Multi-task Learning by Maximizing Statistical Dependence Supplemental Material

Youssef. A. Mejjati
University of Bath

Darren Cosker
University of Bath

Kwang In Kim
University of Bath

We present a discussion on embedding and visualizing multiple tasks using the FSIC similarity measure, along with additional experimental results including the regression performance on two regression benchmark datasets (RF1 and ENB described in the main paper) as well as adaptations of Pentina et al.'s curriculum learning approach (Sec. 2) to ranking and regression settings . We reproduce some content of the main paper to make this supplemental self-contained.

1. Task embedding and visualization

Suppose we have an empirical estimate matrix $F \in \mathbb{R}^{N\times T}$ storing evaluations of multiple task estimators $\{f^1,...,f^T\}$ on N data points $\{\mathbf{x}_1,\dots,\mathbf{x}_N\}$: $[F]_{i,j} = f^j(\mathbf{x}_i)$. The corresponding normalized FSIC matrix is defined as

$$[\Phi(F)]_{i,j} = \frac{\|\hat{\mathbf{u}}^{ij}\|^2}{\|\hat{\mathbf{u}}^{ii}\|\|\hat{\mathbf{u}}^{jj}\|}, \tag{1}$$

$$\hat{\mathbf{u}}^{ij} = \frac{(K^i \circ K^j)\mathbf{1}}{N-1} - \frac{(K^i\mathbf{1})\circ(K^j\mathbf{1})}{N(N-1)}, \tag{2}$$

where $\circ$ denotes element-wise product, $\mathbf{1} = [1,...,1]^\top$, and $[K^i]_{(j,k)} = k^i(q^i_j,[F]_{k,i})$ for the test locations $\{q^i_j\}_{j=1}^{100}$ of the i-th task. We use a Gaussian kernel k^i that guarantees the consistency of individual FSIC estimators $\hat{\mathbf{u}}^{ij}$:

$$k^i(f,g) = \exp\left(-\frac{\|f-g\|^2}{(\sigma^i_k)^2}\right), \tag{3}$$

with a parameter $\sigma^i_k > 0$.

Each entry in the FSIC matrix $\Phi(F)$ is bounded in [0,1] and it can be considered as a measure of similarity between estimators $\mathbf{f}^i$ and $\mathbf{f}^j$ as *data points* in an N-dimensional space: Inspecting Eq. 2 reveals that, evaluated at columns of F, the empirical FSIC estimate $\hat{\mathbf{u}}$ constitutes a positive definite kernel and, therefore, it induces a distance measure on $\mathbb{R}^N$. This facilitates embedding individual estimations into a graph where the graph Laplacian L is constructed based on the *kernel matrix* $\Phi(F)$. Using the graph Laplacian, one could embed all tasks into a low-dimensional visualization space.

Figure 1 shows the results of two-dimensional embedding of 312 tasks in the Birds dataset which provides 11,788 bird images of 200 bird species provided with 312 binary attribute annotations. Each input image is represented based on 1000-dimensional VGG19 features. The embedding is performed based on the FSIC matrix $\Phi(F)$ of the independently trained rank SVM estimates F. First, the normalized graph Laplacian $L = I - D^{-1/2}WD^{-1/2}$ is calculated from the weight matrix W:

$$[W]_{i,j} = \exp\left(-\frac{1-[\Phi(F)]_{i,j}}{\sigma^2_w}\right), \tag{4}$$

with D being a diagonal matrix storing the column sum of W: $[D]_{i,j} = \sum_j [W]_{i,j}$ and σ^2_w is fixed at 5. Thereafter, two dimensional embedding coordinates $[[E]_{i,1},[E]_{i,2}]$ of the i-th attribute are calculated by combining the two eigenvectors $\mathbf{e}^1$ and $\mathbf{e}^2$ of L corresponding to two smallest non-zero eigenvalues λ^1 and λ^2, respectively:

$$E = [\mathbf{e}^1,\mathbf{e}^2]. \tag{5}$$

Visualizing task dependence helps us to understand the nature of the problem: *E.g.*, attributes 'has_throat_color::brown' and 'has_upperparts_color::brown' are indeed statistically related (Fig. 1(top-right)). Furthermore, it gives an insight into identifying potential subsets of tasks that can benefit from MTL. To facilitate this, we performed spectral clustering of the entire dataset (312 tasks) into 15 clusters[1](Fig. 1(top-left)). Figure 1(top-right) highlights the member tasks of the first cluster that we used in the main paper as solid circles: We used a subset of size 10 (blue circles in Figure 1). As shown in the experiments, performing task pre-clustering makes all tasks within a single cluster statistically dependent and therefore renders the simple strategy of uniformly enforcing the task similarity (regMTL) a competitive approach.

2. Additional results

Curriculum learning. Pentina et al.'s **curriculum learning of multiple-tasks (CLMTL)** approach applies a curriculum learning strategy to MTL: Instead of treating all the tasks symmetrically, they form a sequence

[1]For spectral clustering (to 15 clusters), a 15-dimensional data representation formed by the first 15 eigenvectors $E = [\mathbf{e}^1,...,\mathbf{e}^{15}]$ is used.

of tasks where information sharing happens only between consecutive tasks, hence suppressing the influence of outlier tasks and limiting the negative transfer effect.

The key challenge in this approach is to find the order on which the tasks are processed. Given an order π, where $\pi(i)$ contains the index of the task being processed at iteration i, a domain adaptation algorithm is used to obtain $\mathbf{w}^{\pi(i)}$ from $\mathbf{w}^{\pi(i-1)}$. For the classification problem, they constructed the solution $\mathbf{w}^{\pi(i)}$ as an adaptive support vector machine (SVM) that minimizes the energy functional:

$$\mathcal{E}(\mathbf{w}) = \|\mathbf{w}-\mathbf{w}^{\pi(i-1)}\|^2 + \frac{C}{l(i)}\sum_{k=1}^{l(i)}\xi_k^i, \tag{6}$$

$$s.t \quad y_k^i\langle \mathbf{w},\mathbf{x}_k^i\rangle \geq 1-\xi_k^i, \quad \xi_k^i \geq 0, \quad \forall\ 1\leq k\leq l(i), \tag{7}$$

with $f(\mathbf{x})=\mathbf{w}^\top\mathbf{x}$ and $l(i)$ being the number of labeled data points for task i: $\{(\mathbf{x}_1,y_1),...,(\mathbf{x}_{l(i)},y_{l(i)})\}$.

Based on this adaptation strategy, Pentina et al.'s original algorithm identifies the sequence π as to minimize the bound on the generalization error. Their error bound builds upon McAllester's PCA Bayesian bound [6] and therefore, it can be directly applied to only bounded losses, e.g. classification losses. For ranking and regression problems where unbounded losses are typically used:

$$l_{\text{rank}}(\mathbf{x}_i,\mathbf{x}_j;f) = \max(0,1-(f(\mathbf{x}_i)-f(\mathbf{x}_j)))^2, \tag{8}$$

$$l_{\text{regr}}(\mathbf{x}_i,y_i;f) = (f(\mathbf{x}_i)-y_j)^2, \tag{9}$$

the corresponding applications of their algorithm are not straightforward. Instead, we construct adaptations based on their algorithmic construction. First, we observe that at iteration i, their algorithm decides the next task T^i that minimizes the following objective function:

$$\mathcal{O}(T^j) = \frac{1}{l(j)}\sum_{k=1}^{l(j)}\bar{\Phi}\left(\frac{y_k^j\langle \mathbf{w}^j,\mathbf{x}_k^j\rangle}{\|\mathbf{x}_k^j\|}\right) + \frac{\|\mathbf{w}^j-\mathbf{w}^{\pi(j-1)}\|^2}{2\sqrt{\bar{l}}}, \tag{10}$$

where $\bar{l}$ is the harmonic mean of $\{l(j)\}_{j=1}^L$ and $\bar{\Phi}(z) = \frac{1}{2}(1-\text{erf}(\frac{z}{\sqrt{2}}))$ with erf being the *error function*. The original CLMTL algorithm (approximately) minimizes an upper bound on the classification error by selecting the task that minimizes $\mathcal{O}$ in Eq. 10 at each iteration. Now inspecting the objective $\mathcal{O}$, it can be seen that it consists of an increasing function of classification training error (the first term) and the regularization energy (the second term). Our strategy is to replace the first term by the corresponding functions of ranking and regression losses:

$$\mathcal{O}_{\text{regr}}(T^j) = \frac{\lambda}{l(j)}\sum_{k=1}^{l(j)}\bar{\Phi}(l_{\text{regr}}(\mathbf{x}_k,y_k;f)) + \frac{\|\mathbf{w}^j-\mathbf{w}^{\pi(i-1)}\|^2}{2\sqrt{\bar{l}}} \tag{11}$$

$$\mathcal{O}_{\text{rank}}(T^j) = \frac{\lambda}{|P^j|}\sum_{m,n\in P^i}\bar{\Phi}(l_{\text{rank}}(\mathbf{x}_m,\mathbf{x}_n;f)) + \frac{\|\mathbf{w}^j-\mathbf{w}^{\pi(i-1)}\|^2}{2\sqrt{\bar{P}}}, \tag{12}$$

where P^j is the set of ranking labels for task T^j, and $\bar{P}$ is the harmonic mean of $\{|P^j|\}_{j=1}^L$.

The interpretation of CLMTL's operational characteristics applied to these objectives are straightforward: It selects the task whose solution when found, does not deviate significantly from the previous task T^{i-1} and exhibits a small training error. Unfortunately, the solid theoretical interpretation of the original CLMTL is not anymore applicable. It should be noted that we introduced an additional scaling parameter $\lambda>0$ to balance the contributions of the training error and regularization terms. Similarly to the other algorithms that we compare with, the two hyper-parameters λ and C are tuned based on the validation sets.

Results. Overall, we improve upon the baseline independent estimator (Base_1) by adopting MTL approaches (Tables 1-2). While not all datasets and attributes show significant improvement, MTL approaches are on par with or outperform independent estimators. Since not all tasks are equally related (as suggested in Fig. 1), regMTL—which uniformly enforces pairwise task similarity—is further improved by allowing sparsity in task dependence (MTFL,AMTL) and/or task outliers (LRMTL). The performance variations of different MTL algorithms are significant (OSR, PubFig, Shoes, RF1, and ENB), while for SUN and Birds datasets the variation is less significant but noticeable. Among the four recent parametric MTL algorithms (CLMTL, MTFL, LRMTL, AMTL), LRMTL turned out to be the best followed by AMTL, but there was no clear winner indicating the complementary nature of different similarity-enforcing strategies. Also, for Birds dataset where by construction, all tasks are strongly related, the classical regMTL is competitive.

All five existing algorithms use shared parametric forms and extending them for the multiple heterogeneous estimator case is not straightforward. The importance of breaking this limitation can be clearly seen by comparing the results with the heterogeneous baselines (Base_2): Especially, for the OSR and RF1 datasets, by simply adopting heterogeneous estimators including DNNs, GPs and SVMs, even independent training already significantly improved performance. Being able to apply the MTL to these heterogeneous baselines, our algorithm further improves the performance and is consistently ranked among the best three results. In particular, our algorithm constantly improves upon the initial Base_2.

Bonilla et al.'s non-parametric Gaussian process-based MTL approach (GPMTL) [2] produced the best results for the second target attribute of the ENB dataset, improving the baseline with a large margin. However, their results on RF1 indicates that the performance varies significantly across different target attributes. The Spike and slab variational inference (SNS) demonstrated a similar behavior.

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3), 2008. 4

[2] E. V. Bonilla, K. M. Chai, and C. Williams. Multi-task Gaussian process prediction. In *NIPS*, pages 153–160, 2008. 3, 4

[3] J. Chen, J. Liu, and J. Ye. Learning incoherent sparse and low-rank patterns from multiple tasks. *ACM TKDD*, 5(4):22:1–28, 2012. 4

[4] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *KDD*, pages 109–117, 2004. 4

[5] G. Lee, E. Yang, and S. J. Hwang. Asymmetric multi-task learning based on task relatedness and loss. In *PMLR (Proc. ICML)*, pages 230–238, 2016. 4

[6] D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999. 2

[7] A. Pentina, V. Sharmanska, and C. H. Lampert. Curriculum learning of multiple tasks. In *CVPR*, pages 5492–5500, 2015. 4

[8] M. K. Titsias and M. Lázaro-Gredilla. Spike and slab variational inference for multi-task and multiple kernel learning. In *NIPS*, pages 2339–2347, 2011. 4

Table 1. Ranking performances of different MTL algorithms. Kendall's Tau correlations×100 ± std.×100 are presented (higher is better). The three best results are highlighted with **boldface blue**, *italic green*, and plain orange fonts, respectively.

Dataset	Target	Base_1	Base_2	CLMTL [7]	regMTL [4]	MTFL [1]	LRMTL [3]	AMTL [5]	Ours
OSR	1	88.26±0.83	90.26±0.55	88.20±1.04	89.93±0.63	89.09±1.24	*90.67±1.06*	88.34±0.73	**91.95±0.74**
	2	81.30±0.62	*86.01±0.84*	81.40±0.96	85.84±1.02	84.15±1.13	81.28±1.15	81.47±0.58	**86.33±0.92**
	3	71.00±1.04	*75.01±1.42*	71.87±1.04	74.52±2.28	73.89±2.62	73.03±1.31	72.42±1.08	**76.30±1.03**
	4	72.39±1.77	*77.66±1.23*	73.78±1.50	77.41±0.99	75.10±2.08	73.74±2.05	73.35±1.54	**79.01±1.27**
	5	75.52±1.19	79.30±1.08	77.17±1.12	*79.42±1.20*	79.02±1.71	78.08±0.89	77.44±0.83	**82.52±1.13**
	6	76.12±1.27	80.04±1.73	77.66±0.98	*80.12±1.59*	78.23±1.30	76.97±1.74	77.15±1.02	**80.55±1.34**
PubFig	1	64.50±2.53	66.40±3.08	65.40±2.70	60.55±2.59	62.88±2.51	*71.60±1.28*	64.47±2.56	**71.98±2.89**
	2	57.10±3.08	60.07±3.53	54.23±4.59	53.12±3.39	54.12±4.07	*62.14±4.80*	57.10±3.08	**64.71±3.58**
	3	64.22±2.06	66.53±2.77	68.42±2.10	63.97±3.00	65.23±4.05	**72.52±1.43**	68.31±1.50	*72.06±2.70*
	4	61.91±1.37	64.33±2.44	68.83±1.86	63.01±2.00	60.69±2.99	**70.34±2.37**	*69.16±2.04*	69.15±4.53
	5	55.82±3.33	58.48±2.97	62.76±1.58	54.22±3.64	55.36±3.37	*65.08±1.28*	62.09±3.12	**68.65±2.26**
	6	75.12±1.54	77.34±2.54	75.11±1.45	74.65±1.53	72.86±3.40	*77.43±2.18*	75.13±1.55	**78.18±3.31**
	7	58.79±3.03	62.66±3.54	62.64±2.32	57.66±3.74	59.36±3.99	**66.53±1.54**	58.79±3.03	*65.34±4.76*
	8	60.05±1.69	61.91±2.68	57.33±2.88	60.13±1.31	58.08±2.65	**62.63±2.01**	60.05±1.68	*62.54±2.89*
	9	52.44±2.72	*57.09±3.43*	46.96±4.53	53.74±3.54	53.78±3.25	56.65±4.92	52.55±2.87	**57.53±3.80**
	10	58.27±3.86	61.51±2.77	63.93±1.92	59.15±2.18	58.25±3.36	**66.88±1.78**	64.21±3.45	*66.47±3.53*
	11	63.21±1.82	66.81±2.39	69.53±1.82	64.15±1.36	63.91±3.60	*74.05±1.25*	70.07±1.58	**74.99±1.15**
Shoes	1	68.09±1.47	67.87±0.70	68.68±1.88	69.08±1.76	68.43±1.06	*69.72±1.31*	68.84±1.90	**71.93±0.91**
	2	56.39±1.84	*60.27±2.71*	56.81±2.63	59.04±1.72	57.81±2.08	58.02±3.00	57.71±1.88	**60.37±2.52**
	3	30.50±3.65	**34.43±2.66**	30.08±4.65	32.39±3.60	32.55±3.39	30.55±2.60	31.09±3.47	*34.42±2.61*
	4	46.18±1.63	*48.41±2.31*	46.74±1.81	46.85±2.11	46.04±2.72	46.24±1.83	45.81±2.10	**48.47±2.26**
	5	61.44±1.99	61.64±1.98	62.55±1.67	*63.92±1.51*	62.21±2.09	62.77±1.81	63.06±1.40	**64.79±0.99**
	6	61.87±2.30	60.80±3.56	62.34±2.41	**64.40±2.36**	*62.79±2.31*	62.62±1.81	62.46±1.92	62.06±2.14
	7	52.58±1.51	*56.43±2.02*	51.07±1.53	53.64±1.65	52.40±2.54	52.58±2.85	52.95±1.28	**57.15±2.45**
	8	49.56±1.73	48.89±0.86	50.06±1.55	50.34±2.12	**51.98±1.92**	49.68±1.66	50.24±1.30	*50.78±1.50*
	9	61.57±1.97	62.78±2.59	60.02±1.33	62.07±2.44	*63.34±2.33*	62.89±1.81	61.63±1.90	**67.11±1.12**
	10	66.91±1.16	66.83±2.33	68.66±1.23	69.16±1.02	68.10±1.56	*69.34±1.43*	68.86±1.91	**72.09±1.08**
SUN	1	66.18±3.15	68.24±4.32	66.87±4.85	66.52±4.41	68.39±4.39	**72.01±5.65**	65.21±7.56	*70.62±3.35*
	2	71.24±3.11	*75.31±4.06*	71.90±4.41	75.04±3.08	73.78±5.97	**75.89±1.08**	72.60±4.50	73.74±4.53
	3	76.84±1.16	76.77±1.87	**79.77±1.25**	75.74±2.15	75.87±1.90	77.73±2.07	77.62±1.37	*78.78±1.82*
	4	79.03±1.21	80.40±1.32	*82.61±1.32*	79.19±3.37	79.58±1.52	**84.20±0.40**	79.75±3.30	82.49±0.87
	5	79.66±1.42	78.67±2.41	**81.56±1.49**	78.43±1.68	78.42±1.04	80.64±1.63	80.59±1.62	*80.66±1.55*
	6	80.75±0.81	79.76±1.79	*82.34±0.99*	78.78±3.35	79.47±0.95	**82.71±0.90**	81.44±0.79	82.14±1.04
	7	79.76±0.65	79.41±1.00	*79.95±1.06*	79.62±0.88	78.17±1.16	78.76±1.36	79.65±0.90	**80.03±0.81**
	8	83.91±0.53	84.12±0.67	81.20±2.78	**84.21±0.60**	*84.13±0.60*	83.39±1.46	84.08±0.77	83.83±0.72
	9	**63.34±1.10**	63.13±1.01	62.82±1.55	61.91±2.30	62.53±0.73	62.67±0.30	*63.23±1.22*	63.14±0.90
	10	*82.23±3.50*	80.85±5.77	76.39±5.39	78.79±6.76	77.30±7.76	**84.64±2.42**	80.60±5.41	81.31±2.91
Birds	1	56.78±3.06	56.76±3.04	57.36±1.87	**60.43±3.24**	52.51±5.00	55.07±2.48	*59.38±4.26*	58.70±3.69
	2	42.18±3.24	41.23±4.18	44.76±10.96	**54.15±4.33**	52.66±3.51	47.88±5.18	53.43±3.09	*53.88±2.46*
	3	57.74±1.84	57.56±2.48	59.11±1.70	**61.67±2.69**	59.30±2.15	55.45±2.89	60.07±4.44	*60.08±1.53*
	4	46.69±2.74	46.97±2.02	50.82±7.55	54.21±4.17	51.29±1.60	52.74±5.55	**56.48±1.76**	*54.36±1.29*
	5	49.94±3.95	48.56±6.42	**55.07±2.14**	*54.95±3.55*	53.40±4.39	51.41±6.44	51.86±6.61	53.16±2.09
	6	41.78±0.91	43.91±3.11	49.97±8.85	*54.00±3.51*	48.17±5.16	46.14±4.86	**55.75±1.09**	52.31±1.95
	7	46.46±5.68	45.70±5.90	49.77±5.72	48.99±10.49	49.19±1.43	*51.44±3.38*	50.82±5.46	**51.50±1.03**
	8	56.46±0.84	56.53±2.43	56.70±9.76	*62.96±2.32*	58.79±3.03	58.56±4.55	**62.99±1.02**	60.66±1.91
	9	49.89±2.23	49.66±2.16	54.28±2.05	53.92±5.16	53.26±1.64	49.61±3.99	*55.61±2.66*	**55.68±1.15**
	10	54.71±6.15	54.08±6.83	59.33±3.74	*60.27±4.21*	58.73±2.86	58.46±4.70	59.50±6.81	**60.99±1.70**

Table 2. Regression performances of different MTL algorithms. Mean squared error ± std. are presented (lower is better). The three best results are highlighted with **boldface blue**, *italic green*, and plain orange fonts, respectively.

Dataset	Target	Base_1	Base_2	CLMTL [7]	regMTL [4]	MTFL [1]	LRMTL [3]	AMTL [5]	GPMTL [2]	SNS [8]	Ours
RF1	1	21.93±17.51	11.30±2.06	**11.19±0.78**	12.73±0.69	11.71±0.82	11.67±0.64	13.46±4.68	26.28±9.75	15.26±4.68	*11.20±2.13*
	2	0.98±0.36	**0.80±0.20**	0.94±0.11	0.83±0.14	0.85±0.14	1.05±0.12	1.03±0.30	0.84±0.19	*0.81±0.18*	**0.80±0.20**
	3	23.21±25.28	15.82±1.89	*15.13±0.62*	15.82±0.67	15.19±1.19	15.38±0.56	16.43±3.52	25.72±8.56	17.09±4.05	**14.63±1.47**
	4	14.98±4.71	13.41±2.28	*12.52±0.58*	12.80±0.68	12.63±1.13	12.58±0.37	13.06±1.90	16.89±3.97	13.26±2.69	**12.51±1.37**
	5	7.80±0.26	*7.58±0.68*	7.82±0.26	8.40±0.97	7.80±0.53	7.75±0.25	7.77±0.25	10.47±3.97	8.24±1.16	**7.45±0.66**
	6	2.46±0.09	2.35±0.21	2.43±0.07	2.53±0.16	2.56±0.14	2.57±0.07	2.54±0.26	**2.28±0.70**	2.48±0.13	*2.32±0.18*
	7	5.88±0.62	*4.85±1.18*	5.54±0.17	4.90±0.74	4.98±0.81	5.46±0.15	6.05±1.14	7.25±3.35	5.13±1.25	**4.69±1.05**
	8	7.94±8.92	*4.79±0.44*	5.34±0.26	5.45±0.46	5.27±0.23	5.43±0.12	5.24±0.23	6.52±2.69	5.40±0.59	**4.67±0.33**
ENB	1	3.01±0.13	**0.92±0.04**	3.01±0.12	2.22±0.20	1.24±0.10	6.07±0.15	3.01±0.12	*0.96±0.12*	1.07±0.09	**0.92±0.04**
	2	3.26±0.15	1.85±0.15	3.26±0.13	2.44±0.12	1.95±0.20	6.19±0.18	3.26±0.13	**1.76±0.12**	2.01±0.28	*1.82±0.14*

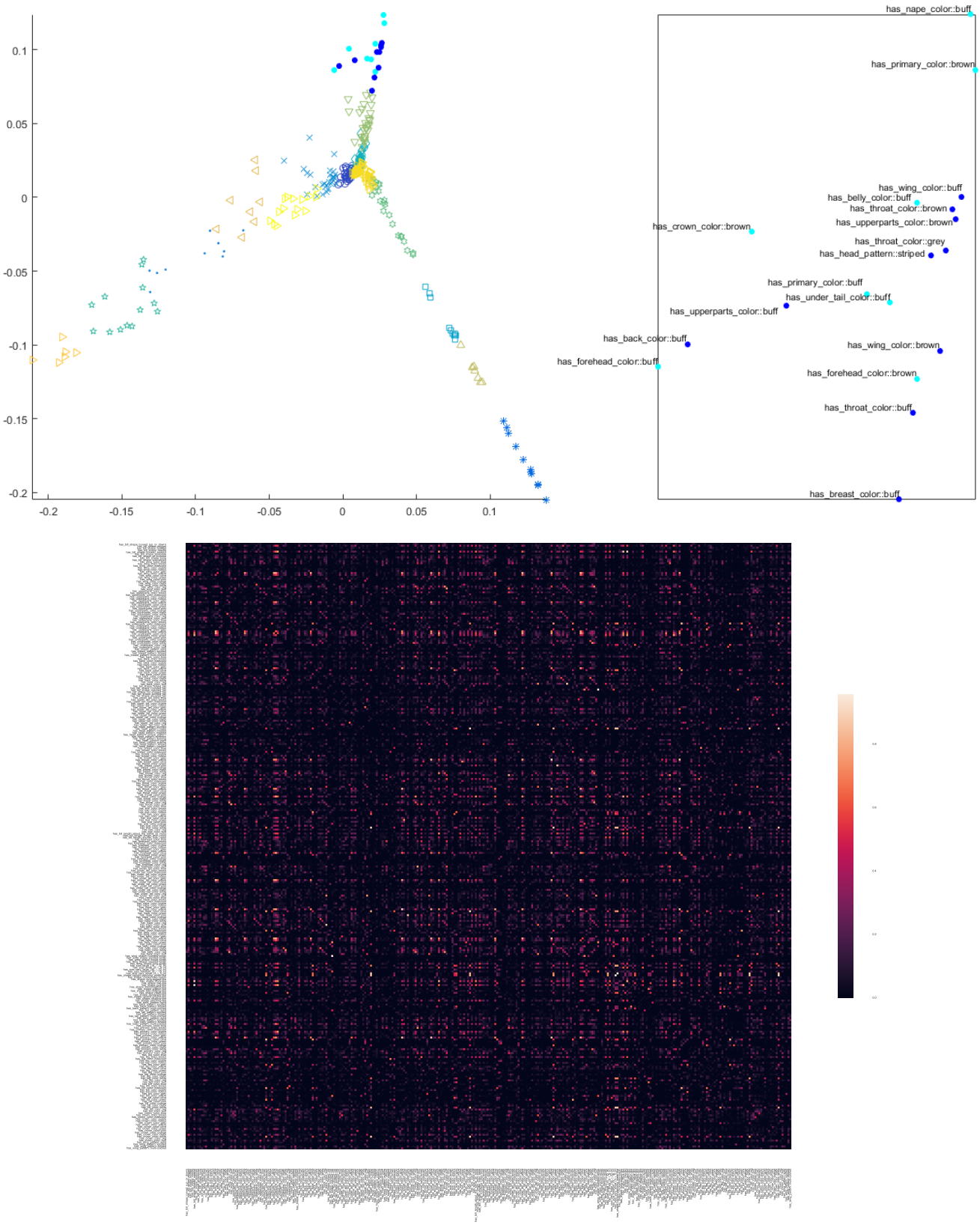

Figure 1. Two-D spectral embedding of 312 tasks in the Birds dataset, and visualization of the FSIC matrix. (Top-left) the color and shape of each entry represents the index of the cluster it belongs to (out of 15 clusters) and (top-right) 18 elements of the first cluster (displayed as solid circles). (Bottom) the FSIC matrix (diagonal entries are set to 0). Readers are advised to check the electronic version of figure.