DIVIDE AND CONQUER NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the learning of algorithmic tasks by mere observation of input-output pairs. Rather than studying this as a black-box discrete regression problem with no assumption whatsoever on the input-output mapping, we concentrate on tasks that are amenable to the principle of *divide and conquer*, and study what are its implications in terms of learning.

This principle creates a powerful inductive bias that we leverage with neural architectures that are defined recursively and dynamically, by learning two scale-invariant atomic operations: how to *split* a given input into smaller sets, and how to *merge* two partially solved tasks into a larger partial solution. Our model can be trained in weakly supervised environments, namely by just observing input-output pairs, and in even weaker environments, using a non-differentiable reward signal. Moreover, thanks to the dynamic aspect of our architecture, we can incorporate the computational complexity as a regularization term that can be optimized by backpropagation. We demonstrate the flexibility and efficiency of the Divide-and-Conquer Network on two combinatorial and geometric tasks: convex hull, clustering and euclidean TSP. Thanks to the dynamic programming nature of our model, we show significant improvements in terms of generalization error and computational complexity.

1 INTRODUCTION

Algorithmic tasks can be described as discrete input-output mappings defined over variable-sized inputs, but this "black-box" vision hides all the fundamental questions that explain how the task can be optimally solved and generalized to arbitrary inputs. Indeed, many tasks have some degree of scale invariance or self-similarity, meaning that there is a mechanism to solve it that is somehow independent of the input size. This principle is the basis of recursive solutions and dynamic programming, and is ubiquitous in most areas of discrete mathematics, from geometry to graph theory. In the case of images and audio signals, invariance principles are also critical for success: CNNs exploit both translation invariance and scale separation with multilayer, localized convolutional operators. In our scenario of discrete algorithmic tasks, we build our model on the principle of *divide and conquer*, which provides us with a form of parameter sharing across scales akin to that of CNNs across space or RNNs across time.

Whereas CNN and RNN models define algorithms with linear complexity, attention mechanisms Bahdanau et al. (2014) generally correspond to quadratic complexity, with notable exceptions Andrychowicz & Kurach (2016). This can result in a mismatch between the intrinsic complexity required to solve a given task and the complexity that is given to the neural network to solve it, which may impact its generalization performance. Our motivation is that learning cannot be 'complete' until these complexities match, and we start this quest by first focusing on problems for which the intrinsic complexity is well known and understood.

Our Divide-and-Conquer Networks (DCN) contain two modules: a *split* phase that is applied recursively and dynamically to the input in a coarse-to-fine way to create a hierarchical partition encoded as a binary tree; and a *merge* phase that traces back that binary tree in a fine-to-coarse way by progressively combining partial solutions; see Figure 1. Each of these phases is parametrized by a single neural network that is applied recursively at each node of the tree, enabling parameter sharing across different scales and leading to good sample complexity and generalisation.

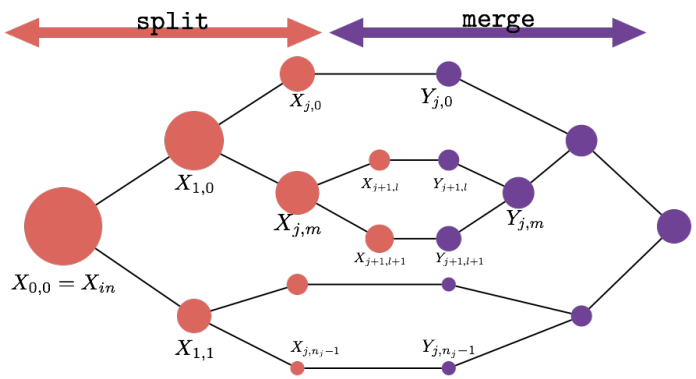

Figure 1: Divide and Conquer Network. The split phase is determined by a dynamic neural network $\mathcal{S}_\theta$ that splits each incoming set into two disjoint sets: $\{X_{j+1,l}, X_{j+1,l+1}\} = \mathcal{S}_\theta(X_{j,m})$, with $X_{j,m} = X_{j+1,l} \sqcup X_{j+1,l+1}$. The merge phase is carried out by another neural network $\mathcal{M}_\phi$ that combines two partial solutions into a solution of the coarser scale: $Y_{j,m} = \mathcal{M}_\phi(Y_{j+1,l}, Y_{j+1,l+1})$; see Section 3 for more details.

In this paper, we attempt to incorporate the scale-invariance prior with the desiderata to only require weak supervision. In particular, we consider two setups: learning from input-output pairs, and learning from a non-differentiable reward signal. Since our split block is inherently discrete, we resort to policy gradient to train the split parameters, while using standard backpropagation for the merge phase; see Section 5. An important benefit of our framework is that the architecture is dynamically determined, which suggests using the computational complexity as a regularization term. As shown in the experiments, computational complexity is a good proxy for generalisation error in the context of discrete algorithmic tasks. We demonstrate our model on algorithmic and geometric tasks with some degree of scale self-similarity: planar convex-hull, k-means and euclidean TSP. Our numerical results on these tasks reaffirm the fact that whenever the structure of the problem has scale invariance, then exploiting it leads to improved generalization and computational complexity over previous approaches.

2 Related Work

Using neural networks to solve algorithmic tasks is an active area of current research, but its models can be traced back to context free grammars Fanty (1994). In particular, dynamic learning appears in works such as Pollack (1991) and Tabor (2000). The current research in the area is dominated by RNNs Joulin & Mikolov (2015); Grefenstette et al. (2015), LSTMs Hochreiter & Schmidhuber (1997), sequence-to-sequence neural models Sutskever et al. (2014); Zaremba & Sutskever (2014), attention mechanisms Vinyals et al. (2015b); Andrychowicz & Kurach (2016) and explicit external memory models Weston et al. (2014); Sukhbaatar et al. (2015); Graves et al. (2014); Zaremba & Sutskever (2015). We refer the reader to Joulin & Mikolov (2015) and references therein for a more exhaustive and detailed account of related work.

Amongst these works, we highlight some that are particularly relevant to us. Neural GPU Kaiser & Sutskever (2015) defines a neural architecture that acts convolutionally with respect to the input and is applied iteratively $o(n)$ times, where n is the input size. It leads to fixed computational machines with total $\Theta(n^2)$ complexity. Neural Programmer-Interpreters Reed & de Freitas (2015) introduce a compositional model based on a LSTM that can learn generic programs. It is trained with full supervision using execution traces. Directly related, Cai et al. (2017) incorporates recursion into the NPI to enhance its capacity and provide learning certificates in the setup where recursive execution traces are available for supervision. Hierarchical attention mechanisms have been explored in Andrychowicz & Kurach (2016). They improve the complexity of the model from $o(n^2)$ of traditional attention to $o(n \log n)$, similarly as our models. Finally, Pointer Networks Vinyals et al. (2015b;a) modify classic attention mechanisms to make them amenable to adapt to variable input-dependent outputs, and illustrate the resulting models on geometric algorithmic tasks. They belong to the $\Theta(n^2)$ category class.

3 PROBLEM SETUP

3.1 SCALE INVARIANT TASKS

We consider tasks consisting in a mapping $\mathcal{T}$ between a variable-sized input set $X = \{x_1, \dots, x_n\}$, $x_j \in \mathcal{X}$ into an ordered set $Y = \{y_1, \dots, y_{m(n)}\}$, $y_j \in \mathcal{Y}$. This setup includes problems where the output size $m(n)$ differs from the input size n, and also problems where Y is a labeling of input elements. In particular, we will study in detail the case where $Y \subseteq X$ (and in particular $\mathcal{Y} \subseteq \mathcal{X}$).

We are interested in tasks that are self-similar across scales, meaning that if we consider the recursive decomposition of $\mathcal{T}$ as

$$\begin{aligned} \forall\, n\,, \forall X\,, |X| = n\,,\ \mathcal{T}(X) &= \mathcal{M}(\mathcal{T}(\mathcal{S}_1(X)), \dots, \mathcal{T}(\mathcal{S}_s(X)))\,, \\ |\mathcal{S}_j(X)| &< n\,,\ \cup_{j \leq s} \mathcal{S}_j(X) = X\,, \end{aligned} \tag{1}$$

where $\mathcal{S}$ splits the input into smaller sets, and $\mathcal{M}$ merges the solved corresponding sub-problems, then both $\mathcal{M}$ and $\mathcal{S}$ are significantly easier to approximate with data-driven models. In other words, the solution of the task for a certain size n is easier to estimate as a function of the partial solutions $\mathcal{T}(\mathcal{S}_j(X))$ than directly from the input. Under this assumption, the task $\mathcal{T}$ can thus be solved by first splitting the input into s strictly smaller subsets $\mathcal{S}_j(X)$, solving $\mathcal{T}$ on each of these subsets, and then appropriately merging the corresponding outputs together. In order words, $\mathcal{T}$ can be solved by recursion. A particularly simple and illustrative case is the binary setup with $s = 2$ and $\mathcal{S}_1(X) \cap \mathcal{S}_2(X) = \emptyset$, that we will adopt in the following for simplicity.

3.2 WEAKLY SUPERVISED RECURSION

Our first goal is to learn how to perform $\mathcal{T}$ for any size n, by observing only input-output example pairs (X^l, Y^l), $l = 1 \dots L$. Throughout this work, we will make the simplifying assumption of binary splitting ($s = 2$), although our framework extends naturally to more general versions. Given an input set X associated with output Y, we first define a split phase that breaks X into a disjoint partition tree $\mathcal{P}(X)$:

$$\mathcal{P}(X) = \{X_{j,k}\,; 0 \leq j < J; 0 \leq k < n_j\}\,, \text{ with } X_{j,k} = X_{j+1,2k} \sqcup X_{j+1,2k+1}\,, \tag{2}$$

and $X = X_{1,0} \sqcup X_{1,1}$. This partition tree is obtained by recursively applying a trainable binary split module $\mathcal{S}_\theta$:

$$\begin{aligned} \{X_{1,0}, X_{1,1}\} &= \mathcal{S}_\theta(X)\,, \text{ with } X = X_{1,0} \sqcup X_{1,1}\,, \\ \{X_{j+1,2k}, X_{j+1,2k+1}\} &= \mathcal{S}_\theta(X_{j,k})\,, \text{ with } X_{j,k} = X_{j+1,2k} \sqcup X_{j+1,2k+1}\,,\ (j < J, k \leq 2^j)\,. \end{aligned} \tag{3}$$

Here, J indicates the number of *scales* or depth of recursion that our model applies for a given input X, and $\mathcal{S}_\theta$ is a neural network that takes a set as input and produces a binary, disjoint partition as output. Eq. (3) thus defines a hierarchical partitioning of the input that can be visualized as a binary tree; see Figure 1. This binary tree is data-dependent and will therefore vary for each input example, dictated by the current choice of parameters for $\mathcal{S}_\theta$.

The second phase of the model takes as input the binary tree partition $\mathcal{P}(X)$ determined by the split phase and produces an estimate $\hat{Y}$. We traverse upwards the dynamic computation tree determined by the split phase using a second trainable block, the merge module $\mathcal{M}_\phi$:

$$\begin{aligned} Y_{J,k} &= \tilde{\mathcal{M}}_\phi(X_{J,k})\ ,\ (1 \leq k \leq 2^J)\,, \\ Y_{j,k} &= \mathcal{M}_\phi(Y_{j+1,2k}, Y_{j+1,2k+1})\ ,\ (1 \leq k \leq 2^j, j < J)\,, \text{and } \hat{Y} = \mathcal{M}_\phi(Y_{1,0}, Y_{1,1})\,. \end{aligned} \tag{4}$$

Here we have denoted by $\tilde{\mathcal{M}}$ the atomic block that transforms inputs at the leaves of the split tree, and $\mathcal{M}_\phi$ is a neural network that takes as input two (possibly ordered) inputs and merges them into another (possibly ordered) output. In the setup where $Y \subseteq X$, we further impose that $Y_{j,k} \subseteq Y_{j+1,2k} \cup Y_{j+1,2k+1}$, to guarantee that the computation load does not diverge with J.

3.3 LEARNING FROM NON-DIFFERENTIABLE REWARDS

Another setup we can address with (1) consists in problems where one can assign a cost (or reward) to a given partitioning of an input set. In that case, Y encodes the labels assigned to each input element.

We also assume that the reward function has some form of self-similarity, in the sense that one can relate the reward associated to subsets of the input to the total reward.

In that case, (3) is used to map an input X to a partition $\mathcal{P}(X)$, determined by the leaves of the tree $\{X_{J,k}\}_k$, that is evaluated by an external black-box returning a cost $\mathcal{L}(\mathcal{P}(X))$. For instance, one may wish to perform graph coloring satisfying a number of constraints. In that case, the cost function would assign $\mathcal{L}(\mathcal{P}(X)) = 0$ if $\mathcal{P}(X)$ satisfies the constraints, and $\mathcal{L}(\mathcal{P}(X)) = |X|$ otherwise.

In its basic form, since $\mathcal{P}(X)$ belongs to a discrete space of set partitions of size super-exponential in $|X|$ and the cost is non-differentiable, optimizing $\mathcal{L}(\mathcal{P}(X))$ over the partitions of X is in general intractable. However, for tasks with some degree of self-similarity, one can expect that the combinatorial explosion can be avoided. Indeed, if the cost function $\mathcal{L}$ is *subadditive*, i.e.,

$$\mathcal{L}(\mathcal{P}(X)) \leq \mathcal{L}(\mathcal{P}(X_{1,0})) + \mathcal{L}(\mathcal{P}(X_{1,1})) \ , \text{ with } \mathcal{P}(X) = \mathcal{P}(X_{1,0}) \sqcup \mathcal{P}(X_{1,1}) \ ,$$

then the hierarchical splitting from (3) can be used as an efficient greedy strategy, since the right hand side acts as a surrogate upper bound that depends only on smaller sets. In our case, since the split phase is determined by a single block $\mathcal{S}_\theta$ that is recursively applied, this setup can be cast as a simple fixed-horizon (J steps) Markov Decision Process, that can be trained with standard policy gradient methods; see Section 5.

3.4 Computational Complexity as Regularization

Besides the prospect of better generalization, the recursion (1) also enables the notion of computational complexity regularization. Indeed, in tasks that are scale invariant the decomposition in terms of $\mathcal{M}$ and $\mathcal{S}$ is not unique in general. For example, in the sorting task with n input elements, one may select the largest element of the array and query the sorting task on the remaining $n-1$ elements, but one can also attempt to break the input set into two subsets of similar size using a pivot, and query the sorting on each of the two subsets. Both cases reveal the scale invariance of the problem, but the latter leads to optimal computational complexity ($\Theta(n \log n)$) whereas the former does not ($\Theta(n^2)$). Therefore, in a trainable divide-and-conquer architecture, one can regularize the split operation to minimize computational complexity; see Appendix A.

4 Neural Models for $\mathcal{S}$ and $\mathcal{M}$

4.1 Split

The split block $\mathcal{S}_\theta$ receives as input a variable-sized set $X = (x_1, \ldots, x_n)$ and produces a binary partition $X = X_0 \sqcup X_1$. We encode such partition with binary labels $z_1 \ldots z_n$, $z_m \in \{0, 1\}$, $m \leq n$. These labels are sampled from probabilities $p_\theta(z_m = 1 \mid X)$ that we now describe how to parametrize. Since the model is defined over sets, we use an architecture that certifies that $p_\theta(z_m = 1 \mid X)$ are invariant by permutation of the input elements. The *Set2set* model Vinyals et al. (2015a) constructs a nonlinear set representation by cascading R layers of

$$h_m^{(1)} = \rho\left(B_{1,0} x_m + B_{2,0}\mu(X)\right) \ , \ h_m^{(r+1)} = \rho\left(B_{1,r} h_m^{(r)} + n^{-1} B_{2,r} \sum_{m' \leq n} h_{m'}^{(r)}\right) \ , \tag{5}$$

with $m \leq n$, $r \leq R$, $h_m^{(r)} \in \mathbb{R}^d$, and $p_\theta(z_m = 1 \mid X) = \text{Sigm}(b^T h_m^{(R)})$. The parameters of $\mathcal{S}_\theta$ are thus $\theta = \{B_0, B_{1,r}, B_{2,r}, b\}$. In order to avoid covariate shifts given by varying input set distributions and sizes, we consider a normalization of the input that standardizes the input variables x_j and feeds the mean and variance $\mu(X) = (\mu_0, \sigma)$ to the first layer. If the input has some structure, for instance X is the set of vertices of a graph, a simple generalization of the above model is to estimate a graph structure specific to each layer:

$$h_m^{(1)} = \rho\left(B_{1,0} x_m + B_{2,0}\mu(X)\right) \ , \ h_m^{(r+1)} = \rho\left(B_{1,r} h_m^{(r)} + n^{-1} B_{2,r} \sum_{m' \leq n} A_{m,m'}^{(r)} h_{m'}^{(r)}\right) \ , \tag{6}$$

where $A^{(r)}$ is a similarity kernel computed as a symmetric function of current hidden variables: $A_{m,m'}^{(r)} = \varphi(h_m^{(r)}, h_{m'}^{(r)})$. This corresponds to the so-called graph neural networks Scarselli et al. (2009); Duvenaud et al. (2015) or neural message passing Gilmer et al. (2017).

Finally, the binary partition tree $\mathcal{P}(X)$ is constructed recursively by first computing $p_\theta(z \mid X)$, then sampling from the corresponding distributions to obtain $X = X_0 \sqcup X_1$, and then applying $\mathcal{S}_\theta$ recursively on X_0 and X_1 until the partition tree leaves have size smaller than a predetermined constant, or the number of scales reaches a maximum value J. We denote the resulting distribution over tree partitions by $\mathcal{P}(X) \sim \mathbf{S}_\theta(X)$.

4.2 MERGE

4.2.1 MERGE MODULE

The merge block $\mathcal{M}_\phi$ takes as input a pair of sequences Y_0, Y_1 and produces an output sequence O. Motivated by our applications, we describe first an architecture for this module in the setup where the output sequence is indexed by elements from the input sequences, although our framework can be extended to more general setups seamlessly. in Section 6.3.

Given an input sequence Y, the merge module computes a stochastic matrix Γ_Y (where each row is a probability distribution) such that the output O is expressed by binarizing its entries and multiplying it by the input:

$$O = \mathcal{M}_\phi(Y_0, Y_1) = \bar{\Gamma} \begin{pmatrix} Y_0 \\ Y_1 \end{pmatrix} \text{ , with } \bar{\Gamma}_{s,i} = \begin{cases} 1 & \text{if } i = \arg\max_{i'} p_s(i') \text{ .} \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Since we are interested in weakly supervised tasks, the target output only exists at the coarsest scale of the partition tree. We thus also consider a generative version $\mathcal{M}^g_\phi$ of the merge block that uses its own predictions in order to sample an output sequence. The initial merge operation at the finest scale $\tilde{\mathcal{M}}$ is defined as the previous merge module applied to the input $(X_{J,k}, \emptyset)$. This merge module operation can be instantiated with Pointer Networks Vinyals et al. (2015b) and with Graph Neural Networks/ Neural Message Passing Gilmer et al. (2017); Kearnes et al. (2016); Bronstein et al. (2016).

Pointer Networks We consider a Pointer Network (PtrNet) Vinyals et al. (2015b) to our input-output interface as our merge block $\mathcal{M}_\phi$. A PtrNet is an auto-regressive model for tasks where the output sequence is a permutation of a subsequence of the input. The model encodes each input sequence $Y_q = (x_{1,q}, \ldots, x_{n_q,q})$, $q = 0, 1$, into a global representation $e_q := e_{q,n_q}$, $q = 0, 1$, by sequentially computing $e_{1,q}, \ldots, e_{n_q,q}$ with an RNN. Then, another RNN decodes the output sequence with initial state $d_0 = \rho(A_0 e_0 + A_1 e_1)$, as described in detail in Appendix E. The trainable parameters ϕ regroup to the RNN encoder and decoder parameters.

Graph Neural Networks Another use case of our Divide and Conquer Networks are problems formulated as paths on a graph, such as convex hulls or the travelling salesman problem. A path on a graph of n nodes can be seen as a binary signal over the $n \times n$ edge matrix. Leveraging recent incarnations of Graph Neural Networks/ Neural Message Passing that consider both node and edge hidden features Gilmer et al. (2017); Kearnes et al. (2016); Bronstein et al. (2016), the merge module can be instantiated with a GNN mapping edge-based features from a bipartite graph representing two partial solutions Y_0, Y_1, to the edge features encoding the merged solution. Specifically, we consider the following update equations:

$$V_m^{(k+1)} = \rho\left(\beta_1 V_m^{(k)} + \sum_{m'} A^{(k)}_{m,m'} V^{(k)}_{m'}\right) \tag{8}$$

$$A^{(k+1)}_{m,m'} = \varphi(V_m^{(k+1)}, V_{m'}^{(k+1)}) \text{ ,} \tag{9}$$

where φ is a symmetric, non-negative function parametrized with a neural network.

4.2.2 RECURSIVE MERGE OVER PARTITION TREE

Given a partition tree $\mathcal{P}(X) = \{X_{j,k}\}_{j,k}$, we perform a merge operation at each node (j, k). The merge operation traverses the tree in a fine-to-coarse fashion. At the leaves of the tree, the sets $X_{J,k}$ are transformed into $Y_{J,k}$ as $Y_{J,k} = \mathcal{M}^g_\phi(X_{J,k}, \emptyset)$, and, while $j > 0$, these outputs are recursively transformed along the binary tree as $Y_{j,k} = \mathcal{M}^g_\phi(Y_{j+1,2k}, Y_{j+1,2k+1})$, $0 < j < J$, using the auto-regressive version, until we reach the scale with available targets: $\hat{Y} = \mathcal{M}_\phi(Y_{1,0}, Y_{1,1})$. At test-time, without ground-truth outputs, we replace the last $\mathcal{M}_\phi$ by its generative version $\mathcal{M}^g_\phi$.

4.2.3 BOOTSTRAPPING THE MERGE PARTITION TREE

The recursive merge defined at (4.2.2) can be viewed as a factorized attention mechanism over the input partition. Indeed, the merge module outputs (20) include the stochastic matrix $\Gamma = (p_1, \ldots, p_S)$ whose rows are the p_s probability distributions over the indexes. The number of rows of this matrix is the length of the output sequence and the number of columns is the length of the input sequence. Since the merge blocks are cascaded by connecting each others outputs as inputs to the next block, given a hierarchical partition of the input $\mathcal{P}(X)$, the overall mapping can be written as

$$\hat{Y} = \left(\prod_{j=0}^{J} \widetilde{\Gamma}_j\right) \begin{bmatrix} Y_{J,0} \\ \vdots \\ Y_{J,n_J} \end{bmatrix}, \text{with } \widetilde{\Gamma}_0 = \bar{\Gamma}_{0,0}\,, \widetilde{\Gamma}_1 = \begin{pmatrix} \bar{\Gamma}_{1,0} & 0 \\ 0 & \bar{\Gamma}_{1,1} \end{pmatrix}, \widetilde{\Gamma}_J = \begin{pmatrix} \bar{\Gamma}_{J,0} & 0 & \cdots \\ 0 & \bar{\Gamma}_{J,1} & \ddots \\ 0 & \cdots & \bar{\Gamma}_{J,n_J} \end{pmatrix}. \tag{10}$$

It follows that the recursive merge over the binary tree is a specific reparametrization of the global permutation matrix, in which the permutation matrix has been decomposed into a product of permutations dictated by the binary tree, indicating our belief that many routing decisions are done locally within the original set. The model is trained with maximum likelihood using the product of the non-binarized stochastic matrices. Lastly, in order to avoid singularities we need to enforce that $\log p_{s,t_s}$ is well-defined and therefore that $p_{s,t_s} > 0$. We thus regularize the quantization step (20) by replacing $0, 1$ with $\epsilon^{1/J}, 1 - n\epsilon^{1/J}$ respectively. We also found useful to binarize the stochastic matrices at fine scales when the model is close to convergence, so gradients are only sent at coarsest scale. For simplicity, we use the notation $p_\phi(Y \mid \mathcal{P}(X)) = \prod_{j=0}^{J} \widetilde{\Gamma}_j = \mathbf{M}_\phi(\mathcal{P}(X))$, where now the matrices $\tilde{\Gamma}_j$ are not binarized.

5 TRAINING

This section describes how the model parameters $\{\theta, \phi\}$ are estimated under two different learning paradigms. Given a training set of pairs $\{(X^l, Y^l)\}_{l \leq L}$, we consider the loss

$$\mathcal{L}(\theta, \phi) = \frac{1}{L} \sum_{l \leq L} \mathbb{E}_{\mathcal{P}(X) \sim \mathbf{S}_\theta(X)} \log p_\phi(\hat{Y} = Y^l \mid \mathcal{P}(X^l))\,, \text{with } \; p_\phi(Y \mid \mathcal{P}(X)) = \mathbf{M}_\phi(\mathcal{P}(X))\,. \tag{11}$$

Section 4.2 explained how the merge phase $\mathbf{M}_\phi$ is akin to a structured attention mechanism. Equations (10) show that, thanks to the parameter sharing and despite the quantizations affecting the finest leaves of the tree, the gradient

$$\nabla_\phi \log p_{\theta,\phi}(Y \mid X) = \mathbb{E}_{\mathcal{P}(X) \sim \mathbf{S}_\theta(X)} \nabla_\phi \log \mathbf{M}_\phi(\mathcal{P}(X)) \tag{12}$$

is well-defined and non-zero almost everywhere. However, since the split parameters are separated from the targets through a series of discrete sampling steps, the same is not true for $\nabla_\theta \log p_{\theta,\phi}(Y \mid X)$. We therefore resort to the identity used extensively in policy gradient methods. For arbitrary F defined over partitions $\mathcal{X}$, and denoting by $f_\theta(\mathcal{X})$ the probability density of the random partition $\mathbf{S}_\theta(X)$, we have

$$\begin{aligned} \nabla_\theta \mathbb{E}_{\mathcal{P}(X) \sim \mathbf{S}_\theta(X)} F(\mathcal{P}(X)) &= \sum_{\mathcal{X}} F(\mathcal{X}) \nabla_\theta f_\theta(\mathcal{X}) = \sum_{\mathcal{X}} F(\mathcal{X}) f_\theta(\mathcal{X}) \nabla_\theta \log f_\theta(\mathcal{X}) = \mathbb{E}_{\mathcal{X} \sim \mathbf{S}_\theta(X)} F(\mathcal{X}) \nabla_\theta \log f_\theta(\mathcal{X}) \\ &\approx \frac{1}{S} \sum_{\tilde{X}^{(s)} \sim \mathbf{S}_\theta(X)} F(\mathcal{X}^{(s)}) \nabla_\theta \log f_\theta(\mathcal{X}^{(s)})\,. \end{aligned} \tag{13}$$

Since the split variables at each node of the tree are conditionally independent given its parent, we can compute $\log f_\theta(\mathcal{P}(X))$ as

$$\log f_\theta(\mathcal{P}(X)) = \sum_{j=1}^{J} \sum_{k \leq n_j} \sum_{m \leq |X_{j,k}|} \log p_\theta(z_{m,j,k} \mid X_{j-1,k/2})\,.$$

By plugging $F(\mathcal{P}(X)) = \log p_\phi(Y \mid \mathcal{P}(X))$ we thus obtain an efficient estimation of $\nabla_\theta \mathbb{E}_{\mathcal{P}(X) \sim \mathbf{S}_\theta(X)} \log p_\phi(Y \mid \mathcal{P}(X))$.

From (13), it is straightforward to train our model in a regime where a given partition $\mathcal{P}(X)$ of an input set is evaluated by a black-box system producing a reward $R(\mathcal{P}(X))$. Indeed, in that case, the loss becomes

$$\mathcal{L}(\theta) = \frac{-1}{L} \sum_{l \leq L} \mathbb{E}_{\mathcal{P}(X) \sim \mathbf{S}_\theta(X)} R(\mathcal{P}(X)) \ , \tag{14}$$

which can be minimized using (13) with $F(\mathcal{P}(X)) = R(\mathcal{P}(X))$.

6 EXPERIMENTS

The hyperparameters used for each experiment can be found at the Appendix. [1]

6.1 CONVEX HULL

The convex hull of a set of n points $X = \{x_1, \ldots, x_n\}$ is defined as the extremal set of points of the convex polytope with minimum area that contains them all. The planar (2d) convex hull is a well known task in discrete geometry and the optimal algorithm complexity is achieved using divide and conquer strategies by exploiting the self-similarity of the problem. The strategy for this task consists of splitting the set of points into two disjoint subsets and solving the problem recursively for each. If the partition is balanced enough, the overall complexity of the algorithm amounts to $\Theta(n \log n)$. The split phase usually takes $\Theta(n \log n)$ because each node involves a median computation to make the balanced partition property hold. The merge phase can be done in linear time on the total number of points of the two recursive solutions, which scales logarithmically with the total number of points when sampled uniformly inside a polytope Dwyer (1988).

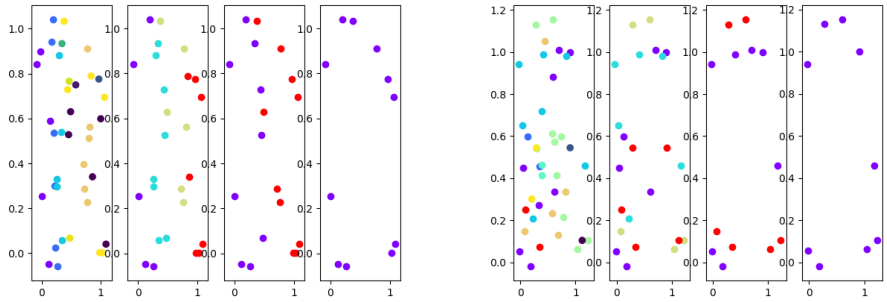

Figure 2: DCN outputs at test time for $n = 50$. The colors indicate the partitions at a given scale. Scales go fine-to-coarse from left to right. *Left:* Split has already converged using the rewards coming from the merge. It gives disjoint partitions to ease the merge work. *Right:* DCN with random split phase. Although the performance degrades due to the non-optimal split strategy, the model is able to output the correct convex hull for most of the cases.

We test the DCN on the setting consisting of n points sampled in the unit square $[0, 1]^2 \subset \mathbb{R}^2$. This is the same setup as Vinyals et al. (2015b). The training dataset has size sampled uniformly from 6 to 50. The training procedure is the following; we first train the baseline pointer network until convergence. Then, we initialize the DCN merge parameters with the baseline and train both split and merge blocks. We use this procedure in order to save computational time for the experiments, however, we observe convergence even when the DCN parameters are initialized from scratch. We supervise the merge block with the product of the continuous Γ matrices. For simplicity, instead of defining the depth of the tree dynamically depending on the average size of the partition, we fix it to 0 for 6-12, 1 for 12-25 and 2 for 25-50; see Figure 2 and Table 1.

[1] Publicly available code to reproduce all results are at `https://???.???.???`

	n=25	n=50	n=100	n=200
Baseline	81.3	65.6	41.5	13.5
DCN Random Split	59.8	37.0	23.5	10.29
DCN	88.1	83.7	73.7	52.0
DCN + Split Reg	**89.8**	**87.0**	**80.0**	**67.2**

Table 1: ConvHull test accuracy results with the baseline PtrNet and different setups of the DCN. The scale J has been set to 3 for n=100 and 4 for n=200. At row 2 we observe that when the split block is not trained we get worse performance than the baseline, however, the generalization error shrinks faster on the baseline. When both blocks are trained jointly, we clearly outperform the baseline. In Row 3 the split is only trained with REINFORCE, and row 4 when we add the computational regularization term (See Supplementary) enforcing shallower trees.

6.2 K-MEANS

We tackle the task of clustering a set of n points with the DCN in the setting described in (14). The problem consists in finding k clusters of the data with respect to the Euclidean distance. In order to make learning more relevant, we consider points in $\mathbb{R}^d$ and use as clustering metric only a subset of the coordinates, so the model needs to learn how to project the data as well. The problem reduces to solving the following combinatorial problem over input partitions $\mathcal{P}(X)$:

$$\min_{\mathcal{P}(X)} -R(\mathcal{P}(X)) = \min_{\mathcal{P}(X)} \sum_{i \in \mathcal{P}(X)} n_i \sigma_i^2 \,, \tag{15}$$

where σ_i^2 is the variance of each subset of the partition $\mathcal{P}(X)$ projected in a fixed subspace, and n_i its cardinality. We only consider the split block for this task because the combinatorial problem is over input partitions. The dataset is constructed by sampling k points in the unit square, then sampling n/k points from gaussians of variance 0.001 centered at each of the k points, and adding 2 iid uniform coordinates. The baseline is a modified version of the split block in which instead of computing binary probabilities we compute a final softmax of dimensionality k in order to produce a labelling over the input. We compare its performance with the DCN with binary splits and $\log k$ scales where we only train with the reward of the output partition at the leaves of the tree. In this case, no split regularization has been added to enforce balanced partitions; see Table 2, and Supplementary material for extended results.The subadditivity property of the cost function is exploited by the hierarchical structure and makes the problem tractable by avoiding the combinatorial explosion.

	k=4, n=40	k=8, n=80	k=16, n=160
Baseline	**0.35**	4.62	16.08
DCN	0.85	**0.86**	**0.86**

Table 2: We show the k-means cost (15) after convergence both for the baseline and the DCN. The dynamic version outperforms the baseline for larger problems, thanks to better sample complexity. The expected number of samples for the baseline to receive meaningful gradients from the cost is much larger than for the DCN.

6.3 TRAVELLING SALESMAN PROBLEM

The TSP is a prime instance of a NP-hard combinatorial optimization task. Due to its important practical applications, several powerful heuristics exist in the metric TSP case, in which edges satisfy the triangular inequality. This motivates data-driven models to either generalize those heuristics to general settings, or improve them. Data-driven approaches to the TSP can be formulated in two different ways. First, one can use both the input graph and the ground truth TSP cycle to train the model to predict the ground truth. Alternatively, one can consider only the input graph and train the model to minimize the cost of the predicted cycle. The latter is more natural since it optimizes the TSP cost directly, but the cost of the predicted cycle is not differentiable w.r.t model parameters. Some authors have successfully used reinforcement learning techniques to address this issue Dai et al. (2017), Bello et al. (2016), although the models suffer from generalization to larger problem instances.

Here we concentrate on that generalization aspect and therefore focus on the supervised setting using ground truth cycles. We compare a baseline model that used the formulation of TSP as a Quadratic Assignment Problem to develop a Graph Neural Network Nowak et al. (2017) with a DCN model that considers split and merge modules given by separate GNNs. More precisely, the split module is a GNN that receives a graph and outputs binary probabilities for each node. This module is applied recursively until a fixed scale J and the baseline is used at every final sub-graph of the partition, resulting in signals over the edges encoding possible partial solutions of the TSP. Finally, the merge module is another GNN that receives a pair of signals encoded as matrices from its leaves and returns a signal over the edges of the complete union graph.

As in Nowak et al. (2017), we generated 20k training examples and tested on 1k other instances. Each one generated by uniformly sampling $\{x_i\}_{i=1}^{20} \in [0,1]^2$. We build a complete graph with $A_{i,j} = d_{\max} - d_2(x_i, x_j)$ as weights. The ground truth cycles are generated with Helsgaun (2006), which has an efficient implementation of the Lin-Kernighan TSP Heuristic. The architecture has 20 layers and 20 feature maps per layer, and alternates between learning node and edge features. The predicted cycles are generated with a beam search strategy with beam size of 40.

	n=10			n=20			n=40			n=80		
	acc	cost	ratio	acc	cost	ratio	acc	cost	ratio	acc	cost	ratio
BS1	**78.71**	**2.89**	**1.01**	27.36	4.74	1.24	15.06	9.12	1.77	13.76	15.07	2.15
BS2	19.45	3.68	1.29	**54.34**	**4.04**	**1.05**	28.42	6.29	1.22	15.72	11.23	1.60
DCN	41.82	3.15	1.11	43.86	4.06	1.06	**35.46**	**6.01**	**1.18**	**29.44**	**9.01**	**1.28**

Table 3: DCN has been trained for $n = 20$ nodes and only one scale ($J = 1$). We used the pre-trained baseline for $n = 10$ as model on both leaves. BS1 and BS2 correspond to the baseline trained for $n = 10$ and $n = 20$ nodes respectively. Although for small n, both baselines outperform the DCN, the scale invariance prior of the DCN is leveraged at larger scales resulting in better results and scalability.

In Table 3 we report preliminary results of the DCN for the TSP problem. Although DCN and BS2 have both been trained on graphs of the same size, the dynamic model outperforms the baseline due to its powerful prior on scale invariance.

Although the scale invariance of the TSP is not as clear as in the previous problems (it is not straightforward how to use two TSP partial solutions to build a larger one), we observe that some degree of scale invariance it is enough in order to improve on scalability. As in previous experiments, the joint work of the split and merge phase is essential to construct the final solution.

7 Conclusions

We have presented a novel neural architecture that can discover and exploit scale invariance in discrete algorithmic tasks, and can be trained with weak supervision. Our model learns how to split large inputs recursively, then learns how to solve each subproblem and finally how to merge partial solutions.

Due to the generality of the DCN, three very different problems have been tackled, some with large and others with weak scale invariance. In all cases, our inductive bias leads to better generalization and computational complexity. In future work, we plan to extend the results of the TSP by increasing the number of splits J and by refining the supervised DCN model with the non-differentiable TSP cost. We also plan to experiment on other NP-hard combinatorial tasks.

References

Andrychowicz, Marcin and Kurach, Karol. Learning efficient algorithms with hierarchical attentive memory. *arXiv preprint arXiv:1602.03218*, 2016.

Bahdanau, Dzmitry, Cho, Kyunghyun, and Bengio, Yoshua. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Bello, Irwan, Pham, Hieu, Le, Quoc V, Norouzi, Mohammad, and Bengio, Samy. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2016.

Bronstein, Michael M, Bruna, Joan, LeCun, Yann, Szlam, Arthur, and Vandergheynst, Pierre. Geometric deep learning: going beyond euclidean data. *arXiv preprint arXiv:1611.08097*, 2016.

Cai, Jonathon, Shin, Richard, and Dawn, Song. Making neural programming architectures generalize via recursion. *ICLR*, 2017.

Dai, Hanjun, Khalil, Elias B, Zhang, Yuyu, Dilkina, Bistra, and Song, Le. Learning combinatorial optimization algorithms over graphs. *arXiv preprint arXiv:1704.01665*, 2017.

Duvenaud, David, Maclaurin, Dougal, Aguilera-Iparraguirre, Jorge, Gómez-Bombarelli, Rafael, Hirzel, Timothy, Aspuru-Guzik, Alán, and Adams, Ryan P. Convolutional networks on graphs for learning molecular fingerprints. In *Neural Information Processing Systems*, 2015.

Dwyer, Rex A. On the convex hull of random points in a polytope. *Journal of Applied Probability*, 25(04):688–699, 1988.

Fanty, Mark. Context-free parsing in connectionist networks. *Parallel natural language processing*, pp. 211–237, 1994.

Gilmer, Justin, Schoenholz, Samuel S, Riley, Patrick F, Vinyals, Oriol, and Dahl, George E. Neural message passing for quantum chemistry. *arXiv preprint arXiv:1704.01212*, 2017.

Graves, Alex, Wayne, Greg, and Danihelka, Ivo. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

Grefenstette, Edward, Hermann, Karl Moritz, Suleyman, Mustafa, and Blunsom, Phil. Learning to transduce with unbounded memory. In *Advances in Neural Information Processing Systems*, pp. 1828–1836, 2015.

Helsgaun, Keld. *An effective implementation of K-opt moves for the Lin-Kernighan TSP heuristic*. PhD thesis, Roskilde University. Department of Computer Science, 2006.

Hochreiter, Sepp and Schmidhuber, Jürgen. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Joulin, Armand and Mikolov, Tomas. Inferring algorithmic patterns with stack-augmented recurrent nets. In *Advances in Neural Information Processing Systems*, pp. 190–198, 2015.

Kaiser, Łukasz and Sutskever, Ilya. Neural gpus learn algorithms. *arXiv preprint arXiv:1511.08228*, 2015.

Kearnes, Steven, McCloskey, Kevin, Berndl, Marc, Pande, Vijay, and Riley, Patrick. Molecular graph convolutions: moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8): 595–608, 2016.

Nowak, Alex, Villar, Soledad, Bandeira, Afonso S, and Bruna, Joan. A note on learning algorithms for quadratic assignment with graph neural networks. *arXiv preprint arXiv:1706.07450*, 2017.

Pollack, Jordan B. The induction of dynamical recognizers. *Machine Learning*, 7(2-3):227–252, 1991.

Reed, Scott and de Freitas, Nando. Neural programmer-interpreters. *arXiv preprint arXiv:1511.06279*, 2015.

Scarselli, Franco, Gori, Marco, Tsoi, Ah Chung, Hagenbuchner, Markus, and Monfardini, Gabriele. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

Sukhbaatar, Sainbayar, Szlam, Arthur, Weston, Jason, and Fergus, Rob. End-to-end memory networks. In *Advances in Neural Information Processing Systems*, pp. 2431–2439, 2015.

Sutskever, Ilya, Vinyals, Oriol, and Le, Quoc V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pp. 3104–3112, 2014.

Tabor, Whitney. Fractal encoding of context-free grammars in connectionist networks. *Expert Systems*, 17(1):41–56, 2000.

Vinyals, Oriol, Bengio, Samy, and Kudlur, Manjunath. Order matters: Sequence to sequence for sets. *arXiv preprint arXiv:1511.06391*, 2015a.

Vinyals, Oriol, Fortunato, Meire, and Jaitly, Navdeep. Pointer networks. In *Advances in Neural Information Processing Systems*, pp. 2692–2700, 2015b.

Weston, Jason, Chopra, Sumit, and Bordes, Antoine. Memory networks. *arXiv preprint arXiv:1410.3916*, 2014.

Zaremba, Wojciech and Sutskever, Ilya. Learning to execute. *arXiv preprint arXiv:1410.4615*, 2014.

Zaremba, Wojciech and Sutskever, Ilya. Reinforcement learning neural turing machines-revised. *arXiv preprint arXiv:1505.00521*, 2015.

A REGULARIZATION WITH COMPUTATIONAL COMPLEXITY

As discussed previously, an added benefit of dynamic computation graphs is that one can consider computational complexity as a regularization criteria. We describe how computational complexity can be controlled in the split module.

A.1 SPLIT REGULARIZATION

We verify from Subsection 4.1 that the cost of running each split block $\mathcal{S}$ is linear on the input size. It results that the average case complexity $C_{\mathcal{S}}(n)$ of the whole split phase on an input of size n satisfies the following recursion:

$$\mathbb{E}C_{\mathcal{S}}(n) = \mathbb{E}\{C_{\mathcal{S}}(\alpha_s n) + C_{\mathcal{S}}((1-\alpha_s)n)\} + \Theta(n) \,, \tag{16}$$

where α_s are the fraction of input elements that are respectively sent to each output. Since this fraction is input-dependent, the average case is obtained by taking expectations with respect to the underlying input distribution. Assuming without loss of generality that $\mathbb{E}(\alpha_s) \geq 0.5$, we verify that the resulting complexity is of the order of $\mathbb{E}C_{\mathcal{S}}(n) \simeq \frac{n \log n}{\log \mathbb{E}\alpha_s^{-1}}$, which confirms the intuition that balanced partition trees ($\alpha_s = 0.5$) will lead to improved computational complexity. We can enforce α_s to be close to 0.5 by maximizing the variance of the split probabilities $p_\theta(z \mid X)$ computed by $\mathcal{S}_\theta$:

$$\mathcal{R}(\mathcal{S}) = -\left[M^{-1} \sum_{m \leq M} p_\theta(z \mid X)^2 - M^{-2} \left(\sum_m p_\theta(z \mid X) \right)^2 \right] \,. \tag{17}$$

B TRAINING DETAILS

The split parameters are updated with the RMSProp algorithm with initial learning rate of 0.01 and the merge parameters with Adam with initial learning rate of 0.001. Learning rates are updated as lr/k where k is the epoch number.

B.1 CONVEX-HULL

Training and test datasets have 1M and 4096 examples respectively. We use a batch size of 128. The split block has 5 layers with 15 hidden units. The merge block is a GRU with 512 hidden units. The number of scales of the DCN depends on the input size. Use 0 for 6-12, 1 for 12-25 and 2 for 25-50 for training. Use 1 for 25, 2 for 50, 3 for 100 and 4 for 200 at test time. The merge block is trained using the product of the continuous Γ matrices.

B.2 CLUSTERING

Training and test datasets have 1.3K and 4096 examples respectively. We use a batch size of 256. The split block has 10 layers with 30 hidden units.

B.3 TSP

Training and test datasets have 20K and 1K examples respectively. We use a batch size of 32. Both the split and the merge have 20 layers with 20 hidden units. The number of scales of the DCN is fixed to 1, both for training and test time.

C K-MEANS EXPERIMENTS

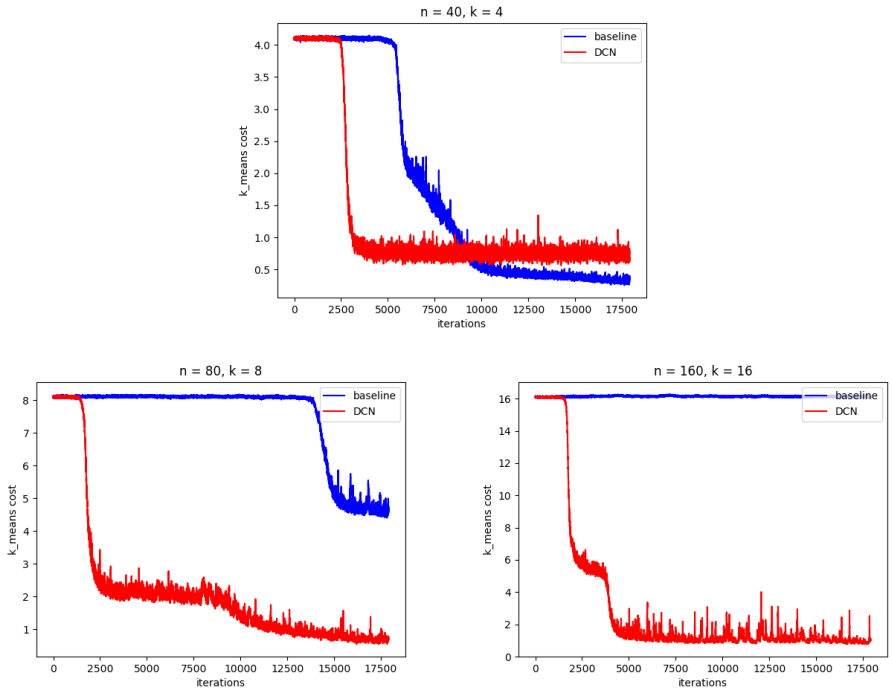

Figure 3: Training dynamics of the k-means cost both for the baseline and the DCN. Although we observe better performance of the baseline for the setting with smallest n and k, the training time it is always larger than the DNC due to the larger sample complexity.

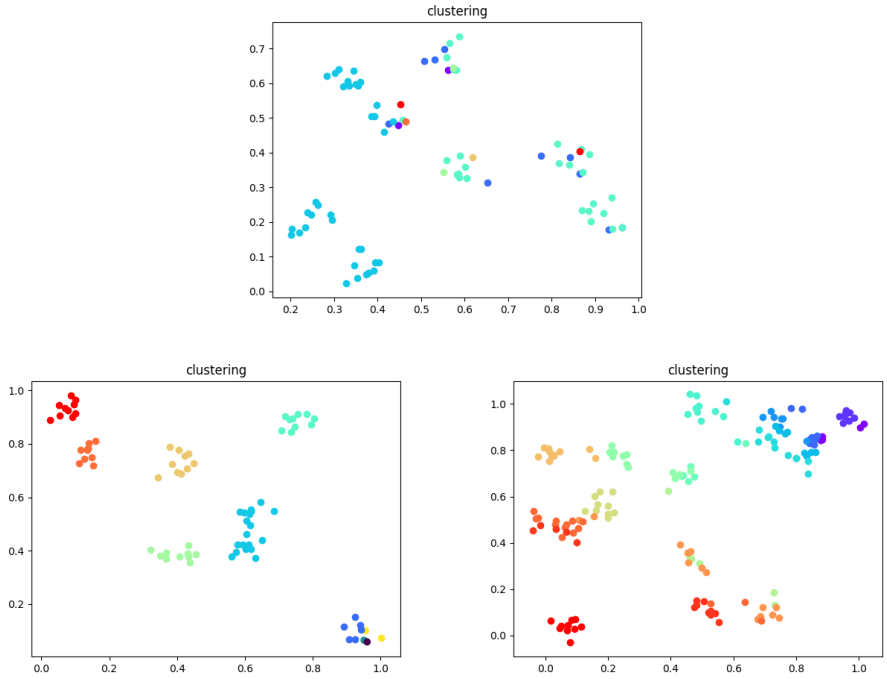

Figure 4: Examples of predictions. *Top*: Baseline; *Bottom:* DCN for 8, and 16 clusters.

D CONVEX HULL EXPERIMENTS

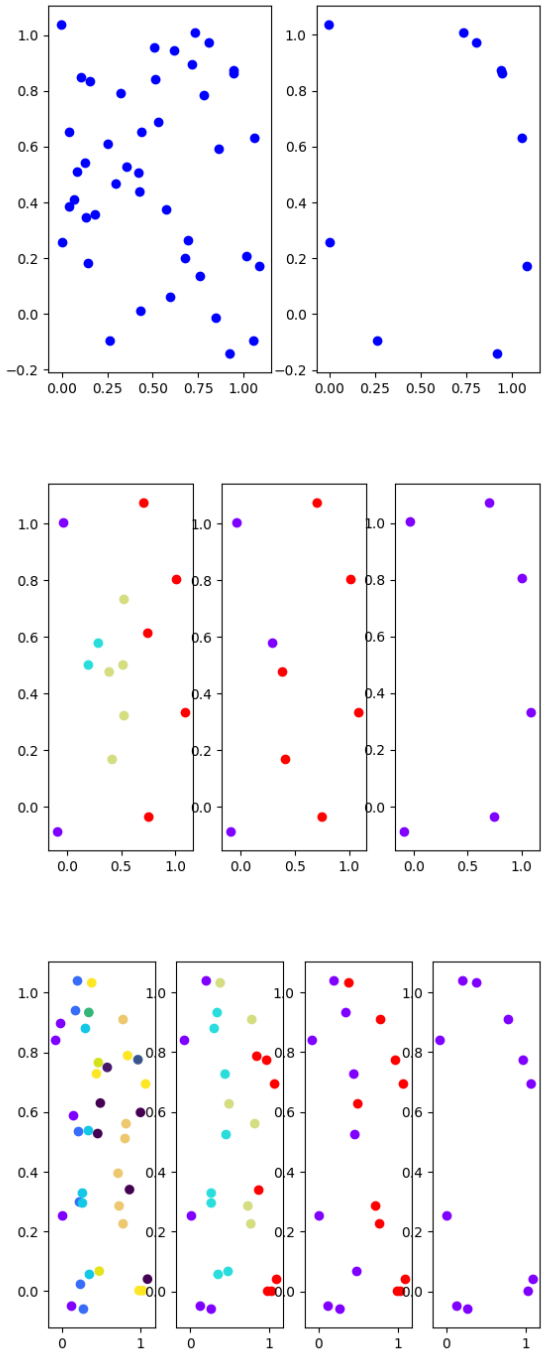

Figure 5: Output examples of the DCN at test time. Top: Baseline. Middle: DCN 1 scale. Bottom: DCN 2 scales.

E DETAILS OF POINTER NETWORK MODULE

The merge block $\mathcal{M}_\phi$ takes as input a pair of sequences Y_0, Y_1 and produces an output sequence O. We describe first the architecture for this module, and then explain on how it is modified to perform the finest scale computation $\tilde{\mathcal{M}}_\phi$. We modify a Pointer Network (PtrNet) Vinyals et al. (2015b) to our input-output interface as our merge block $\mathcal{M}_\phi$. A PtrNet is an auto-regressive model for tasks where the output sequence is a permutation of a subsequence of the input. The model encodes each input sequence $Y_q = (x_{1,q}, \ldots, x_{n_q,q})$, $q = 0, 1$, into a global representation $e_q := e_{q,n_q}$, $q = 0, 1$, by sequentially computing $e_{1,q}, \ldots, e_{n_q,q}$ with an RNN. Then, another RNN decodes the output sequence with initial state $d_0 = \rho(A_0 e_0 + A_1 e_1)$. The trainable parameters ϕ regroup to the RNN encoder and decoder parameters.

Suppose first that one has a target sequence $T = (t_1 \ldots t_S)$ for the output of the merge. In that case, we use a conditional autoregressive model of the form

$$\left\{ \begin{array}{ll} e_{q,i} = f_{\text{enc}}(e_{q,i-1}, y_{q,i}) & i = 1, \ldots, n_q \ , q = 0, 1 \ , \\ d_s = f_{\text{dec}}(d_{s-1}, t_{s-1}) & s = 1, \ldots, S \end{array} \right. \tag{18}$$

The conditional probability of the target given the inputs is computed by performing attention over the embeddings $e_{q,i}$ and interpreting the attention as a probability distribution over the input indexes:

$$\left\{ \begin{array}{ll} u^s_{q,i} = \phi_V^T \text{tanh}\left(\phi_e e_{q,i} + \phi_d d_s\right) & s = 0, \ldots S \ , q = 0, 1 \ , i \leq n_q \ , \\ p_s = \text{softmax}(u^s_\cdot) \ . & \end{array} \right. , \tag{19}$$

leading to $\Gamma = (p_1, \ldots, p_S)$. The output O is expressed in terms of Γ by binarizing its entries and multiplying it by the input:

$$, \ O = \mathcal{M}_\phi(Y_0, Y_1) = \bar{\Gamma} \left(\begin{array}{c} Y_0 \\ Y_1 \end{array} \right) \ , \text{ with } \ \bar{\Gamma}_{s,i} = \left\{ \begin{array}{ll} 1 & \text{if } i = \arg\max_{i'} p_s(i') \ . \\ 0 & \text{otherwise.} \end{array} \right. \tag{20}$$

However, since we are interested in weakly supervised tasks, the target output only exists at the coarsest scale of the partition tree. We thus also consider a generative version $\mathcal{M}^g_\phi$ of the merge block that uses its own predictions in order to sample an output sequence. Indeed, in that case, we replace equation (18) by

$$\left\{ \begin{array}{ll} e_{q,i} = f_{\text{enc}}(e_{q,i-1}, y_{q,i}) & i = 1, \ldots, n_q \ , q = 0, 1 \ , \\ d_s = f_{\text{dec}}(d_{s-1}, y_{o_{s-1}}) & s = 1, \ldots, S \end{array} \right. \tag{21}$$

where o_s is computed as $o_s = x_{\arg\max p_s}$, $s \leq S$. The initial merge operation at the finest scale $\tilde{\mathcal{M}}$ is defined as the previous merge module applied to the input $(X_{J,k}, \emptyset)$. We describe next how the successive merge blocks are connected so that the whole system can be evaluated and run.