary}, and we can rely solely on the model’s \textit{learned indicators}. In Example\( \ref{example:1} \) the potentially expensive cyclist detection algorithm is not required at runtime.

\textbf{Example 1} A developer notices that their self-driving car is not detecting bicyclists at night. Upon error analysis, they diagnose that their state-of-the-art object detection model, trained on an automobile detection dataset \((\mathcal{X}, \mathcal{Y})\) of images, is indeed underperforming on safety-critical night and cyclist slices. They write two SFs: \( \lambda_1 \) to classify night vs. day, based on pixel intensity; and \( \lambda_2 \) to detect bicycles, which calls a pretrained object detector for a bicycle (with or without a rider). These functions rely on external resources only available during offline training. Given these SFs, the developer relies on Slice-based Learning to improve model performance on specified data subsets.

3.2 Model Architecture

The Slice-based Learning architecture has six components. The key intuition is that we will train a standard prediction model, which we call the \textit{base task}. We will then learn a representation for each slice that effectively explains how its predictions should differ from the representation of the base task. The architecture will then use an attention mechanism to combine these representations to make a slice-aware prediction.

With this intuition in mind, the six components (Figure\( \ref{fig:architecture} \)) are: (a) a \textbf{backbone}, (b) a set of \( k \) \textbf{slice-indicator heads} and (c) \( k \) corresponding \textbf{slice-specific representations}, (d) a \textbf{shared slice prediction head}, (e) a combined, \textbf{slice-aware representation}, and (f) a \textbf{prediction head}. Each SRAM operates over any backbone architecture and represents a path through components (b) through (e). In Figure\( \ref{fig:ablation} \) we show ablations of these architecture components by varying the components considered in the reweighting mechanism. In the following, we describe the architecture in as a binary classification task (\( c = 1 \)).

\textbf{(a) Backbone:} Our approach is agnostic to the neural network architecture, which we call the \textit{backbone}, denoted \( f_{\theta} \); this is used primarily for feature extraction (e.g., the latest transformer for textual data, CNN for image data). The backbone maps data points \( x \) to a representation \( z \in \mathbb{R}^d \).

\textbf{(b) Slice indicator heads:} For each slice, a learned indicator head will be used to reweight (e) the “expert” slice representations (c) based on the likelihood that an example is in the corresponding slice.
Each indicator head maps the backbone representation, $z$, to a logit indicating slice-membership: $\{q_i\}_{i=1,\ldots,k} \in \{0, 1\}$. Each slice indicator head is supervised by a corresponding $\lambda_i$, and an “expert” (c) slice-specific representation, which is trained only on examples belonging to the slice using a (d) shared slice prediction head. An attention mechanism reweights these representations into a combined, (e) slice-aware representation. A final (f) prediction head makes model predictions based on the slice-aware representation.

Figure 2: Model Architecture: A developer writes SFs ($\lambda_1,\ldots,\lambda_k$) over input data and specifies any (a) backbone architecture (e.g. ResNet [14], BERT [11]) as a feature extractor. Extracted features are shared parameters for $k$ slice-residual attention modules; each learns a (b) slice indicator head, which is supervised by a corresponding $\lambda_i$, and an “expert” (c) slice-specific representation, which is trained only on examples belonging to the slice using a (d) shared slice prediction head. An attention mechanism reweights these representations into a combined, (e) slice-aware representation.

(f) Prediction head Finally, we use our slice-aware representation $z'$ as the input to a final linear layer, $f(\cdot)$, which we term the prediction head, to make a prediction on the original, base task. During inference, the prediction head is responsible for making the final prediction. During training, we
Reweighting Mechanism | Performance (F1 score) | Overall | S0 | S1 | S2 | S3
--- | --- | --- | --- | --- | --- | ---
Uniform | 77.1 | 57.1 | 68.6 | 73.6 | 72.0 |
Ind. Output | 78.1 | 52.6 | 71.0 | 76.4 | 78.6 |
Pred. Confidence | 79.3 | 61.1 | 69.2 | 78.7 | 78.6 |
Attention | 82.7 | 66.7 | 77.4 | 89.1 |

Figure 3: Architecture Ablation: Using a synthetic, binary classification dataset (Figure, left) with four randomly specified data subsets as slices (Figure, middle), we specify corresponding, noisy SFs (Figure, right) and ablate specific model components by modifying the reweighting mechanism for slice-specific representations. We compare uniform weighting, using only the indicator outputs, using only prediction head confidences, and incorporating all components of our described attention mechanism. Attention performs most consistently on slices without worsening overall performance.

minimize the cross entropy between the prediction head’s output and the base task’s ground truth labels, $y$: $\ell_{\text{base}} = L_{\text{CE}}(f(z'), y)$.

The overall training loss combines loss values for all task heads: $\ell_{\text{train}} = \ell_{\text{base}} + \ell_{\text{ind}} + \ell_{\text{pred}}$

3.3 Synthetic experiments

To understand the properties of Slice-based Learning, we validate our model and its components (Figure 2) on a set of synthetic data with the above baselines. In our setup (Figure 1), we construct a dataset $X \in \mathbb{R}^2$ with a 2-way classification problem in which over 95% of the data are linearly separable. We introduce two minor perturbations along the decision boundary, which we define as critical slices, $S_1$ and $S_2$. Intuitively, examples that fall within these slices follow different distributions ($P(Y|X, s_i)$) relative to the overall data ($P(Y|X)$). For all models, the shared backbone is defined as a 2-layer MLP architecture with a backbone representation size $d = 13$ and a final ReLU non-linearity. OURS is initialized with a slice-representation size $h = 13$.

The model learns the slice-conditional label distribution $P(Y|S_i, X)$ from noisy SF inputs. We show in Figure 4 that the slices at the perturbed decision boundary cannot be learned in the general case, by a VANILLA model. As a result, we define two SFs, $\lambda_1$ and $\lambda_2$, to target the slices of interest. Because our attention-based model (OURS) is slice-aware, Figure 4 validates that it outperforms VANILLA, which has no notion of slices. Intuitively, if the model knows “where” in the data space an example lives (as defined by SFs), it can condition on slice information as it makes a final prediction. In Figure 5, we observe our model’s ability to cope with noisy SF inputs: the indicator is robust to moderate amounts of noise by ignoring noisy examples (middle); with extremely noisy inputs, it disregards poorly-defined SFs by assigning relatively uniform weights.

The model does not degrade overall performance. The primary goal of the slice-aware model is to improve slice-specific performance without degrading the model’s existing capabilities. We show in Figure 4 that OURS improves the overall score by 1.36 F1 points by learning the proportionally smaller perturbations in the decision boundary in addition to the more general linear boundary. Further, we note that we do not regress on individual slice heads.

Learning slice weights with features $P(Y|s_i, X)$ improves over doing so with only supervision source information $P(Y|s_i)$, A core assumption of our approach asserts that if the model learns improved slice-conditional weights via $\lambda_1$, downstream slice-specific performance will improve. Data programming (DP) [22] is a popular weak supervision approach deployed at numerous Fortune 500 companies [1], in which the weights of heuristics are learned solely from labeling source information. We emphasize that our setting provides the model with strictly more information—in the data’s feature representations—to learn such weights; we show in Figure 4 (right) that increasing representation size allows us to significantly outperform DP.
<table>
<thead>
<tr>
<th>Method</th>
<th>Performance (F1 score)</th>
<th>Overall</th>
<th>$S_1$</th>
<th>$S_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>VANILLA</td>
<td>96.56</td>
<td>52.94</td>
<td>68.75</td>
<td></td>
</tr>
<tr>
<td>DP [27]</td>
<td>96.88</td>
<td>44.12</td>
<td>43.75</td>
<td></td>
</tr>
<tr>
<td>HARD PARAM</td>
<td>96.72</td>
<td>50.00</td>
<td>75.00</td>
<td></td>
</tr>
<tr>
<td>MoE [16]</td>
<td>98.48</td>
<td>88.24</td>
<td>87.50</td>
<td></td>
</tr>
<tr>
<td>OURS</td>
<td>97.92</td>
<td>91.18</td>
<td>81.25</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: **Scaling with hidden feature representation dimensions.** We plot the quality versus the hidden dimension size. The slice-aware model (OURS) improves over HARD PARAM on both slices at a fixed hidden dimension size, while being close to MoE. Note: MoE has significantly more parameters in general as it copies the entire model.

Figure 5: **Coping with Noise:** We test the robustness of our approach on a simple synthetic example. In each panel, we show noisy SFs (left) as binary points and the corresponding slice indicator’s output (right) as a heatmap of probabilities. We show that the indicator assigns low relative probabilities on noisy (40%, middle) samples and ignores a very noisy (80%, right) SF, assigning relatively uniform scores to all samples.

**Attention weights learn from noisy $\lambda_i$ to combine slice residual representations.** Given slice information, the model achieves improvements over methods that do not aggregate slice information, as defined by each noisy $\lambda_i$. Both the indicator outputs ($Q$) and prediction confidence ($\text{abs}(P)$) is robustly combined in the attention mechanism. Even a noisy indicator will be upweighted if the predictions are high confidence, and if the indicator has high signal, even a slice expert making poor predictions can benefit from the underlying features. We show in Figure 4c that our attention-based architecture is able to combine slice-specific residual representations and (OURS) sees improvements over VANILLA of 38.2 slice-level F1 averaged across $S_1$ and $S_2$ (Table 4).

**Our model demonstrates similar expressivity to MoE with much less cost.** We come within 6.25 slice-level F1 averaged across $S_1$ and $S_2$ of MoE with approximately half as many parameters (Table 4). With large backbones, characterized by $M$ parameters, and a large number of slices, $k$, MoE requires a quadratically larger number of parameters because we initialize an entire backbone for each of the slices. In contrast, all other models scale linearly in parameters with $M$.

4 Experiments

We evaluate the efficacy of **Slice-based Learning** to improve slice-level performance for natural language understanding and computer vision tasks. We compare to several baselines and demonstrate that, using the same backbone architecture, our approach successfully models slice importance and significantly improves slice-level performance without impacting overall model performance.

4.1 Datasets and Slices

**Natural language understanding.** We select slices based on independently-conducted error analyses [18] (Appendix 1.2). In **Corpus of Linguistic Acceptability (CoLA)** [35], the task is to
We consider four baselines that capture alternatives we have seen in practice or the literature. Weights and fine-tuned for a fixed number of epochs and the optimal hyperparameters. We perform SuperGlue Table 1: Application Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>CoLA (Matthews Corr. [22])</th>
<th>RTE (F1 Score)</th>
<th>CYDET (F1 Score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall (std)</td>
<td>Max Lift</td>
<td>Overall (std)</td>
<td>Max Lift</td>
</tr>
<tr>
<td>VANILLA</td>
<td>57.8 (±1.3)</td>
<td>67.0 (±1.6)</td>
<td>39.4 (±5.4)</td>
</tr>
<tr>
<td>HARD PARAM [6]</td>
<td>57.4 (±2.1)</td>
<td>+12.7</td>
<td>67.9 (±1.8)</td>
</tr>
<tr>
<td>MANUAL</td>
<td>57.9 (±1.2)</td>
<td>+6.3</td>
<td>69.4 (±1.8)</td>
</tr>
<tr>
<td>MoE [10]</td>
<td>57.7 (±0.9)</td>
<td>+20.0</td>
<td>69.2 (±1.5)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Overview</th>
<th>Vanilla</th>
<th>Hard Param</th>
<th>Manual</th>
<th>MoE</th>
<th>OOM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slice Lift</td>
<td>58.3 (±0.7)</td>
<td>+19.0</td>
<td>+2.5</td>
<td>69.5 (±0.8)</td>
<td>+10.9</td>
</tr>
</tbody>
</table>

Table 1: Application Datasets: We compare our model to baselines averaged over 5 runs with different seeds in natural language understanding and computer vision applications. We report the overall score and maximum relative improvement (i.e. Lift) over the VANILLA model for each of the slice-aware approaches. For some trials of MoE, our system ran out of GPU memory (i.e. OOM).

predict whether a sentence is linguistically acceptable (i.e. grammatically); success is measured using Matthews correlation coefficient [22]. Natural slices might occur as questions or long sentences; such examples might consist of non-standard or challenging sentence structure. Since ground truth test labels are not available for this task (they are held out in evaluation servers [34]), we sample to create data splits with 7.2K/1.3K/1K train/valid/test sentences, respectively. To properly evaluate slices of interest, we ensure that the proportions of examples in ground truth slices are consistent across splits.

In Recognizing Textual Entailment (RTE) [34, 9, 2, 13, 3], the task is to predict whether or not a premise sentence entails a hypothesis sentence. Similar to CoLA, we create our own data splits and use 2.25K/0.25K/0.275K train/valid/test sentences, respectively. Finally, in a user study where we work with practitioners tackling the recently published SuperGlue [33] benchmark, we leverage Slice-based Learning to improve state-of-the-art quality on evaluation test servers.

Computer Vision. In the image domain, we evaluate on an autonomous vehicle dataset called Cyclist Detection for Autonomous Vehicles (CYDET) [20]. We leverage clips in a self-driving video dataset to detect whether a cyclist (person plus bicycle) is present at each frame. We select one independent clip for evaluation, and the remainder for training; for valid/test splits, we select alternating batches of five frames each from the evaluation clip. We preprocess the dataset with an open-source implementation of Mask R-CNN [21] to provide metadata (e.g. presence of traffic lights, benches), which serve as slice indicators for each frame.

4.2 Baselines

We consider four baselines that capture alternatives we have seen in practice or the literature. VANILLA: A vanilla neural network backbone is trained with a final prediction head to make predictions. This baseline represents the de-facto approach used in deep learning modeling tasks; it is unaware of the notion of slices and, as a result, neglects to model them.

MoE: We train a mixture of experts [16], where each expert is a separate VANILLA model trained on a data subset specified by \( s_i \). A gating network [32] is then trained to combine expert predictions into a final prediction.

HARD PARAM: In the style of multi-task learning, we model slices as separate task heads with a shared backbone trained via hard parameter sharing. Each slice task performs the same prediction task, but they are trained on subsets of data corresponding to \( s_i \). In this approach, backpropagation from different slice tasks is intended to encourage a slice-aware representation bias [30, 6].

MANUAL UPWEIGHT: To simulate the manual effort required to upweight hyperparameters for tuning slice-specific representations, we leverage the same architecture as HARD PARAM and grid search over a set of multipliers \( \alpha \in \{2, 20, 50, 100\} \) for loss terms of underperforming slices (i.e. \( \text{score}_{\text{overall}} - \text{score}_{\text{slice}} \geq 5 \) F1 points in VANILLA).

For each of the datasets, we first train the backbone model with a standard hyperparameter search over learning rate and \( \ell_2 \) regularization values. Then, each method is initialized from the backbone weights and fine-tuned for a fixed number of epochs and the optimal hyperparameters. We perform all empirical experiments on Google’s Cloud infrastructure using NVIDIA V100 GPUs.
4.3 Weak Supervision Baselines

To contextualize our contributions in the weak supervision literature, we compare directly to Data Programming (DP) [25], a popular approach for reweighting user-specified heuristics using supervision source information [27]. We consider two text-based relation extraction datasets: Chemical-Disease Relations (CDR), in which we identify causal links between chemical and disease entities in a dataset of PubMed abstracts, and Spouses [8], in which we identify mentions of spousal relationships using preprocessed pairs of person mentions from news articles (via Spacy [15]).

In both datasets, we leverage the exact noisy linguistic patterns and distant supervision heuristics provided in the open-source implementation of DP, and instead model them as slicing functions. Rather than voting on a particular class, we repurpose the provided labeling functions as binary slice indicators for our model. We then train our slice-aware model on the probabilistic labels aggregated from these heuristics.

4.4 Discussion

Slice-aware models improve slice-specific performance. We see in Table 1 that each slice-aware model (HARD PARAM, MANUAL, MOE, OURS) largely improves over the naive model.

OURS improves overall performance. We also observe that OURS improves overall performance for each of the datasets. This is likely because the chosen slices were explicitly modeled from error analysis papers, and explicitly modeling “error” slices led to improved overall performance.

OURS learns slice-specific representations consistently. While HARD PARAM and MANUAL perform well on some slices, they exhibit much higher variance compared to OURS and MOE (as denoted by the error bars in Table 1). These models lack an attention mechanism to combine and reweight slice representations in a consistent and measured way; instead, they rely purely on representation bias from slice-specific heads to improve slice-level performance. Because these representations are not modeled explicitly, improvements are largely driven by chance, and this approach risks worsening performance on other slices or overall.

OURS improves performance with a parameter-efficient representation. For CoLA and RTE experiments, we used the BERT-base [11] architecture with 110M parameters; for CyDet, we used ResNet-18 [14]. For each additional slice, OURS requires a 1.07% and 1.05% increase in relative parameter count in the BERT and ResNet architectures, respectively. As a comparison, HARD PARAM requires the same relative increase in parameters per slice. MOE on the other hand, increases relative number of parameters by 100% for both architectures. With limited increase in model size, OURS is able to outperform or match all other baselines, including MOE, which requires an order of magnitude more parameters.

OURS improves state-of-the-art quality models with slice-aware representations. In a submission to evaluation servers of the SuperGLUE benchmark, we leverage the same BERT-large architecture of previous submissions and observe improvements of +3.8/+2.8 avg. F1/acc. on CB, +2.4 acc. on COPA, +2.5 acc. on WiC; this amounts to an aggregate 2.7 point increase in overall benchmark score.

OURS improves over current weak supervision methods. Treating the noisy heuristics as slicing functions, we observe lifts of up to 1.3 F1 overall and 15.9 F1 on heuristically-defined slices. We reproduce the DP [25] setup to obtain overall scores of F1=41.9 on Spouses and F1=56.4 on CDR. Using Slice-based Learning, we improve to 42.8 (+0.9) and 57.7 (+1.3) F1, respectively. Intuitively, we can explain this improvement, because OURS has access to features of the data belonging to slices whereas DP relies only on the source information of each heuristic.

5 Conclusion

We introduced the challenge of improving slice-specific metrics without damaging the overall model quality, and introduced the first programming abstraction and machine learning model to support these actions. We demonstrated that the model could be used to push the state-of-the-art quality. In our analysis, we can explain consistent gains in the Slice-based Learning paradigm because our attention mechanism has access to a rich set of deep features, whereas existing weak supervision
paradigms have no way to access this information. We view this work in the context of programming models that sit on top of traditional modeling approaches in machine learning systems.

Acknowledgements We would like to thank Braden Hancock, Feng Niu, and Charles Srisuwananukorn for many helpful discussions, tests, and collaborations throughout the development of slicing. We gratefully acknowledge the support of DARPA under Nos. FA87501720095 (D3M), FA86501827865 (SDH), FA86501827882 (ASED), NIH under No. U54EB020405 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity) and CCF1563078 (Volume to Velocity), ONR under No. N000141712266 (Unifying Weak Supervision), the Moore Foundation, NXP, LETI-CEA, Intel, Microsoft, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, and American Family Insurance, Google Cloud, Swiss Re, and members of the Stanford DAWN project: Teradata, Facebook, Google, Ant Financial, NEC, SAP, VMware, and Infosys. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

1 Appendix

1.1 Model Characteristics

We include a further exploration of model characteristics to supplement Section 3.3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Slice-aware</th>
<th>No manual tuning</th>
<th>Weighted slice info.</th>
<th>Avoids copies of model (M params)</th>
<th>Num. Params</th>
</tr>
</thead>
<tbody>
<tr>
<td>VANILLA</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>O(M+r)</td>
<td></td>
</tr>
<tr>
<td>HARD PARAM</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>O(M+kr)</td>
<td></td>
</tr>
<tr>
<td>MANUAL</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>O(M+kr)</td>
<td></td>
</tr>
<tr>
<td>MOE</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>O(kM+kr)</td>
<td></td>
</tr>
<tr>
<td>Ours</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>O(M+krh)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Model characterizations: We characterize each model’s advantages/limitations and compute the number of parameters for each baseline model, given k slices, M backbone parameters, feature representation z dimension r, and slice-specific representation pi dimension h.

1.2 Slicing Function (SF) Construction

We walk through specific examples of SFs written for a number of our applications.

Textual SFs For text-based applications (CoLA, RTE), we write SFs over pairs of sentences for each task. Following dataset convention, we denote the first sentence as the premise and the second as the hypothesis where appropriate. Then, SFs are written, drawing largely from existing error analysis [18]. For instance, we might expect certain questions to be especially difficult to formulate in a language acceptability task. So, we write the following SF to heuristically target where questions:

```python
def SF_where_question(premise, hypothesis):
    # triggers if "where" appears in sentence
    sentences = premise + hypothesis
    return "where" in sentences.lower()
```

In some cases, we write SFs over both sentences at once. For instance, to capture possible errors in article references (e.g. the Big Apple vs a big apple), we identify multiple of the same article across sentences:

```python
def SF_has_multiple_articles(premise, hypothesis):
    # triggers if a sentence has more than one occurrence of the same article
    sentences = premise + hypothesis
    multiple_a = sum([int(x == "a") for x in sentences.split()]) > 1
    multiple_an = sum([int(x == "an") for x in sentences.split()]) > 1
    multiple_the = sum([int(x == "the") for x in sentences.split()]) > 1
    return multiple_a or multiple_an or multiple_the
```
Image-based SFs For computer vision applications, we leverage image metadata and bounding box attributes, generated from an off-the-shelf Mask R-CNN [21], to target slices of interest.

```python
def SF_bus(image):
    # triggers if a "bus" appears in the predictions of the noisy detector
    outputs = noisy_detector(image)
    return "bus" in outputs
```

We note that these potentially expensive detectors are run offline, and our model uses learned indicators at inference time. Despite the detectors’ noisy predictions, our model is able to reweight representations appropriately.

1.3 CoLA SFs

CoLA is a language acceptability task based on linguistics and grammar for individual sentences. We draw from error analysis which introduces several linguistically important slices for language acceptability via a series of challenge tasks. Each task consists of synthetically generated examples to measure model evaluation on specific slices. We heuristically define SFs to target subsets of data corresponding to each challenge, and include the full set of SFs derived from each category of challenge tasks:

- **Wh-words:** This task targets sentences containing *who, what, where, when, why, how*. We exclude *why* and *how* below because the CoLA dataset does not have enough examples for proper training and evaluation of these slices.

```python
def SF_where_in_sentence(sentence):
    return "where" in sentence
def SF_who_in_sentence(sentence):
    return "who" in sentence
def SF_what_in_sentence(sentence):
    return "what" in sentence
def SF_when_in_sentence(sentence):
    return "when" in sentence
```

- **Definite-Indefinite Articles:** This challenge task measures the model based on different combinations of definite (*the*) and indefinite (*a, an*) articles in a sentence (i.e. swapping definite for indefinite articles and vice versa). We target containing multiple uses of a definite (*the*) or indefinite article (*a, an*):

```python
def SF_has_multiple_articles(sentence):
    # triggers if a sentence has more than one occurrence of the same article
    multiple_indefinite = sum([int(x == "a") for x in sentence.split()]) > 1 or sum([int(x == "an") for x in sentence.split()]) > 1
    multiple_definite = sum([int(x == "the") for x in sentence.split()]) > 1
    return multiple_indefinite or multiple_definite
```

- **Coordinating Conjunctions:** This task seeks to measure correct usage of coordinating conjunctions (*and, but, or*) in context. We target the presence of these words in both sentences.

```python
def and_in_sentence(sentence):
    return "and" in sentence
def but_in_sentence(sentence):
    return "but" in sentence
def or_in_sentence(sentence):
    return "or" in sentence
```

- **End-of-Sentence:** This task measures a model’s ability to identify coherent sentences or sentence chunks after removing punctuation. We heuristically target this slice by identifying particularly short sentences and those that end with verbs and adverbs. We use helper functions from Spacy [15] to process parts of speech.

```python
def SF_short_sentence(sentence):
    # triggered if sentence has fewer than 5 tokens
    return len(sentence.split()) < 5

# spacy helper
def get_spacy_pos(sentence):
    import spacy
    nlp = spacy.load("en_core_web_sm")
```
1.4 RTE SFs

Similar to CoLA, we use challenge tasks from NLI-based error analysis [18] to write SFs over the textual entailment (RTE) dataset.

- **Prepositions**: The authors swap around a manually-curated list of prepositions. Because the list in its entirety spans a large proportion of the RTE dataset, we separate them into temporal and possessive segments.

```python
def SF_has_temporal_preposition(premise, hypothesis):
    temporal_prepositions = ["after", "before", "past"]
    sentence = premise + sentence
    return any([p in sentence for p in temporal_prepositions])

def SF_has_possessive_preposition(premise, hypothesis):
    possessive_prepositions = ["inside of", "with", "within"]
    sentence = premise + sentence
    return any([p in sentence for p in possessive_prepositions])
```

- **Comparatives**: The authors choose sentences with specific comparative words and mutate/negate them in their challenge task. We directly target keywords identified in their approach.

```python
def SF_is_comparative(premise, hypothesis):
    comparative_words = ["more", "less", "better", "worse", "bigger", "smaller"]
    sentence = premise + hypothesis
    return any([p in sentence for p in comparative_words])
```

- **Quantification**: The quantification challenge tests natural language understanding with common quantifiers. We target common quantifiers in both the combined premise/hypothesis and in only the hypothesis.

```python
def is_quantification(premise, hypothesis):
    quantifiers = ["all", "some", "none"]
    sentence = premise + hypothesis
    return any([p in sentence for p in quantifiers])

def is_quantification_hypothesis(premise, hypothesis):
    quantifiers = ["all", "some", "none"]
    return any([p in hypothesis for p in quantifiers])
```

- **Spatial Expressions**: This task identifies spatial relations between entities (i.e. A is to the left of B). We exclude this task from our slices, as they do not account for enough examples in the RTE dataset.

- **Negation**: This challenge task identifies whether natural language inference models can handle negations. We heuristically target this slice via a list of common negation words.

```python
def SF_common_negation(premise, hypothesis):
    # Words from https://www.grammarly.com/blog/negatives/
    negation_words = ["no", "not", "none", "no one", "nobody", "nothing", "neither", "nowhere", "never", "hardly", "scarcely", "barely", "doesn't", "isn't", "wasn't", "shouldn't",
```

12
"wouldn't",
"couldn't",
"won't",
"can't",
"don't",
]

sentence = premise + hypothesis
return any([x in negation_words for x in sentence])

• **Premise/Hypothesis Length**: Finally, separate from the cited error analysis, we target different length hypotheses and premises as an additional set of slicing tasks. In our own error analysis of the RTE model, we found these represented intuitive slices (i.e. long premises are typically harder to parse for key information, shorter hypotheses share syntactical structure).

```python
def SF_short_hypothesis(premise, hypothesis):
    return len(hypothesis.split()) < 5

def SF_long_hypothesis(premise, hypothesis):
    return len(hypothesis.split()) > 100

def SF_short_premise(premise, hypothesis):
    return len(premise.split()) < 15

def SF_long_premise(premise, hypothesis):
    return len(premise.split()) > 100
```

### 1.5 CYDet SFs

For the cyclist detection dataset, we identify subsets that correspond to other objects in the scene using a noisy detector (i.e. an off-the-shelf Mask R-CNN [21]).

```python
# define noisy detector
def noise_detector(image):
    probs = mask_rcnn.forward(image)

    # threshold predictions
    preds = []
    for object in classes:
        if probs[object] > 0.5:
            preds.append(object)
    return preds

# Cyclist Detection SFs
def SF_bench(image):
    outputs = noisy_detector(image)
    return "bench" in outputs

def SF_truck(image):
    outputs = noisy_detector(image)
    return "truck" in outputs

def SF_car(image):
    outputs = noisy_detector(image)
    return "car" in outputs

def SF_bus(image):
    outputs = noisy_detector(image)
    return "bus" in outputs

def SF_person(image):
    outputs = noisy_detector(image)
    return "person" in outputs

def SF_traffic_light(image):
    outputs = noisy_detector(image)
    return "traffic light" in outputs

def SF_fire_hydrant(image):
    outputs = noisy_detector(image)
    return "fire hydrant" in outputs

def SF_stop_sign(image):
    outputs = noisy_detector(image)
    return "stop sign" in outputs

def SF_bicycle(image):
    outputs = noisy_detector(image)
```
1.6 Slice-specific Metrics

We visualize slice-specific metrics across each application dataset, for each method of comparison. We report the corresponding aggregate metrics in Figure 1 (below).

In CoLA, we see that MoE and OURS exhibit the largest slice-specific gains, and also overfit on the same slice ends with adverb. In RTE, we see that OURS improves performance on all slices except common negation, where it falls less than a point below VANILLA. On CYDET, we see the largest gains for OURS on bench and bus slices—in particular, we are able to improve in cases where the model might able to use the presence of these objects to make more informed decisions about whether a cyclist is present. Note: because the MoE model on CYDET encounters an “Out of Memory” error, the corresponding (blue) data bar is not available for this dataset.
Figure 6: For each application dataset (Section 4.1) we report all relative, slice-level metrics compared to VANILLA for each model.
References


