

# Loss Functions for Multiset Prediction

Sean Welleck<sup>1,2</sup>, Zixin Yao<sup>1</sup>, Yu Gai<sup>1</sup>, Jialin Mao<sup>1</sup>, Zheng Zhang<sup>1</sup>, Kyunghyun Cho<sup>2,3</sup>

<sup>1</sup>New York University Shanghai

<sup>2</sup>New York University

<sup>3</sup>CIFAR Azrieli Global Scholar

{wellecks,zixin.yao,yg1246,jialin.mao,zz,kyunghyun.cho}@nyu.edu

## Abstract

We study the problem of multiset prediction. The goal of multiset prediction is to train a predictor that maps an input to a multiset consisting of multiple items. Unlike existing problems in supervised learning, such as classification, ranking and sequence generation, there is **no known order among items in a target multiset**, and **each item in the multiset may appear more than once**, making this problem extremely challenging. In this paper, we propose a novel multiset loss function by viewing this problem from the perspective of sequential decision making. The proposed multiset loss function is empirically evaluated on two families of datasets, one synthetic and the other real, with varying levels of difficulty, against various baseline loss functions including reinforcement learning, sequence, and aggregated distribution matching loss functions. The experiments reveal the effectiveness of the proposed loss function over the others.

## 1 Introduction

A relatively less studied problem in machine learning, particularly supervised learning, is the problem of multiset prediction. The goal of this problem is to learn a mapping from an arbitrary input to a multiset<sup>1</sup> of items. This problem appears in a variety of contexts. For instance, in the context of high-energy physics, one of the important problems in a particle physics data analysis is to count how many physics objects, such as electrons, muons, photons, taus, and jets, are in a collision event [6]. In computer vision, object counting and automatic alt-text can be framed as multiset prediction [30, 14].

In multiset prediction, a learner is presented with an arbitrary input and the associated multiset of items. It is assumed that there is no predefined order among the items, and that there are no further annotations containing information about the relationship between the input and each of the items in the multiset. These properties make the problem of multiset prediction unique from other well-studied problems. It is different from sequence prediction, because there is no known order among the items. It is not a ranking problem, since each item may appear more than once. It cannot be transformed into classification, because the number of possible multisets grows exponentially with respect to the maximum multiset size.

In this paper, we **view multiset prediction as a sequential decision making process**. Under this view, the problem reduces to finding a policy that **sequentially predicts one item at a time, while the outcome is still evaluated based on the aggregate multiset of the predicted items**. We first propose an oracle policy that assigns non-zero probabilities only to prediction sequences that result exactly in the target, ground-truth multiset given an input. This oracle is optimal in the sense that its prediction never decreases the precision and recall regardless of previous predictions. That is, its decision is optimal in any state (i.e., prediction prefix). We then propose a novel *multiset loss* which minimizes

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

<sup>1</sup>A set that allows **multiple instances**, e.g.  $\{x, y, x\}$ . See Appendix A for a detailed definition.

the KL divergence between the oracle policy and a parametrized policy at every point in a decision trajectory of the parametrized policy.

We compare the proposed multiset loss against an extensive set of baselines. They include a sequential loss with an arbitrary rank function, sequential loss with an input-dependent rank function, and an aggregated distribution matching loss and its one-step variant. We also test policy gradient, as was done in [30] recently for multiset prediction. Our evaluation is conducted on two sets of datasets with varying difficulties and properties. According to the experiments, we find that the proposed multiset loss outperforms all the other loss functions.

## 2 Multiset Prediction

A multiset prediction problem is a generalization of classification, where a target is not a single class but a multiset of classes. The goal is to find a mapping from an input  $x$  to a multiset  $\mathcal{Y} = \{y_1, \dots, y_{|\mathcal{Y}|}\}$ , where  $y_k \in \mathcal{C}$ . Some of the core properties of multiset prediction are; (1) the input  $x$  is an arbitrary vector, (2) there is no predefined order among the items  $y_i$  in the target multiset  $\mathcal{Y}$ , (3) the size of  $\mathcal{Y}$  may vary depending on the input  $x$ , and (4) each item in the class set  $\mathcal{C}$  may appear more than once in  $\mathcal{Y}$ . Formally,  $\mathcal{Y}$  is a multiset  $\mathcal{Y} = (\mu, \mathcal{C})$ , where  $\mu : \mathcal{C} \rightarrow \mathbb{N}$  gives the number of occurrences of each class  $c \in \mathcal{C}$  in the multiset. See Appendix A for a further review of multisets.

As is typical in supervised learning, in multiset prediction a model  $f_\theta(x)$  is trained on a dataset  $\{(x_i, \mathcal{Y}_i)\}_{i=1}^N$ , then evaluated on a separate test set  $\{(x_i, \mathcal{Y}_i)\}_{i=1}^n$  using evaluation metrics  $m(\cdot, \cdot)$  that compare the predicted and target multisets, i.e.  $\frac{1}{n} \sum_{i=1}^n m(\hat{\mathcal{Y}}_i, \mathcal{Y}_i)$ , where  $\hat{\mathcal{Y}}_i = f_\theta(x_i)$  denotes a predicted multiset. For evaluation metrics we use exact match  $\text{EM}(\hat{\mathcal{Y}}, \mathcal{Y}) = \mathbb{I}[\hat{\mathcal{Y}} = \mathcal{Y}]$ , and the  $F_1$  score. Refer to Appendix A for multiset definitions of exact match and  $F_1$ .

## 3 Related Problems in Supervised Learning

Variants of multiset prediction have been studied earlier. We now discuss a taxonomy of approaches in order to differentiate our proposal from previous work and define strong baselines.

### 3.1 Set Prediction

**Ranking** A ranking problem can be considered as learning a mapping from a pair of input  $x$  and one of the items  $c \in \mathcal{C}$  to its score  $s(x, c)$ . All the items in the class set are then sorted according to the score, and this sorted order determines the rank of each item. Taking the top- $K$  items from this sorted list results in a predicted set (e.g. [7]). Similarly to multiset prediction, the input  $x$  is arbitrary, and the target is a set without any prespecific order. However, ranking differs from multiset prediction in that it is unable to handle multiple occurrences of a single item in the target set.

**Multi-label Classification via Binary Classification** Multi-label classification consists of learning a mapping from an input  $x$  to a subset of classes identified as  $\mathbf{y} \in \{0, 1\}^{|\mathcal{C}|}$ . This problem can be reduced to  $|\mathcal{C}|$  binary classification problems by learning a binary classifier for each possible class. Representative approaches include binary relevance, which assumes classes are conditionally independent, and probabilistic classifier chains which decompose the joint probability as  $p(\mathbf{y}|x) = \prod_{c=1}^{|\mathcal{C}|} p(y_c | \mathbf{y}_{<c}, x)$  [4, 19, 27, 8]. Since each  $p(y_c | \mathbf{y}_{<c}, x)$  models *binary* membership of a particular class, their predictions form a set  $\hat{\mathbf{y}} \in \{0, 1\}^{|\mathcal{C}|}$  rather than a multiset  $\hat{\mathbf{y}} \in \mathbb{N}^{|\mathcal{C}|}$ .

### 3.2 Parallel Prediction

**Power Multiset Classification** A brute-force approach based on the Combination Method in multi-label classification [27, 19], is to transform the class set  $\mathcal{C}$  into a set  $M(\mathcal{C})$  of all possible multisets, then train a multi-class classifier  $\pi$  that maps an input  $x$  to one of the elements in  $M(\mathcal{C})$ . However, the number of all possible multisets grows exponentially in the maximum size of a target multiset,<sup>2</sup> rendering this approach infeasible in practice.

<sup>2</sup>The number of all possible multisets of size  $\leq K$  is  $\sum_{k=1}^K \frac{(|\mathcal{C}|+k-1)!}{k!(|\mathcal{C}|-1)!}$ .

**One Step Distribution Matching** Instead of considering the target multiset as an actual multiset, one can convert it into a distribution over the class set, using each item’s multiplicity. That is, we consider a target multiset  $\mathcal{Y}$  as a set of samples from a single, underlying distribution  $q^*$  over the class set  $C$ , empirically estimated as  $q^*(c|x) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} I_{y=c}$ , where  $I$  is an indicator function. A model then outputs a point  $q_\theta(\cdot|x)$  in a  $|C|$ -dimensional simplex and is trained by minimizing a divergence between  $q_\theta(\cdot|x)$  and  $q^*(c|x)$ . The model also predicts the size  $\hat{l}_\theta(x)$  of the target multiset, so that each unique  $c \in C$  has a predicted cardinality  $\hat{\mu}(c) = \text{round}(q_\theta^c(x) \cdot \hat{l}_\theta(x))$ . An un-normalized variant could directly regress the cardinality of each class.

A major weakness of these methods is the lack of modeling dependencies among the items in the predicted multiset, a known issue in multi-label classification [4, 16]. We test this approach in the experiments ( $\mathcal{L}_{1\text{-step}}$ ) and observe substantially worse prediction accuracy than other baselines.

### 3.3 Sequential Methods

**Sequence prediction** A sequence prediction problem is characterized as finding a mapping from an input  $x$  to a sequence of classes  $\mathcal{Y}_{\text{seq}} = (y_1, \dots, y_{|\mathcal{Y}|})$ . It is different from multiset prediction since a sequence has a predetermined order of items, while a multiset is an unordered collection. Multiset prediction can however be treated as sequence prediction by defining an ordering for each multiset. Each target multiset  $\mathcal{Y}$  is then transformed into an ordered sequence  $\mathcal{Y}_{\text{seq}} = (y_1, \dots, y_{|\mathcal{Y}|})$ , a model predicts a sequence  $\hat{\mathcal{Y}}_{\text{seq}} = (\hat{y}_1, \dots, \hat{y}_{|\mathcal{Y}|})$ , and a per-step loss  $\mathcal{L}_{\text{seq}}$  is minimized using  $\mathcal{Y}_{\text{seq}}$  and  $\hat{\mathcal{Y}}_{\text{seq}}$ .

Recently, multi-label classification (i.e. set prediction) was posed as sequence prediction with RNNs [29, 16], improving upon methods that do not model conditional label dependencies. However, these approaches and the  $\mathcal{L}_{\text{seq}}$  approach outlined above require a pre-specified rank function which orders output sequences (e.g. class prevalence in [29]).

Because multiset prediction does not come with such a rank function by definition, we must choose a (often ad-hoc) rank function, and performance can significantly vary based on the choice. Vinyals et al. [28] observed this variation in sequence-based set prediction (also observed in [16, 29]), which we confirm for multisets in section 5.3. This shows the importance of our proposed method, which does not require a fixed label ordering.

Unlike  $\mathcal{L}_{\text{seq}}$ , our multiset loss  $\mathcal{L}_{\text{multiset}}$  proposed below is permutation invariant with respect to the order of the target multiset, and is thus not susceptible to performance variations from choosing a rank function, since such a choice is not required. We use  $\mathcal{L}_{\text{seq}}$  as a baseline in Experiment 3, finding that it underperforms the proposed  $\mathcal{L}_{\text{multiset}}$ .

**Aggregated Distribution Matching** As in one-step distribution matching, a multiset is treated as a distribution  $q_*$  over classes. The sequential variant predicts a sequence of classes  $(y_1, \dots, y_{|\mathcal{Y}|})$  by sampling from a predicted distribution  $q_\theta^{(t)}(y_t|y_{<t}, x)$  at each step  $t$ . The per-step distributions  $q_\theta^{(t)}$  are averaged into an aggregate distribution  $q_\theta$ , and a divergence between  $q_*$  and  $q_\theta$  is minimized. We test  $L_1$  distance and KL-divergence in the experiments ( $\mathcal{L}_{\text{dm}}^p, \mathcal{L}_{\text{dm}}^{\text{KL}}$ ).

A major issue with this approach is that it may assign non-zero probability to an incorrect sequence of predictions due to the aggregated distribution’s invariance to the order of predictions. This is reflected in an increase in the entropy of  $q_\theta^{(t)}$  over time, discussed in Experiment 3.

**Reinforcement Learning** In [30], an approach based on reinforcement learning (RL) was proposed for multiset prediction. In this approach, a policy  $\pi_\theta$  samples a multiset as a sequential trajectory, and the goal is finding  $\pi_\theta$  whose trajectories maximize a reward function designed specifically for multiset prediction. REINFORCE [31] is used to minimize the resulting loss function, which is known to be difficult due to high variance [18]. We test the RL method in the experiments ( $\mathcal{L}_{\text{RL}}$ ).

### 3.4 Domain-Specific Methods

In computer vision, object counting and object detection are instances of multiset prediction. Typical object counting approaches in computer vision, e.g. [14, 33, 17], model the counting problem as density estimation over image space, and assume that each object is annotated with a dot specifying its location. Object detection methods (e.g. [24, 21, 20, 9]) also require object location annotations. Since these approaches exploit the fact the input is an image and rely on additional annotated

information, they are not directly comparable to our method which only assumes annotated class labels and is agnostic to the input modality.

## 4 Multiset Loss Function for Multiset Prediction

In this paper, we propose a novel loss function, called *multiset loss*, for the problem of multiset prediction. This loss function is best motivated by treating the multiset prediction problem as a sequential decision making process with a model being considered a policy  $\pi$ . This policy, parametrized by  $\theta$ , takes as input the input  $x$  and all the previously predicted classes  $\hat{y}_{<t}$  at time  $t$ , and outputs the distribution over the next class to be predicted. That is,  $\pi_\theta(y_t|\hat{y}_{<t}, x)$ .

We first define a free label multiset at time  $t$ , which contains all the items that remain to be predicted after  $t - 1$  predictions by the policy, as

**Definition 1** (Free Label Multiset).

$$\mathcal{Y}_t \leftarrow \mathcal{Y}_{t-1} \setminus \{\hat{y}_{t-1}\},$$

where  $\hat{y}_{t-1}$  is the prediction made by the policy at time  $t - 1$ .

We then construct an oracle policy  $\pi_*$ . This oracle policy takes as input a sequence of predicted labels  $\hat{y}_{<t}$ , the input  $x$ , and the free label multiset with respect to its predictions,  $\mathcal{Y}_t = \mathcal{Y} \setminus \{\hat{y}_{<t}\}$ . It outputs a distribution whose entire probability (1) is evenly distributed over all the items in the free label multiset  $\mathcal{Y}_t$ . In other words,

**Definition 2** (Oracle).

$$\pi_*(y_t|\hat{y}_{<t}, x, \mathcal{Y}_t) = \begin{cases} \frac{1}{|\mathcal{Y}_t|}, & \text{if } y_t \in \mathcal{Y}_t \\ 0, & \text{otherwise} \end{cases}$$

An interesting and important property of this oracle is that it is optimal given any prefix  $\hat{y}_{<t}$  with respect to both precision and recall. This is intuitively clear by noticing that the oracle policy allows only a correct item to be selected. We call this property the optimality of the oracle.

**Remark 1.** Given an arbitrary prefix  $\hat{y}_{<t}$ ,

$$\text{Prec}(\hat{y}_{<t}, \mathcal{Y}) \leq \text{Prec}(\hat{y}_{<t} \cup \hat{y}, \mathcal{Y}) \text{ and } \text{Rec}(\hat{y}_{<t}, \mathcal{Y}) \leq \text{Rec}(\hat{y}_{<t} \cup \hat{y}, \mathcal{Y}),$$

for any  $\hat{y} \sim \pi_*(\hat{y}_{<t}, x, \mathcal{Y}_t)$ .

From the remark above, it follows that the oracle policy is optimal in terms of precision and recall.

**Remark 2.**

$$\text{Prec}(\hat{y}_{\leq|\mathcal{Y}|}, \mathcal{Y}) = 1 \text{ and } \text{Rec}(\hat{y}_{\leq|\mathcal{Y}|}, \mathcal{Y}) = 1, \text{ for all } \hat{y}_{\leq|\mathcal{Y}|} \sim \prod_{t=1}^{|\mathcal{Y}|} \pi_*(y_t|\hat{y}_{<t}, x, \mathcal{Y}_t).$$

It is trivial to show that sampling from such an oracle policy would never result in an incorrect prediction. That is, this oracle policy assigns zero probability to any sequence of predictions that is not a permutation of the target multiset.

**Remark 3.**

$$\prod_{t=1}^{|\mathcal{Y}|} \pi_*(y_t|\hat{y}_{<t}, x) = 0, \text{ if multiset}(y_1, \dots, y_{|\mathcal{Y}|}) \neq \mathcal{Y},$$

where multiset equality refers to exact match, as defined in Appendix 1.

In short, this oracle policy tells us at each time step  $t$  which of all the items in the class set  $C$  must be selected. By selecting an item according to the oracle, the free label multiset decreases in size. Since the oracle distributes equal probability over items in the free label multiset, the oracle policy's entropy decreases over time.

**Remark 4** (Decreasing Entropy).

$$\mathcal{H}(\pi_*^{(t)}) > \mathcal{H}(\pi_*^{(t+1)}),$$

where  $\mathcal{H}(\pi_*^{(t)})$  denotes the Shannon entropy of the oracle policy at time  $t$ ,  $\pi_*(y|\hat{y}_{<t}, x, \mathcal{Y}_t)$ .

Proofs of the remarks above can be found in Appendix B–D.

The oracle’s optimality allows us to consider a step-wise loss between a parametrized policy  $\pi_\theta$  and the oracle policy  $\pi_*$ , because the oracle policy provides us with an optimal decision regardless of the quality of the prefix generated so far. We thus propose to minimize the KL divergence from the oracle policy to the parametrized policy at each step separately. This divergence is defined as

$$\text{KL}(\pi_*^t \parallel \pi_\theta^t) = \underbrace{\mathcal{H}(\pi_*^t)}_{\text{const. w.r.t. } \theta} - \sum_{y_j \in |\mathcal{Y}_t|} \frac{1}{|\mathcal{Y}_t|} \log \pi_\theta(y_j | \hat{y}_{<t}, x),$$

where  $\mathcal{Y}_t$  is formed using predictions  $\hat{y}_{<t}$  from  $\pi_\theta$ , and  $\mathcal{H}(\pi_*^t)$  is the entropy of the oracle policy at time step  $t$ . This entropy term can be safely ignored when learning  $\pi_\theta$ , since it is constant with respect to  $\theta$ . We then define a per-step loss function as  $\mathcal{L}^t(x, \mathcal{Y}, \hat{y}_{<t}, \theta) = \text{KL}(\pi_*^t \parallel \pi_\theta^t) - \mathcal{H}(\pi_*^t)$ . The KL divergence may be replaced with another divergence.

It is intractable to minimize this per-step loss for every possible state  $(\hat{y}_{<t}, x)$ , since the size of the state space grows exponentially with respect to the size of a target multiset. We thus propose here to minimize the per-step loss only for the state, defined as a pair of the input  $x$  and the prefix  $\hat{y}_{<t}$ , visited by the parametrized policy  $\pi_\theta$ . That is, we generate an entire trajectory  $(\hat{y}_1, \dots, \hat{y}_T)$  by executing the parametrized policy until either all the items in the target multiset have been predicted or the predefined maximum number of steps have passed. Then, we compute the loss function at each time  $t$  based on  $(x, \hat{y}_{<t})$ , for all  $t = 1, \dots, T$ . The final loss function is the sum of all these per-step loss functions:

**Definition 3** (Multiset Loss Function).

$$\mathcal{L}_{\text{multi}}(x, \mathcal{Y}, \theta) = - \sum_{t=1}^T \frac{1}{|\mathcal{Y}_t|} \sum_{y_j \in \mathcal{Y}_t} \log \pi_\theta(y_j | \hat{y}_{<t}, x),$$

where  $T$  is the smaller of the smallest  $t$  for which  $\mathcal{Y}_t = \emptyset$  and the predefined maximum value.

By Remarks 2 and 3, minimizing this loss function maximizes F1 and exact match.

**Execution Strategies** As was shown in [22], the use of the parametrized policy  $\pi_\theta$  instead of the oracle policy  $\pi_*$  allows the upper bound on the learned policy’s error to be linear with respect to the size of the target multiset. If the oracle policy had been used, the upper bound would have grown quadratically with respect to the size of the target multiset. To confirm this empirically, we test the following three alternative strategies for executing the parametrized policy  $\pi_\theta$ : (1) Greedy search:  $\hat{y}_t = \arg \max_y \log \pi_\theta(y | \hat{y}_{<t}, x)$ , (2) Stochastic sampling:  $\hat{y}_t \sim \pi_\theta(y | \hat{y}_{<t}, x)$ , and (3) Oracle sampling:  $\hat{y}_t \sim \pi_*(y | \hat{y}_{<t}, x, \mathcal{Y}_t)$ . After training, the learned policy is evaluated by greedily selecting each item from the policy.

**Variable-Sized Multisets** In order to predict variable-sized multisets with the proposed loss functions, we introduce a **termination policy**  $\pi_s$ , which outputs a stop distribution given the predicted sequence of items  $\hat{y}_{<t}$  and the input  $x$ . Because the size of the target multiset is known during training, we simply train this termination policy in a supervised way using a binary cross-entropy loss. At evaluation time, we simply threshold the predicted stop probability at a predefined threshold (0.5). An alternative method for supporting variable-sized multisets is discussed in Appendix E.

**Relation to Learning to Search** Our framing of multiset prediction as a sequential task based on learning to imitate an oracle policy is inspired by the Learning to Search (L2S) approach to structured prediction [3, 2]. Recently, Leblond et al. [12] proposed SeaRNN, adapting L2S to modern recurrent models. Our proposal can be seen as designing an oracle and loss with favorable properties for multiset prediction, using a learned roll-in  $\pi_\theta$ , and directly setting a cost vector equal to the oracle’s distribution, avoiding the expensive per-step roll-out in SeaRNN. We believe that applying the general L2S framework to novel problem settings is an important research direction.

## 5 Experiments and Analysis

### 5.1 Datasets

**MNIST Multi** MNIST Multi is a class of synthetic datasets. Each dataset consists of multiple 100x100 images, each of which contains a varying number of digits from the original MNIST [13].

Table 1: Influence of rank function choice

	<b>MNIST Multi (4)</b>		<b>COCO Easy</b>	
	EM	F1	EM	F1
<b>Random</b>	0.920	0.977	0.721	0.779
<b>Area</b>	0.529	0.830	0.700	0.763
<b>Spatial</b>	0.917	0.976	0.675	0.738

Table 2: Execution Strategies

	<b>COCO Medium</b>	
	EM	F1
<b>Greedy</b>	$0.475 \pm 0.006$	$0.645 \pm 0.016$
<b>Stochastic</b>	$0.475 \pm 0.004$	$0.649 \pm 0.009$
<b>Oracle</b>	$0.469 \pm 0.002$	$0.616 \pm 0.009$

We vary the size of each digit and also add clutter. In the experiments, we consider the following variants of MNIST Multi:

- **MNIST Multi (4):**  $|\mathcal{Y}| = 4$ ; 20-50 px digits
- **MNIST Multi (1-4):**  $|\mathcal{Y}| \in 1, \dots, 4$ ; 20-50 px digits
- **MNIST Multi (10):**  $|\mathcal{Y}| = 10$ ; 20 px digits

Each dataset has a training set with 70,000 examples and a test set with 10,000 examples. We randomly sample 7,000 examples from the training set to use as a validation set, and train with the remaining 63,000 examples.

**MS COCO** As a real-world dataset, we use Microsoft COCO [15] which includes natural images with multiple objects. Compared to MNIST Multi, each image in MS COCO has objects of more varying sizes and shapes, and there is a large variation in the number of object instances per image which spans from 1 to 91. The problem is made even more challenging with many overlapping and occluded objects. To better control the difficulty, we create the following two variants:

- **COCO Easy:**  $|\mathcal{Y}| = 2$ ; 10,230 examples, 24 classes
- **COCO Medium:**  $|\mathcal{Y}| \in 1, \dots, 4$ ; 44,121 training examples, 23 classes

In both of the variants, we only include images whose  $|\mathcal{Y}|$  objects are large and of common classes. An object is defined to be large if the object’s area is above the 40-th percentile across the training set of MS COCO. After reducing the dataset to have  $|\mathcal{Y}|$  large objects per image, we remove images containing only objects of rare classes. A class is considered rare if its frequency is less than  $\frac{1}{|\mathcal{C}|}$ , where  $\mathcal{C}$  is the class set. These two stages ensure that only images with a proper number of large objects are kept. We do not use fine-grained annotation (pixel-level segmentation and bounding boxes) except for creating input-dependent rank functions for the  $\mathcal{L}_{\text{seq}}$  baseline (see Appendix F.2).

For each variant, we hold out a randomly sampled 15% of the training examples as a validation set. We form separate test sets by applying the same filters to the COCO validation set. The test set sizes are 5,107 for COCO Easy and 21,944 for COCO Medium.

## 5.2 Models

**MNIST Multi** We use three convolutional layers of channel sizes 10, 10 and 32, followed by a convolutional long short-term memory (LSTM) layer [32]. At each step, the feature map from the convolutional LSTM layer is average-pooled spatially and fed to a softmax classifier. In the case of the one-step variant of aggregate distribution matching, the LSTM layer is skipped.

**MS COCO** We use a ResNet-34 [10] pretrained on ImageNet [5] as a feature extractor. The final feature map from this ResNet-34 is fed to a convolutional LSTM layer, as described for MNIST Multi above. We do not finetune the ResNet-34 based feature extractor.

In all experiments, for predicting variable-sized multisets we use the termination policy approach since it is easily applicable to all of the baselines, thus ensuring a fair comparison. Conversely, it is unclear how to extend the special class approach to the distribution matching baselines.

**Training and evaluation** For each loss, a model was trained for 200 epochs (350 for MNIST Multi 10). After each epoch, exact match was computed on the validation set. The model with the highest validation exact match was used for evaluation on the test set. See Appendix E for more details.

When evaluating a trained policy, we use greedy decoding. Each predicted multiset is compared against the ground-truth target multiset, and we report both the exact match accuracy (EM) and F-1 score (F1), as defined in Appendix 1.

Table 3: Loss function comparison

	(a) MNIST Variants						(b) MS COCO Variants			
	Multi (4)		Multi (1-4)		Multi (10)		Easy		Medium	
	EM	F1	EM	F1	EM	F1	EM	F1	EM	F1
$\mathcal{L}_{\text{multi}}$	<b>0.950</b>	<b>0.987</b>	<b>0.953</b>	<b>0.981</b>	<b>0.920</b>	<b>0.992</b>	0.702	<b>0.788</b>	<b>0.481</b>	<b>0.639</b>
$\mathcal{L}_{\text{RL}}$	0.912	0.977	0.945	0.980	0.665	0.970	0.672	0.746	0.425	0.564
$\mathcal{L}_{\text{dm}}^1$	0.921	0.978	0.918	0.969	0.239	0.714	0.533	0.614	0.221	0.085
$\mathcal{L}_{\text{dm}}^{\text{KL}}$	0.908	0.974	0.908	0.962	0.256	0.874	<b>0.714</b>	0.763	0.444	0.591
$\mathcal{L}_{\text{seq}}$	0.906	0.973	0.891	0.952	0.592	0.946	0.709	0.774	0.457	0.592
$\mathcal{L}_{1\text{-step}}$	0.210	0.676	0.055	0.598	0.032	0.854	0.552	0.664	0.000	0.446

### 5.3 Experiment 1: Influence of a Rank Function on Sequence Prediction

First, we investigate the sequence loss function  $\mathcal{L}_{\text{seq}}$  from Sec. 3.3, while varying a rank function. We test three alternatives: a random rank function<sup>3</sup> and two input-dependent rank functions  $r_{\text{spatial}}$  and  $r_{\text{area}}$ .  $r_{\text{spatial}}$  orders labels in left-to-right, top-to-bottom order, and  $r_{\text{area}}$  orders labels by decreasing object area; see Appendix F for more detail. We compare these rank functions on MNIST Multi (4) and COCO Easy validation sets.

We present the results in Table 1. It is clear from the results that the performance of the sequence prediction loss function is dependent on the choice of a rank function. In the case of MNIST Multi, the area-based rank function was far worse than the other choices. However, this was not true on COCO Easy, where the spatial rank function was worst among the three. In both cases, we have observed that the random rank function performed best, and from here on, we use the random rank function in the remaining experiments. This set of experiments firmly suggests the need of an order-invariant multiset loss function, such as the proposed multiset loss function.

### 5.4 Experiment 2: Execution Strategies for the Multiset Loss Function

In this set of experiments, we compare the three execution strategies for the proposed multiset loss function, illustrated in Sec. 3. They are **greedy** decoding, **stochastic** sampling and **oracle** sampling. We test them on the most challenging dataset, COCO Medium, and report the mean and standard deviation for the evaluation metrics across 5 runs.

As shown in Table 2, greedy decoding and stochastic sampling, both of which consider states that are likely to be visited by the parametrized policy, outperform the oracle sampling, which only considers states on optimal trajectories. This is particularly apparent in the F1 score, which can be increased even after visiting a state that is not on an optimal trajectory. The results are consistent with the theory from [22, 2]. The performance difference between the first two strategies was not significant, so from here on we choose the simpler method, greedy decoding, when training a model with the proposed multiset loss function.

### 5.5 Experiment 3: Loss Function Comparison

We now compare the proposed multiset loss function against the five baseline loss functions: reinforcement learning  $\mathcal{L}_{\text{RL}}$ , aggregate distribution matching— $\mathcal{L}_{\text{dm}}^1$  and  $\mathcal{L}_{\text{dm}}^{\text{KL}}$ —, its one-step variant  $\mathcal{L}_{1\text{-step}}$ , and sequence prediction  $\mathcal{L}_{\text{seq}}$ , introduced in Section 3. Refer to Appendix F for additional details.

**MNIST Multi** We present the results on the MNIST Multi variants in Table 3 (a). On all three variants and according to both metrics, the proposed multiset loss function outperforms all the others. The reinforcement learning based approach closely follows behind. Its performance, however, drops as the number of items in a target multiset increases. This is understandable, as the variance of policy gradient grows as the length of an episode grows. A similar behaviour was observed with sequence prediction as well as aggregate distribution matching. We were not able to train any decent models with the one-step variant of aggregate distribution matching. This was true especially in terms of exact match (EM), which we attribute to the one-step variant not being capable of modelling dependencies among the predicted items.

<sup>3</sup>The random rank function is generated before training and held fixed. We verified that generating a new random rank function for each batch significantly decreased performance.

**MS COCO** Similar to the results on the variants of MNIST Multi, the proposed multiset loss function matches or outperforms all the others on the two variants of MS COCO, as presented in Table 3 (b). On COCO Easy, with only two objects to predict per example, both aggregated distribution matching (with KL divergence) and the sequence loss functions are as competitive as the proposed multiset loss. The other loss functions significantly underperform these three loss functions, as they did on MNIST Multi. The performance gap between the proposed loss and the others, however, grows substantially on the more challenging COCO Medium, which has more objects per example. The proposed multiset loss outperforms the aggregated distribution matching with KL divergence by 3.7 percentage points on exact match and 4.8 on F1. This is analogous to the experiments on MNIST Multi, where the performance gap increased when moving from four to ten digits.

## 5.6 Analysis: Entropy Evolution

Recall from Remark 4 that the entropy of the oracle policy’s predictive distribution strictly decreases over time, i.e.,  $\mathcal{H}(\pi_*^{(t)}) > \mathcal{H}(\pi_*^{(t+1)})$ . This naturally follows from the fact that there is no pre-specified rank function, because the oracle policy cannot prefer any item from the others in a free label multiset. Hence, we examine here how the policy learned based on each loss function compares to the oracle policy in terms of per-step entropy. We consider the policies trained on MNIST Multi (10), where the differences among them were most clear. As shown in Fig. 1, the policy trained on MNIST Multi (10) using the proposed multiset loss closely follows the oracle policy. The entropy decreases as the predictions are made. The decreases can be interpreted as concentrating probability mass on progressively smaller free labels sets. The variance is quite small, indicating that this strategy is uniformly applied for any input.

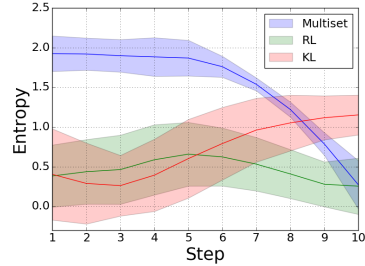


Figure 1: Comparison of per-step entropies of predictive distributions compared over the validation set.

The policy trained on MNIST Multi (10) using the proposed multiset loss closely follows the oracle policy. The entropy decreases as the predictions are made. The decreases can be interpreted as concentrating probability mass on progressively smaller free labels sets. The variance is quite small, indicating that this strategy is uniformly applied for any input.

The policy trained with reinforcement learning retains a relatively low entropy across steps, with a decreasing trend in the second half. We carefully suspect the low entropy in the earlier steps is due to the greedy nature of policy gradient. The policy receives a high reward more easily by choosing one of many possible choices in an earlier step than in a later step. This effectively discourages the policy from exploring all possible trajectories during training.

On the other hand, the policy found by aggregated distribution matching ( $\mathcal{L}_{\text{dm}}^{\text{KL}}$ ) has the opposite behaviour. The entropy in general grows as more predictions are made. To see why this is sub-optimal, consider the final step. Assuming the first nine predictions were correct, there is only one correct class left for the final prediction. The high entropy, however, indicates that the model is placing a significant amount of probability on incorrect sequences. Such a policy may result because  $\mathcal{L}_{\text{dm}}^{\text{KL}}$  cannot properly distinguish between policies with increasing and decreasing entropies. The increasing entropy also indicates that the policy has learned a rank function implicitly and is fully relying on it. We conjecture this reliance on an inferred rank function, which is by definition sub-optimal, resulted in lower performance of aggregate distribution matching.

## 6 Conclusion

We have extensively investigated the problem of multiset prediction in this paper. We rigorously defined the problem, and proposed to approach it from the perspective of sequential decision making. In doing so, an oracle policy was defined and shown to be optimal, and a new loss function, called *multiset loss*, was introduced as a means to train a parametrized policy for multiset prediction. The experiments on two families of datasets, MNIST Multi variants and MS COCO variants, have revealed the effectiveness of the proposed loss function over other loss functions including reinforcement learning, sequence, and aggregated distribution matching loss functions. This success brings in new opportunities of applying machine learning to various new domains, including high-energy physics.

## Acknowledgments

KC thanks support by eBay, TenCent, NVIDIA and CIFAR. This work was supported by Samsung Electronics (Improving Deep Learning using Latent Structure).



## References

- [1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
- [2] Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé, III, and John Langford. Learning to search better than your teacher. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37, ICML'15*, pages 2058–2066. JMLR.org, 2015.
- [3] Hal Daumé, John Langford, and Daniel Marcu. Search-based structured prediction. *Machine Learning*, 75(3):297–325, Jun 2009.
- [4] Krzysztof Dembczyński, Weiwei Cheng, and Eyke Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10*, pages 279–286, USA, 2010. Omnipress.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.
- [6] W Ehrenfeld, R Buckingham, J Cranshaw, T Cuhadar Donszelmann, T Doherty, E Gallas, J Hrivnac, D Malon, M Nowak, M Slater, F Viegas, E Vinek, Q Zhang, and the ATLAS Collaboration. Using tags to speed up the atlas analysis process. *Journal of Physics: Conference Series*, 331(3):032007, 2011.
- [7] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013.
- [8] S. Hamid Rezatofighi, Vijay Kumar B G, Anton Milan, Ehsan Abbasnejad, Anthony Dick, and Ian Reid. Deepsetnet: Predicting sets with deep neural networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. *arXiv preprint arXiv:1703.06870*, 2017.
- [10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [11] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [12] Rémi Leblond, Jean-Baptiste Alayrac, Anton Osokin, and Simon Lacoste-Julien. Searnn: Training rnns with global-local losses, 2017.
- [13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- [14] V. Lempitsky and A. Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, 2010.
- [15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.
- [16] Jinseok Nam, Eneldo Loza Mencía, Hyunwoo J Kim, and Johannes Fürnkranz. Maximizing subset accuracy with recurrent neural networks in multi-label classification. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5419–5429. Curran Associates, Inc., 2017.
- [17] Daniel Oñoro-Rubio and Roberto J. López-Sastre. Towards perspective-free object counting with deep learning. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 615–629, Cham, 2016. Springer International Publishing.
- [18] J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks*, 21(4):682–697, May 2008.
- [19] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. Classifier chains for multi-label classification. *Machine Learning*, 85(3):333, Jun 2011.

- [20] Mengye Ren and Richard S. Zemel. End-to-end instance segmentation with recurrent attention. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [21] Bernardino Romera-Paredes and Philip H. S. Torr. Recurrent instance segmentation. *arXiv preprint arXiv:1511.08250*, 2015.
- [22] Stéphane Ross, Geoffrey J Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pages 627–635, 2011.
- [23] D. Singh, A. M. Ibrahim, T. Yohanna, and J.N Singh. An overview of the applications of multisets. *Novi Sad Journal of Mathematics*, 37(3):73–92, 2007.
- [24] Russell Stewart, Mykhaylo Andriluka, and Andrew Y. Ng. End-to-end people detection in crowded scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [25] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.
- [26] Apostolos Syropoulos. *Mathematics of Multisets*, pages 347–358. Springer, Berlin, Heidelberg, 2001.
- [27] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *Int J Data Warehousing and Mining*, 2007:1–13, 2007.
- [28] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets, 2015.
- [29] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [30] Sean Welleck, Kyunghyun Cho, and Zheng Zhang. Saliency-based sequential image attention with multiset prediction. In *Advances in neural information processing systems*, 2017.
- [31] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.
- [32] SHI Xingjian, Zhouong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015.
- [33] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma. Single-image crowd counting via multi-column convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, June 2016.

## A Definitions

We review definitions of multiset and exact match, and present multiset versions of precision, recall, and F1. For a comprehensive overview of multisets, refer to [26, 23].

**Multiset** A multiset is a set that allows for multiple instances of elements. Multisets are unordered, i.e.  $\{x, x, y\}$  and  $\{x, y, x\}$  are equal. We now introduce the formal definition and convenient ways of representing a multiset.

Formally, a multiset is a pair  $\mathcal{Y} = (C, \mu)$ , where  $C = \{c_1, \dots, c_p\}$  is a *ground set*, and  $\mu : C \rightarrow \mathbb{N}_{\geq 0}$  is a *multiplicity function* that maps each  $c_i \in C$  to the number of times it occurs in the multiset. The multiset cardinality is defined as  $|\mathcal{Y}| = \sum_{c \in C} \mu(c)$ .

A multiset can be *enumerated* by numbering each element instance and representing the multiset as a size  $|\mathcal{Y}|$  set:  $\mathcal{Y} = \{c_1^{(1)}, c_1^{(2)}, \dots, c_1^{(\mu(c_1))}, c_2^{(1)}, \dots, c_2^{(\mu(c_2))}, \dots, c_p^1, \dots, c_p^{(\mu(c_p))}\}$ . This allows for notation such as  $\sum_{c \in \mathcal{Y}}$ .

An additional compact notation is  $\mathcal{Y} = \{y_1, y_2, \dots, y_{|\mathcal{Y}|}\}$ , where each  $y_i$  is an auxiliary variable referring to an underlying element  $c \in C$  of the ground set.

For instance, the multiset  $\mathcal{Y} = \{\text{cat}, \text{cat}, \text{dog}\}$  can be defined as  $\mathcal{Y} = (C, \mu)$ , where  $C = \{c_1 = \text{cat}, c_2 = \text{dog}, c_3 = \text{fish}\}$ ,  $\mu(\text{cat}) = 2, \mu(\text{dog}) = 1, \mu(\text{fish}) = 0$ , and can be written as  $\mathcal{Y} = \{c_1^{(1)} = \text{cat}, c_1^{(2)} = \text{cat}, c_2^{(1)} = \text{dog}\}$  or  $\mathcal{Y} = \{y_1 = \text{cat}, y_2 = \text{cat}, y_3 = \text{dog}\}$ .

For multiset analogues of common set operations (e.g. union, intersection, difference), and the notion of a subset, see [26, 23].

**Exact Match (EM)** Two multisets *exactly match* when their elements and multiplicities are the same. For example,  $\{x, y, x\}$  exactly matches  $\{y, x, x\}$ , while  $\{x, y, x\}$  does not exactly match  $\{z, y, z\}$  or  $\{x, y\}$ .

Formally, let  $\hat{\mathcal{Y}} = (C, \mu_{\hat{\mathcal{Y}}})$ ,  $\mathcal{Y} = (C, \mu_{\mathcal{Y}})$  be multisets over a common ground set  $C$ . Then  $\hat{\mathcal{Y}}$  and  $\mathcal{Y}$  exactly match if and only if  $\mu_{\hat{\mathcal{Y}}}(c) = \mu_{\mathcal{Y}}(c)$  for all  $c \in C$ . The evaluation metric  $\text{EM}(\hat{\mathcal{Y}}, \mathcal{Y})$  is 1 when  $\hat{\mathcal{Y}}$  and  $\mathcal{Y}$  exactly match, and 0 otherwise.

Note that exact match is the same as multiset equality, i.e.  $\hat{\mathcal{Y}} = \mathcal{Y}$ , as defined in [23].

**Precision** Precision gives the ratio of correctly predicted elements to the number of predicted elements. Specifically, let  $\hat{\mathcal{Y}} = (C, \mu_{\hat{\mathcal{Y}}})$ ,  $\mathcal{Y} = (C, \mu_{\mathcal{Y}})$  be multisets. Then

$$\text{Prec}(\hat{\mathcal{Y}}, \mathcal{Y}) = \frac{\sum_{y \in \hat{\mathcal{Y}}} I_{y \in \mathcal{Y}}}{|\hat{\mathcal{Y}}|}.$$

The summation and membership are done by **enumerating** the multiset. For example, the multisets  $\hat{\mathcal{Y}} = \{a, a, b\}$  and  $\mathcal{Y} = \{a, b\}$  are enumerated as  $\hat{\mathcal{Y}} = \{a^{(1)}, a^{(2)}, b^{(1)}\}$  and  $\mathcal{Y} = \{a^{(1)}, b^{(1)}\}$ , respectively. Then clearly  $a^{(1)} \in \mathcal{Y}$  but  $a^{(2)} \notin \mathcal{Y}$ .

Formally, precision can be defined as

$$\text{Prec}(\hat{\mathcal{Y}}, \mathcal{Y}) = 1 - \frac{\sum_{c \in C} \max(\mu_{\hat{\mathcal{Y}}}(c) - \mu_{\mathcal{Y}}(c), 0)}{|\hat{\mathcal{Y}}|}$$

where the summation is now over the ground set  $C$ . Intuitively, **precision decreases by  $\frac{1}{|\hat{\mathcal{Y}}|}$  each time an extra class label is predicted.**

**Recall** Recall gives the ratio of correctly predicted elements to the number of ground-truth elements. Recall is defined analogously to precision, as:

$$\text{Rec}(\hat{\mathcal{Y}}, \mathcal{Y}) = \frac{\sum_{y \in \hat{\mathcal{Y}}} I_{y \in \mathcal{Y}}}{|\mathcal{Y}|}.$$

Formally,

$$\text{Rec}(\hat{\mathcal{Y}}, \mathcal{Y}) = 1 - \frac{\sum_{c \in C} \max(\mu_{\mathcal{Y}}(c) - \mu_{\hat{\mathcal{Y}}}(c), 0)}{|\mathcal{Y}|}.$$

Intuitively, **recall decreases by  $\frac{1}{|\mathcal{Y}|}$  each time an element of  $\mathcal{Y}$  is not predicted.**

**F1** The F1 score is the harmonic mean of precision and recall:

$$F_1(\hat{\mathcal{Y}}, \mathcal{Y}) = 2 \cdot \frac{\text{Prec}(\hat{\mathcal{Y}}, \mathcal{Y}) \cdot \text{Rec}(\hat{\mathcal{Y}}, \mathcal{Y})}{\text{Prec}(\hat{\mathcal{Y}}, \mathcal{Y}) + \text{Rec}(\hat{\mathcal{Y}}, \mathcal{Y})}.$$

## B Proof of Remark 1

*Proof.* Note that the precision with  $\hat{y}_{<t}$  is defined as

$$\text{Prec}(\hat{y}_{<t}, \mathcal{Y}) = \frac{\sum_{y \in \hat{y}_{<t}} I_{y \in \mathcal{Y}}}{|\hat{y}_{<t}|}.$$

Because  $\hat{y} \sim \pi_*(\hat{y}_{<t}, x, \mathcal{Y}_t) \in \mathcal{Y}_t$ ,

$$\text{Prec}(\hat{y}_{\leq t}, \mathcal{Y}) = \frac{1 + \sum_{y \in \hat{y}_{<t}} I_{y \in \mathcal{Y}}}{1 + |\hat{y}_{<t}|}.$$

Then,

$$\text{Prec}(\hat{y}_{\leq t}, \mathcal{Y}) - \text{Prec}(\hat{y}_{<t}, \mathcal{Y}) = \frac{1 - \text{Prec}(\hat{y}_{<t}, \mathcal{Y})}{1 + |\hat{y}_{<t}|} \geq 0,$$

because  $0 \leq \text{Prec}(\hat{y}_{<t}, \mathcal{Y}) \leq 1$  and  $|\hat{y}_{<t}| \geq 0$ . The equality holds when  $\text{Prec}(\hat{y}_{<t}, \mathcal{Y}) = 1$ .

Similarly, we start with the definition of the recall:

$$\text{Rec}(\hat{y}_{<t}, \mathcal{Y}) = \frac{\sum_{y \in \hat{y}_{<t}} I_{y \in \mathcal{Y}}}{|\mathcal{Y}|}.$$

Because  $\hat{y} \sim \pi_*(\hat{y}_{<t}, x, \mathcal{Y}_t) \in \mathcal{Y}_t$ ,

$$\text{Rec}(\hat{y}_{\leq t}, \mathcal{Y}) = \frac{1 + \sum_{y \in \hat{y}_{<t}} I_{y \in \mathcal{Y}}}{|\mathcal{Y}|}.$$

Since the denominator is identical,

$$\text{Rec}(\hat{y}_{\leq t}, \mathcal{Y}) - \text{Rec}(\hat{y}_{<t}, \mathcal{Y}) = \frac{1}{|\mathcal{Y}|} \geq 0.$$

□

## C Proof of Remark 2

*Proof.* When  $t = 1$ ,

$$\text{Prec}(\hat{y}_{\leq 1}, \mathcal{Y}) = 1,$$

because  $\hat{y}_1 \sim \pi_*(\emptyset, x, \mathcal{Y}_1) \in \mathcal{Y}$ . From Remark 1, we know that

$$\text{Prec}(\hat{y}_{\leq t}, \mathcal{Y}) = \text{Prec}(\hat{y}_{<t}, \mathcal{Y}),$$

when  $\text{Prec}(\hat{y}_{<t}, \mathcal{Y}) = 1$ . By induction,  $\text{Prec}(\hat{y}_{\leq |\mathcal{Y}|}, \mathcal{Y}) = 1$ .

From the proof of Remark 1, we know that the recall increases by  $\frac{1}{|\mathcal{Y}|}$  each time, and we also know that

$$\text{Rec}(\hat{y}_{\leq 1}, \mathcal{Y}) = \frac{1}{|\mathcal{Y}|},$$

when  $t = 1$ . After  $|\mathcal{Y}| - 1$  steps of executing the oracle policy, the recall becomes

$$\text{Rec}(\hat{y}_{\leq |\mathcal{Y}|}, \mathcal{Y}) = \frac{1}{|\mathcal{Y}|} + \sum_{t'=2}^{|\mathcal{Y}|} \frac{1}{|\mathcal{Y}|} = 1.$$

□

## D Proof of Remark 4

*Proof.* Given a multiset  $\mathcal{Y}$  with  $|\mathcal{Y}| \leq M$ , define  $\mathbf{C} = \{c_i^{(m)} | 1 \leq i \leq |\mathcal{C}|, 1 \leq m \leq M\}$ , where  $c_i^{(m)}$  is interpreted as the  $m$ 'th instance of class  $c_i$ . Writing  $\mathcal{Y}$  in enumerated form it is clear that  $\mathcal{Y} \subset \mathbf{C}$ . Let  $t$  range from 1 to  $|\mathcal{Y}|$  and define  $\mathcal{Y}_t$  as in Definition 1.

Now, define the oracle policy as a distribution over  $\mathbf{C}$ , according to Definition 2:

$$\pi_*^{(t)}(y = c_i^{(m)} | \hat{y}_{<t}, x, \mathcal{Y}_t) = \begin{cases} \frac{1}{|\mathcal{Y}_t|}, & \text{if } c_i^{(m)} \in \mathcal{Y}_t \\ 0, & \text{otherwise} \end{cases}.$$

Therefore,

$$\begin{aligned}
\mathcal{H}(\pi_*^{(t)}) &= - \sum_{i=1}^{|C|} \sum_{m=1}^M \pi_*^{(t)}(y = c_i^{(m)}) \log \pi_*^{(t)}(y = c_i^{(m)}) \\
&= - \sum_{c \in \mathcal{Y}_t} \frac{1}{|\mathcal{Y}_t|} \log \frac{1}{|\mathcal{Y}_t|} \\
&= \frac{1}{|\mathcal{Y}_t|} \sum_{c \in \mathcal{Y}_t} \log |\mathcal{Y}_t| \\
&= \log |\mathcal{Y}_t|
\end{aligned}$$

where  $0 \log 0$  is defined as 0 in the first step.

Now, observe that  $|\mathcal{Y}_t| > |\mathcal{Y}_{t+1}|$  since  $\hat{y}_t \sim \pi_*^{(t)}$  is in  $\mathcal{Y}_t$  with probability 1 and  $\mathcal{Y}_{t+1} \leftarrow \mathcal{Y}_t \setminus \{\hat{y}_t\}$  by definition. Hence

$$\mathcal{H}(\pi_*^{(t)}) = \log |\mathcal{Y}_t| > \log |\mathcal{Y}_{t+1}| = \mathcal{H}(\pi_*^{(t+1)}).$$

□

## E Model Descriptions

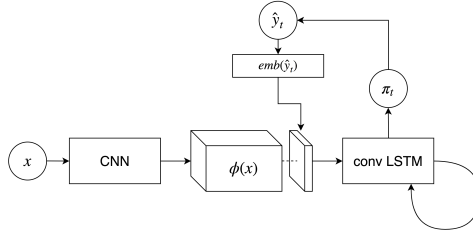


Figure 2: Graphical illustration of a predictor used throughout the experiments.

**Model** An input  $x$  is first processed by a tower of convolutional layers, resulting in a feature volume of size  $w' \times h'$  with  $d$  feature maps, i.e.,  $H = \phi(x) \in \mathbb{R}^{w' \times h' \times d}$ . At each time step  $t$ , we resize the previous prediction’s embedding  $\text{emb}(\hat{y}_{t-1}) \in \mathbb{R}^{(w') \times (h')}$  to be a  $w' \times h'$  tensor and concatenate it with  $H$ , resulting in  $\tilde{H} \in \mathbb{R}^{w' \times h' \times (d+1)}$ . This feature volume is then fed into a stack of convolutional LSTM layers. The output from the final convolutional LSTM layer  $C \in \mathbb{R}^{w' \times h' \times q}$  is spatially average-pooled, i.e.,  $c = \frac{1}{w'h'} \sum_{i=1}^{w'} \sum_{j=1}^{h'} C_{i,j,\cdot} \in \mathbb{R}^q$ . This feature vector  $c$  is then turned into a conditional distribution over the next item after affine transformation followed by a softmax function. When the one-step variant of aggregated distribution matching is used, we skip the convolutional LSTM layers, i.e.,  $c = \frac{1}{w'h'} \sum_{i=1}^{w'} \sum_{j=1}^{h'} H_{i,j,\cdot} \in \mathbb{R}^d$ .

See Fig. 2 for the graphical illustration of the entire network. See Table 4 for the details of the network for each dataset.

**Preprocessing** For MNIST Multi, we do not preprocess the input at all. In the case of MS COCO, input images are of different sizes. Each image is first resized so that its larger dimension has 600 pixels, then along its other dimension is zero-padded to 600 pixels and centered, resulting in a 600x600 image.

**Training** The model is trained end-to-end, except ResNet-34 which remains fixed after being pretrained on ImageNet. For all the experiments, we train a neural network using Adam [11] with a fixed learning rate of 0.001,  $\beta$  of (0.9, 0.999) and  $\epsilon$  of 1e-8. The learning rate was selected based on the validation performance during the preliminary experiments, and the other parameters are the default values. For MNIST Multi, the batch size was 64, and for COCO was 32. For the selection strategy experiments, 5 runs with different random seeds were used.

Table 4: Network Architectures

Data	MNIST Multi	MS COCO
CNN	conv $5 \times 5$ feat 10 max-pool $2 \times 2$ conv $5 \times 5$ feat 10 max-pool $2 \times 2$ conv $5 \times 5$ feat 32 max-pool $2 \times 2$	ResNet-34
emb( $\hat{y}_{t-1}$ )	81	361
ConvLSTM	conv $3 \times 3$ feat 32 conv $3 \times 3$ feat 32	conv $3 \times 3$ feat 512 conv $3 \times 3$ feat 512

**Feedforward Alternative** While we use a recurrent model in the experiments, the multiset loss can be used with a feedforward model as follows. A key use of the recurrent hidden state is to retain the previously predicted labels, i.e. to remember the full conditioning set  $\hat{y}_1, \dots, \hat{y}_{t-1}$  in  $p(y_t | \hat{y}_1, \dots, \hat{y}_{t-1})$ . Therefore, the proposed loss can be used in a feedforward model by encoding  $\hat{y}_1, \dots, \hat{y}_{t-1}$  in the input  $x_t$ , and running the feedforward model for  $|\hat{\mathcal{Y}}|$  steps, where  $|\hat{\mathcal{Y}}|$  is determined using a termination policy or the Special Class method detailed below. Note that compared to the recurrent model, this approach involves additional feature engineering.

**Termination Policy Alternative: Special Class** An alternative strategy to support predicting variable-sized multisets is to introduce a special item to the class set, called  $\langle \text{END} \rangle$ , and add it to the final free label multiset  $\mathcal{Y}_{|\mathcal{Y}|+1} = \{\langle \text{END} \rangle\}$ . Thus, the parametrized policy is trained to predict this special item  $\langle \text{END} \rangle$  once all the items in the target multiset have been predicted. This is analogous to NLP sequence models which predict an end of sentence token [25, 1], and was used in [30] to predict variable-sized multisets.

## F Additional Experimental Details

### F.1 Baseline Loss Functions

#### F.1.1 $\mathcal{L}_{1\text{-step}}$

The corresponding loss function for the one-step distribution matching baseline introduced in 3.1.1,  $\mathcal{L}_{1\text{-step}}$ , is:

$$\mathcal{L}_{1\text{-step}}(x, \mathcal{Y}, \theta) = \sum_{c \in C} q_*(c|x) \log q_\theta(c|x) + \lambda(\hat{l}_\theta(x) - |\mathcal{Y}|)^2,$$

where  $\lambda > 0$  is a coefficient for balancing the contributions from the two terms.

#### F.1.2 $\mathcal{L}_{\text{seq}}$

First define a rank function  $r$  that maps from one of the unique items in the class set  $c \in C$  to a unique integer. That is,  $r : C \rightarrow \mathbb{Z}$ . This function assigns the rank of each item and is used to order items  $y_i$  in a target multiset  $\mathcal{Y}$ . This results in a sequence  $\mathcal{S} = (s_1, \dots, s_{|\mathcal{Y}|})$ , where  $r(s_i) \geq r(s_j)$  for all  $j > i$ , and  $s_i \in \mathcal{Y}$ .

With this target sequence  $\mathcal{S}$  created from  $\mathcal{Y}$  using the rank function  $r$ , the sequence loss function is defined as

$$\mathcal{L}_{\text{seq}}(x, \mathcal{S}, \theta) = - \sum_{t=1}^{|\mathcal{S}|} \log \pi_\theta(s_t | s_{<t}, x).$$

Minimizing this loss function is equivalent to maximizing the conditional log-probability of the sequence  $\mathcal{S}$  given  $x$ .

### F.1.3 $\mathcal{L}_{\text{dm}}$

In distribution matching, we consider the target multiset  $\mathcal{Y}$  as a set of samples from a single, underlying distribution  $q^*$  over the class set  $C$ . This underlying distribution can be empirically estimated by counting the number of occurrences of each item  $c \in C$  in  $\mathcal{Y}$ . That is,

$$q_*(c|x) = \frac{1}{|\mathcal{Y}|} \sum_{y \in \mathcal{Y}} I_{y=c},$$

where  $I$  is the indicator function.

Similarly, we can construct an aggregated distribution computed by the parametrized policy, here denoted as  $q_\theta(c|x)$ . To do so, the policy predicts  $(y_1, \dots, y_{|\mathcal{Y}|})$  by sampling from a predicted distribution  $q_\theta^{(t)}(y_t|y_{<t}, x)$  at each step  $t$ . The per-step distributions  $q_\theta^{(t)}$  are then averaged to form the aggregate distribution  $q_\theta$ .

Learning is equivalent to minimizing a divergence between  $q_*$  and  $q_\theta$ . The  $\mathcal{L}_{\text{dm}}^p$  baseline uses

$$\mathcal{L}_{\text{dm}}^p(x, \mathcal{Y}, \theta) = \|q_* - q_\theta\|_p,$$

where  $q_*$  and  $q$  are the vectors representing the corresponding categorical distributions, and  $p = 1$  in the experiments. The  $\mathcal{L}_{\text{dm}}^{\text{KL}}$  baseline uses KL divergence:

$$\mathcal{L}_{\text{dm}}^{\text{KL}}(x, \mathcal{Y}, \theta) = - \sum_{c \in C} q_*(c|x) \log q_\theta(c|x).$$

### F.1.4 $\mathcal{L}_{\text{RL}}$

Instead of assuming the existence of an oracle policy, this approach solely relies on a reward function  $r$  designed specifically for multiset prediction. The reward function is defined as

$$r(\hat{y}_t, \mathcal{Y}_t) = \begin{cases} 1, & \text{if } \hat{y}_t \in \mathcal{Y}_t \\ -1, & \text{otherwise} \end{cases}$$

The goal is then to maximize the sum of rewards over a trajectory of predictions from a parametrized policy  $\pi_\theta$ . The final loss function is

$$\mathcal{L}_{\text{RL}} = -\mathbb{E}_{\hat{y} \sim \pi_\theta} \left[ \sum_{t=1}^T r(\hat{y}_{<t}, \mathcal{Y}_t) - \lambda \mathcal{H}(\pi_\theta(\hat{y}_{<t}, x)) \right] \quad (1)$$

where the second term inside the expectation is the negative entropy multiplied with a regularization coefficient  $\lambda$ . The second term encourages exploration during training. As in [30], we use REINFORCE [31] to stochastically minimize the loss function above with respect to  $\pi_\theta$ . This loss function is optimal in that the return, i.e., the sum of the step-wise rewards, is maximized when both the precision and recall are maximal (= 1).

## F.2 Input-Dependent Rank Function

For the  $\mathcal{L}_{\text{seq}}$  baseline, a domain-specific, input-dependent rank function can be defined to transform the target multiset into a sequence. A representative example is an image input with bounding box annotations. Here, we present two input-dependent rank functions in such a case.

First, a spatial rank function  $r_{\text{spatial}}$  assigns an integer rank to each item in a given target multiset  $\mathcal{Y}$  such that

$$r_{\text{spatial}}(y_i|x) < r_{\text{spatial}}(y_j|x), \\ \text{if } \text{pos}_x(x_i) < \text{pos}_x(x_j) \text{ and } \text{pos}_y(x_i) < \text{pos}_y(x_j),$$

where  $x_i$  and  $x_j$  are the objects corresponding to the items  $y_i$  and  $y_j$ .

Second, an area rank function  $r_{\text{area}}$  decides the rank of each label in a target multiset according to the size of the corresponding object inside the input image:

$$r_{\text{area}}(y_i|x) < r_{\text{area}}(y_j|x), \text{ if } \text{area}(x_i) < \text{area}(x_j).$$

The area may be determined based on the size of a bounding box or the number of pixels, depending on the level of annotation.