Many Task Learning With Task Routing

Gjorgji Strezoski, Nanne van Noord and Marcel Worring
University of Amsterdam
ICCV 2019

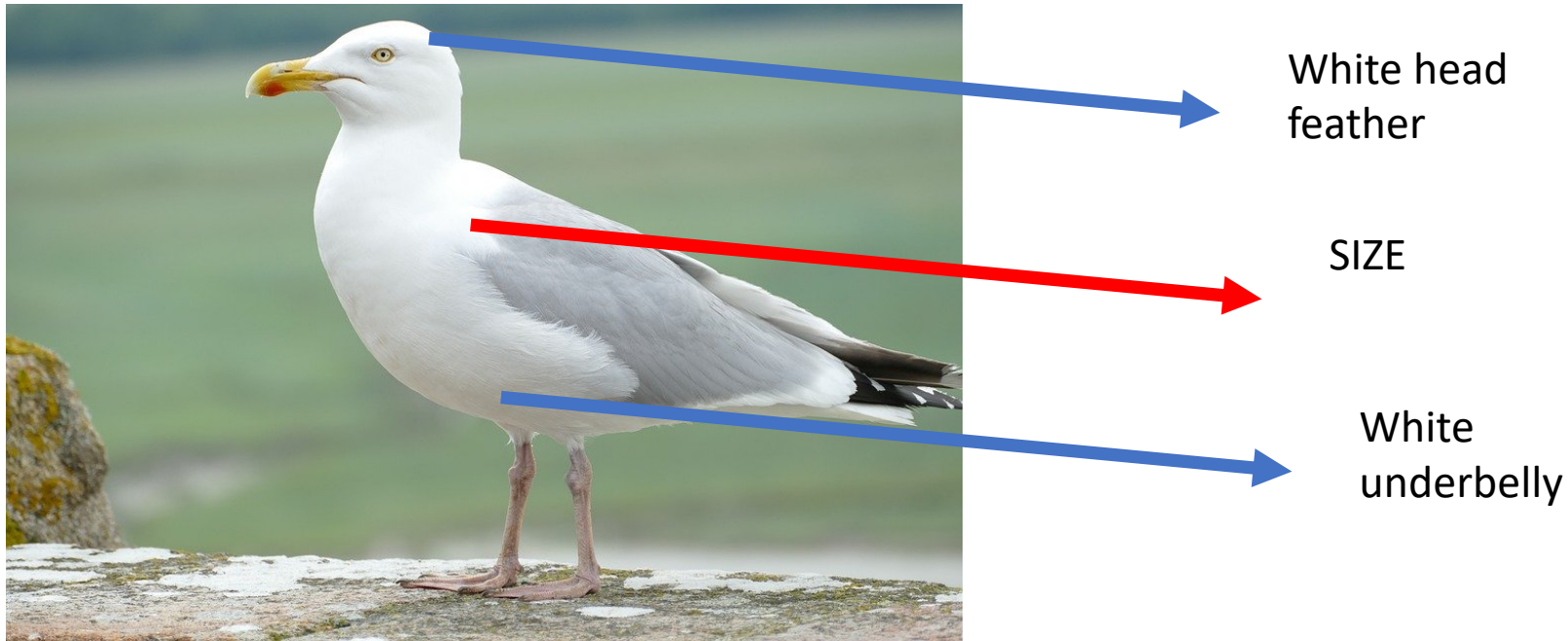
Presented by Huidong Xie

Background

- MTL (multi-task learning); STL (single-task learning)
- STL models usually have an abundance of parameters that have the capacity to fit to more than one task. The goal of MTL is to use these redundant parameters.
- Asymmetric MTL: training multiple small, auxiliary tasks for one main tasks. It is similar to transfer learning but asymmetric MTL is trained simultaneously.
- Symmetric MTL: aims to improve the performance of all tasks simultaneously. **They focus on this.**

Background

- Previous works are vulnerable to noisy and irrelevant tasks, which deteriorate performance. This is because of low feature robustness and the assumption that all tasks are positively related to each other.
- They solve this issue by randomizing the sharing structure and enforcing tasks to use different routes in the model.



Background

- Most previous methods have firm constraints in terms of how a model should be defined, structured and initialized.
- Their method is applicable to any deep MTL model with no structural adjustments.
- Also, since their method does not require structural adjustments, resources consumption will not increase significantly with the number of tasks.

Task Routing Layer

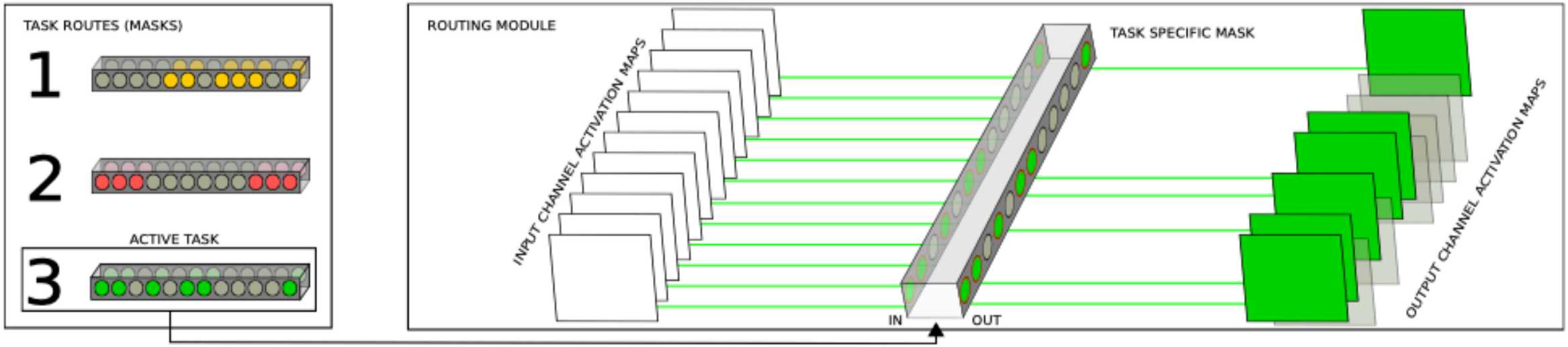


Figure 2: Operation of the TRL over the output from a convolutional layer (white channels). The current active task is used to select the mask (bright green are 1, and dark green are 0). After the element-wise multiplication across channels, the ones that remain are colored bright green and the nullified channels are brown and transparent.

Task Routing Layer

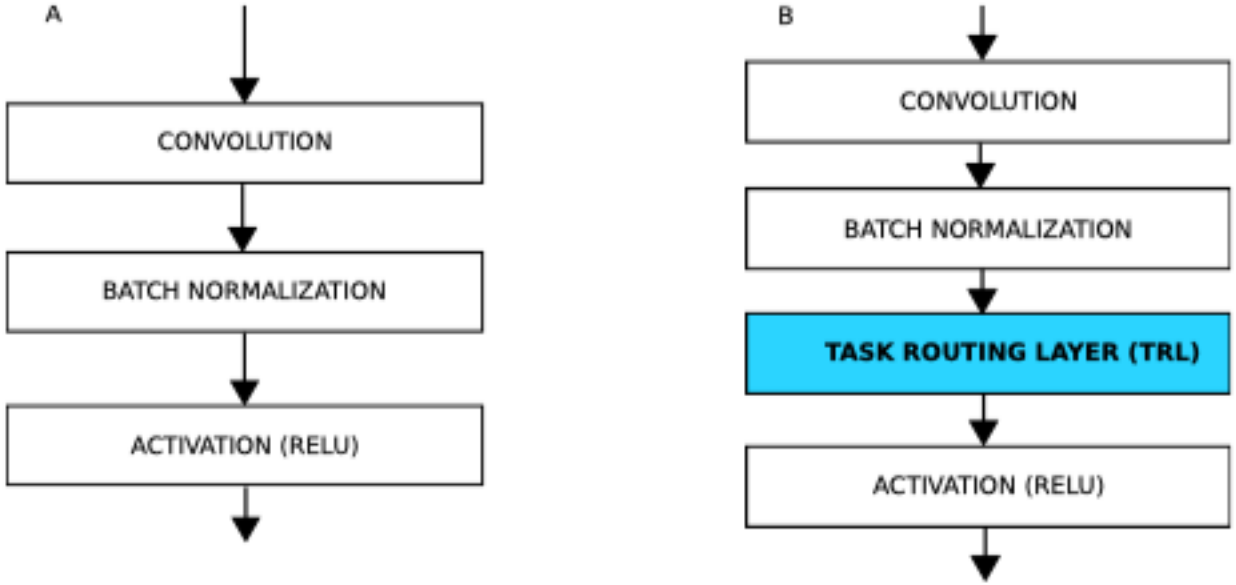


Figure 3: TRL placement (blue block) within a convolutional block. Section A (on the left) shows the convolutional block before adding the TRL, and Section B after.

- There is a hyper-parameter σ is a sharing ratio.
- It defines how many units are task specific, and how many are shared between tasks. $\sigma = 0$ means no sharing.

Training

Algorithm 1 Training epoch for TRL

```
procedure TRAIN(X)
    for X in X_Train do                    ▷ Training loop
        A ← sample(task_set)
        set_active_task(A)
        forward(X)
```

- A task is randomly selected, and the corresponding mask is applied.
- This setting is global, so it will not affect existing ways of propagations or back-propagations.
- Mask is fixed and not trainable.

Datasets

- UCSD Birds contains 11,788 bird images over 200 bird species with 312 binary attribute annotations.
- Visual Decathlon (VD) is a benchmark that evaluates the ability to capture simultaneously 10 different visual domains. They evaluate the performance by assign a max score of 10,000 (1,000 per task) based on per-task accuracies using the official challenge metric.
- FashionMNIST and CIFAR-10 are used as proof of concept of their method.
- CelebA consists of more than 200,000 face images with binary annotations on 40 facial attributions.
- UT-Zappos50K consists more than 50,000 catalog images collected from the web. It contains 4 annotations (show type, gender, height of the heel and the shoe closing mechanism.)

Experimental details

- They used VGG-11, VGG-16 or ResNet50 as their default network setting. They add the Task Routing Layer after each convolutional layer.
- Batch-size 64, normalization by dataset mean.
- SGD with learning rate 0.01 and momentum 0.5.

Result

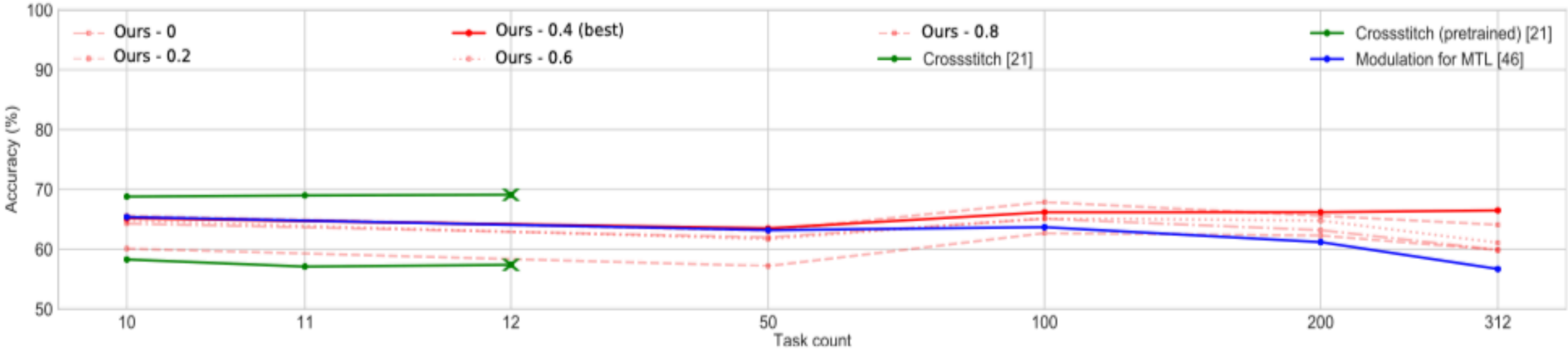


Figure 4: Accuracy comparison on the UCSD-Birds dataset on 10, 50, 100, 200, 312 task between our method (in red) with $\sigma = [0, 1]$, cross-stitch networks [21] (in green) and modulation for MTL [48] (in blue). Cross-stitch networks scale up to 12 tasks, where as modulation for MTL and our approach fit to the full number of tasks. The best performing sharing ratio $\sigma = 0.4$ is set in strong red, and the other σ values in light red.

Result

Table 1: Average scores on the VD challenge. Best overall approach is highlighted in gray.

Run	VD Score	Aircraft	Cifar-100	Daimler	DTD	GTSRB	ImageNet-12	Omniglot	SVHN	UCF-101	VGG-Flowers
ResAdapt [25] ($\sigma = 0$)	2851.31	299.88	195.96	155.41	261.51	472.6	224.15	337.05	**282.8**	231.69	390.26
Ours $\sigma = 0.2$	2873.84	302.1	200.01	162.79	267.22	472.2	210.2	344.12	265.4	250.02	399.78
Ours $\sigma = 0.4$	**2919.26**	**305.2**	204.12	**165.89**	**273.28**	469.2	**228.39**	345.08	272.77	252.12	**403.21**
Ours $\sigma = 0.6$	2870.26	287.2	206.12	148.89	256.28	474.2	223.39	**350.08**	260.77	261.12	402.21
Ours $\sigma = 0.8$	2806.26	285.2	208.12	139.89	253.28	455.2	222.39	338.08	249.77	**263.12**	391.21
Ours $\sigma = 1$	2768.26	282.2	**214.12**	132.89	256.28	445.2	207.39	339.08	239.77	261.12	390.21

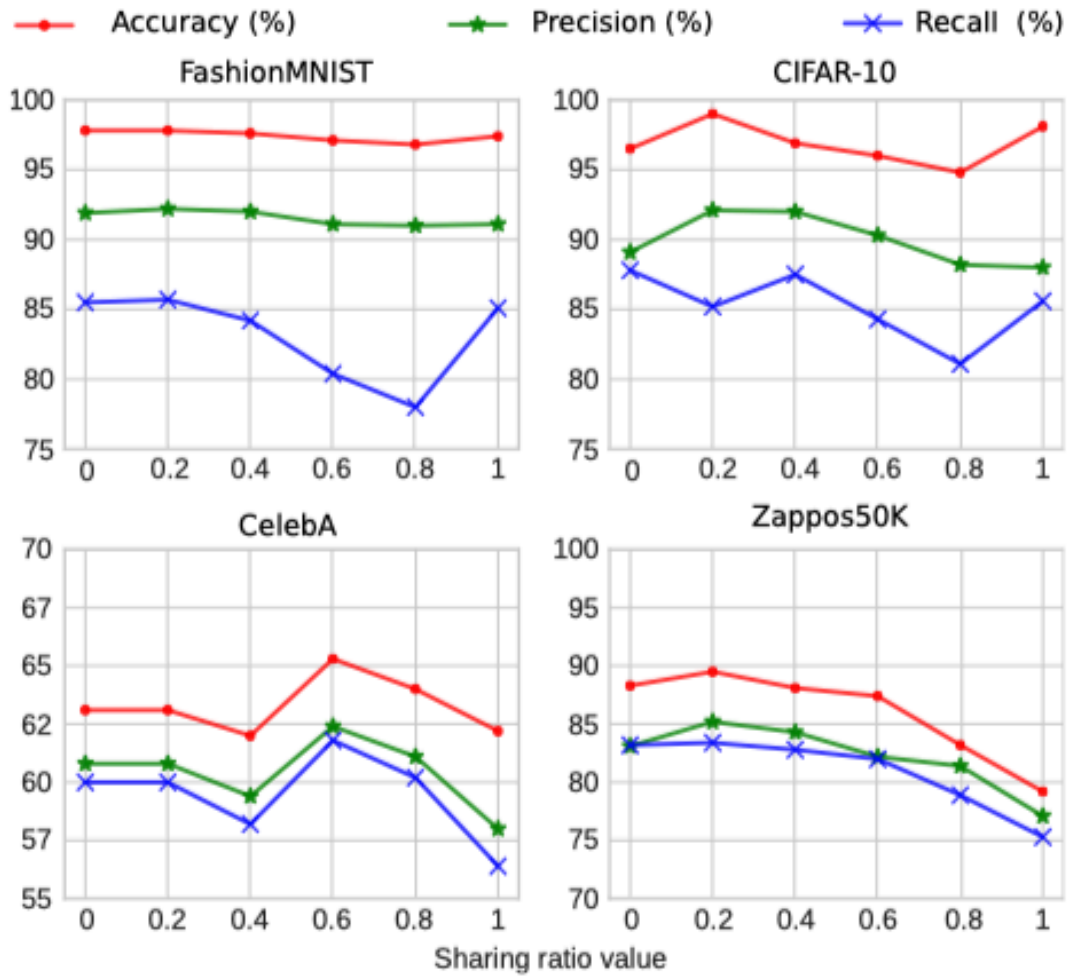


Figure 5: Effects of the sharing ratio σ value on accuracy, precision and recall in Fashion-MNIST (top left), CIFAR10 (top right), CelebA (bottom left) and Zappos50K (bottom right). Partially sharing units between tasks is beneficial to performance with sharing ratios $\sigma = [0.2, 0.6]$ compared to a fully shared network $\sigma = 1$ or many distinct subnetworks without sharing $\sigma = 0$.

- Precision = tp/(tp+fp)
- Recall = tp/(tp+fn)
- Accuracy = (tp+tn)/(tp+tn+fp+fn)

Result

Table 2: Average scores over 5 runs on the on FashionMNIST, CIFAR-10, Zappos50K and CelebA (10 and 40 Tasks) on the complete dataset with the complete sharing ratio scope $\sigma = [0, 1]$. Because for $\sigma = 0$ no sharing occurs and for $\sigma = 1$ our approach reverts to hard shared MTL, we group them separately. The overall best performing approach across all datasets is highlighted in gray and the best approaches per dataset are in bold. Fields marked with *n/a* signify experiments for which the method could not scale to the task count.

Dataset	FashionMNIST			CIFAR-10			Zappos50K			CelebA			CelebA (Full)		
Number of Tasks	10			10			4			10			40		
Approach	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Cross-stitch	**98.1**±1.14	91.4±1.02	86.1±0.23	98.5±0.17	91.6±1.18	85.9±1.07	84.7±2.23	82.2±1.12	81.8±1.29	71.5±1.96	68.0±1.38	67.0±0.83	n/a	n/a	n/a
Modulation	96.9±2.04	91.0±1.14	80.1±0.54	63.2±1.13	57.4±2.14	53.2±3.10	63.7±2.76	60.4±1.94	59.8±2.02	71.9±1.66	70.2±2.18	69.4±2.63	64.1±1.43	61.0±1.81	60.4±1.45
Ours σ = Adapt	96.3±0.04	90.6±0.04	84.1±0.05	98.1±0.06	88.3±0.03	85.9±0.05	88.3±0.11	81.7±0.02	80.6±0.04	71.9±0.07	68.2±0.08	66.3±0.17	63.0±0.08	59.0±0.15	57.1±0.11
Ours $\sigma = 0$	97.8±0.25	91.9±0.44	85.5±0.32	96.5±0.42	89.1±0.98	**87.8**±0.24	88.3±0.31	83.1±0.54	83.2±0.42	70.1±0.08	68.0±0.22	67.4±0.78	63.1±0.33	60.8±0.05	60.0±0.21
Ours $\sigma = 1$	97.4±0.01	91.1±0.07	85.1±0.04	98.1±0.03	88.0±0.03	85.6±0.01	79.2±0.10	77.1±0.09	75.3±0.08	69.9±0.13	67.2±0.10	66.8±0.06	62.2±0.07	58.0±0.07	56.4±0.11
Ours $\sigma = 0.2$	97.8±0.06	**92.2**±0.11	**85.7**±0.03	**99.0**±0.03	**92.1**±0.08	85.2±0.06	**89.5**±0.03	**85.2**±0.04	**83.4**±0.12	**73.2**±0.15	**71.4**±0.11	**70.8**±0.13	63.1±0.12	60.8±0.12	60.0±0.13
Ours $\sigma = 0.4$	97.6±0.05	92.0±0.08	84.2±0.07	96.9±0.09	92.0±0.09	87.5±0.15	88.1±0.17	84.3±0.18	82.8±0.14	73.0±0.14	71.4±0.12	70.2±0.12	62.0±0.25	59.4±0.24	58.2±0.24
Ours $\sigma = 0.6$	97.1±0.10	91.1±0.08	80.4±0.08	96.0±0.06	90.3±0.05	84.3±0.07	87.4±0.11	82.2±0.17	82.0±0.13	72.7±0.05	71.0±0.04	69.6±0.09	**65.3**±0.22	**62.4**±0.18	**61.8**±0.17
Ours $\sigma = 0.8$	96.8±0.08	91.0±0.08	78.0±0.03	94.8±0.09	88.2±0.11	81.1±0.10	83.2±0.04	81.4±0.01	78.9±0.01	71.4±0.05	70.1±0.06	69.1±0.08	64.0±0.14	61.1±0.10	60.2±0.09

Result

Table 3: Average scores using the routing module over an increasing number tasks for the UCSD-Birds dataset and a sharing ratio of $\sigma = [0, 1]$. Because for $\sigma = 0$ no sharing occurs and for $\sigma = 1$ our approach reverts to hard shared MTL, we group them separately. Fields marked with *n/a* signify experiments for which the method could not scale to the task count. The pretrained cross-stitch networks experiment is marked with a star (*). The overall best performing method is highlighted in gray and the best performing model per task setting is set in bold.

Dataset	UCSD-Birds														
Number of tasks	10			50			100			200			312		
Approach	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall	Accuracy	Precision	Recall
Cross-stitch [21]	58.3	55.6	54.2	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Cross-stitch [21] *	**68.8**	**67.4**	**67.0**	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a	n/a
Modulation [48]	65.4	59.8	55.2	63.2	57.4	53.2	63.7	60.4	59.8	61.2	58.6	57.3	56.7	51.8	50.2
Ours $\sigma = 0$	64.3	62.4	55.3	62.0	60.6	54.6	65.1	62.7	61.1	63.2	60.2	58.8	59.9	57.2	56.1
Ours $\sigma = 1$	62.3	57.4	51.8	58.6	56.8	54.2	60.7	58.1	57.8	60.0	58.6	56.8	59.6	53.9	52.2
Ours $\sigma = 0.2$	65.6	62.9	57.0	63.1	62.9	57.2	**67.8**	63.3	60.9	65.6	63.6	63.2	64.1	61.6	60.2
Ours $\sigma = 0.4$	65.1	62.7	55.9	**63.5**	**63.0**	**59.9**	66.2	**63.8**	**61.2**	**66.2**	**64.2**	**63.7**	**66.5**	**62.3**	**61.8**
Ours $\sigma = 0.6$	64.9	62.1	54.8	61.7	59.9	59.0	65.2	60.9	59.5	64.8	62.0	59.8	61.1	59.2	59.0
Ours $\sigma = 0.8$	60.1	55.0	50.2	57.2	52.2	50.0	62.7	60.4	59.8	62.3	59.2	58.0	59.9	55.1	54.2