Supplementary material for submission "From N to N+1: Multiclass Transfer Incremental Learning"

Ilja Kuzborskij[1,2], Francesco Orabona[3], and Barbara Caputo[1]

[1] Idiap Research Institute, Switzerland,
[2] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland,
[3] Toyota Technological Institute at Chicago, USA,
ilja.kuzborskij@idiap.ch, francesco@orabona.com, barbara.caputo@idiap.ch

1. Closed-form Leave-One-Out (LOO) prediction in Multiclass Least-Squares Support Vector Machine (LSSVM)

We follow closely proof given by Cawley [1] with generalization to multiclass scenario.

Additional notation:
$\mathbf{X}$ – Sample matrix, where each column is a sample
$\mathbf{Y}$ – Encoded One-Versus-All (OVA) label matrix, where each label code is a column
$\boldsymbol{A}$ – model parameter matrix, where each model parameters form a column
$\boldsymbol{A}^{(i)}$ – i-th row of a matrix $\boldsymbol{A}$
$\boldsymbol{A}^{(-i)}$ – all, but i-th row of a matrix $\boldsymbol{A}$
$\mathbf{b}$ – transfer parameter vector
In the following we will assume that solution in terms of $\boldsymbol{A}$ and $\mathbf{b}$ is given by solving

$$\begin{bmatrix} \boldsymbol{A} \\ \mathbf{b}^\top \end{bmatrix} = \mathbf{M}^{-1} \begin{bmatrix} \mathbf{Y} - \mathbf{X}^\top \boldsymbol{W}'\boldsymbol{\beta} \\ \mathbf{0} \end{bmatrix} \tag{1}$$

To derive LOO prediction formula, we need to solve Equation 1 when one of $\mathbf{X}$ elements is missing. For this reason we dissect matrix $\mathbf{M}$ as follows (notice r.h.s.)

$$\begin{bmatrix} \mathbf{X}^\top\mathbf{X} + \frac{1}{C}\mathbf{I} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix} = \mathbf{M} = \begin{bmatrix} m_{11} & \mathbf{m_1}^\top \\ \mathbf{m_1} & \mathbf{M_1} \end{bmatrix}$$

Consequently, we recover closed-form solution

$$\begin{bmatrix} \boldsymbol{A}^{(-1)} \\ \mathbf{b}^{\top(-1)} \end{bmatrix} = \mathbf{M_1}^{-1}(\mathbf{Y}^{(-1)} - [\mathbf{X}^{\top(-1)}\boldsymbol{W}' \;\; \mathbf{X}^{\top(-1)}\boldsymbol{W}'\boldsymbol{\beta}])$$

Using parameters $\begin{bmatrix} \boldsymbol{A}^{(-1)} \\ \mathbf{b}^{\top(-1)} \end{bmatrix}$ we obtain prediction on the missing sample

$$\begin{aligned} \tilde{\mathbf{Y}}^{(1)} =& \mathbf{m_1}^\top \begin{bmatrix} \boldsymbol{A}^{(-1)} \\ \mathbf{b}^{\top(-1)} \end{bmatrix} + [\mathbf{X}^{\top(1)}\boldsymbol{W}' \;\; \mathbf{X}^{\top(1)}\boldsymbol{W}'\boldsymbol{\beta}] \\ =& \mathbf{m_1}\mathbf{b}^\top\mathbf{M_1}^{-1}(\mathbf{Y}^{(-1)} - [\mathbf{X}^{\top(-1)}\boldsymbol{W}' \;\; \mathbf{X}^{\top(-1)}\boldsymbol{W}'\boldsymbol{\beta}]) \\ &+ [\mathbf{X}^{\top(1)}\boldsymbol{W}' \;\; \mathbf{X}^{\top(1)}\boldsymbol{W}'\boldsymbol{\beta}] \end{aligned} \tag{2}$$

Noting, that predictions with respect to all, but first element are

$$\begin{aligned} &\begin{bmatrix} \mathbf{m_1} & \mathbf{M_1} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} \\ \mathbf{b}^\top \end{bmatrix} \\ =& \mathbf{M_1}^{-1}(\mathbf{Y}^{(-1)} - [\mathbf{X}^{\top(-1)}\boldsymbol{W}' \;\; \mathbf{X}^{\top(-1)}\boldsymbol{W}'\boldsymbol{\beta}]) \end{aligned}$$

we rewrite Equation 2 as

$$\begin{aligned} \tilde{\mathbf{Y}}^{(1)} =& \mathbf{m_1}^\top\mathbf{M_1}^{-1} \begin{bmatrix} \mathbf{m_1} & \mathbf{M_1} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} \\ \mathbf{b}^\top \end{bmatrix} \\ &+ [\mathbf{X}^{\top(1)}\mathbf{W}' \;\; \mathbf{X}^{\top(1)}\mathbf{W}'\boldsymbol{\beta}] \\ =& \mathbf{m_1}^\top\mathbf{M_1}^{-1}\mathbf{m_1}\boldsymbol{A}^{(1)} + \mathbf{m_1}^\top \begin{bmatrix} \boldsymbol{A}^{(-1)} \\ \mathbf{b}^\top \end{bmatrix} \\ &+ [\mathbf{X}^{\top(1)}\mathbf{W}' \;\; \mathbf{X}^{\top(1)}\mathbf{W}'\boldsymbol{\beta}] \end{aligned} \tag{3}$$

Noting, that, first equation in the System 1 is

$$\mathbf{Y}^{(1)} - [\mathbf{X}^{\top(1)}\mathbf{W}' \;\; \mathbf{X}^{\top(1)}\mathbf{W}'\boldsymbol{\beta}] = m_{11}\boldsymbol{A}^{(1)} + \mathbf{m_1}^\top \begin{bmatrix} \boldsymbol{A}^{(-1)} \\ \mathbf{b}^\top \end{bmatrix}$$

we rearrange and plug $\mathbf{m_1}^\top \begin{bmatrix} \boldsymbol{A}^{(-1)} \\ \mathbf{b}^\top \end{bmatrix}$ into Equation 3 to arrive at

$$\begin{aligned} \tilde{\mathbf{Y}}^{(1)} =& \mathbf{m_1}^\top\mathbf{M_1}^{-1}\mathbf{m_1}\boldsymbol{A}^{(1)} + \mathbf{Y}^{(1)} \\ &- [\mathbf{X}^{\top(1)}\mathbf{W}' \;\; \mathbf{X}^{\top(1)}\mathbf{W}'\boldsymbol{\beta}] - m_{11}\boldsymbol{A}^{(1)} \\ &+ [\mathbf{X}^{\top(1)}\mathbf{W}' \;\; \mathbf{X}^{\top(1)}\mathbf{W}'\boldsymbol{\beta}] \\ =& \mathbf{Y}^{(1)} + (\mathbf{m_1}^\top\mathbf{M_1}^{-1}\mathbf{m_1} - m_{11})\boldsymbol{A}^{(1)} \end{aligned}$$

Expressing $\mathbf{M}^{-1}$ by Schur complement lemma, we observe, that inverse of complement $\mu = m_{11} - \mathbf{m_1}^\top \mathbf{M_1}^{-1} \mathbf{m_1}$ is the first matrix element.

$$\mathbf{M^{-1}} = \begin{bmatrix} \mu^{-1} & -\mu^{-1}\mathbf{m}_1\mathbf{M}_1^{-1} \\ \mathbf{M}_1^{-1} + \mu^{-1}\mathbf{M}_1^{-1}\mathbf{m}_1^\top\mathbf{m}_1\mathbf{M}_1^{-1} & -\mu^{-1}\mathbf{M}_1^{-1}\mathbf{m}_1^\top \end{bmatrix}$$

Combining this fact with insensitivity of system to row-wise permutations, for the i-th sample we have:

$$\tilde{\mathbf{Y}}^{(i)} = \mathbf{Y}^{(i)} - \frac{\boldsymbol{A}^{(i)}}{\mathbf{M}_{ii}^{-1}}$$

2. Supplementary experimental results

Here, additional experimental results are provided for Caltech-256 and Animals with Attributes (AwA) datasets. Results are presented with respect to 5 and 20 unrelated, mixed and related class semantic categories. Accuracies are partitioned:

- by *linear* setting with one feature and *non-linear* setting with multiple features by kernel averaging
- with respect to *no-transfer* and *transfer* baselines
- with respect to N source classes and $N+1$ target class

Each figure showing accuracy on $N+1$ classes, features two smaller figures beneath, corresponding to accuracies on N source classes and $N+1$ target class.

Figure 1 depicts 5-class results for Caltech-256 dataset:

- *Linear* baselines and MULticlass Transfer Incremental LEarning (MULTIpLE), SIFT feature in all cases
- MULTIpLE compared to *no-transfer* baselines in Subfigures 1a – unrelated, 1b – mixed, 1c – related
- MULTIpLE compared to *transfer* baselines in Subfigures 1g – unrelated, 1h – mixed, 1i – related

Figure 2 depicts 5-class results for Caltech-256 dataset:

- *Non-linear* baselines and MULTIpLE, average of Radial Basis Function (RBF) kernels over oriented and unoriented PHOG shape descriptors, SIFT appearance descriptors, region covariance and local binary patterns totalling in 14 descriptor types [2] and RBF hyperparameters $\gamma \in \{2^i : -5 \leq i \leq 8\}$
- MULTIpLE compared to *no-transfer* baselines in Subfigures 2a – unrelated, 2b – mixed, 2c – related
- MULTIpLE compared to *transfer* baselines in Subfigures 2g – unrelated, 2h – mixed, 2i – related

Figure 3 depicts 20-class results for Caltech-256 dataset:

- *Non-linear* baselines and MULTIpLE, feature setting as in Figure 2
- MULTIpLE compared to *no-transfer* baselines in Subfigures 3a – unrelated, 3b – mixed, 3c – related
- MULTIpLE compared to *transfer* baselines in Subfigures 3g – unrelated, 3h – mixed, 3i – related

Figure 4 depicts 5-class results for AwA dataset:

- *Linear* baselines and MULTIpLE, PHOG feature in all cases
- MULTIpLE compared to *no-transfer* baselines in Subfigures 4a – unrelated, 4b – mixed, 4c – related
- MULTIpLE compared to *transfer* baselines in Subfigures 4g – unrelated, 4h – mixed, 4i – related

Figure 5 depicts 5-class results for AwA dataset:

- *Non-linear* baselines and MULTIpLE, average of RBF kernels over SIFT, rgSIFT, SURF, PHOG, RGB color histograms, local self-similarity histograms [3] and RBF hyperparameters $\gamma \in \{2^i : -5 \leq i \leq 8\}$
- MULTIpLE compared to *no-transfer* baselines in Subfigures 5a – unrelated, 5b – mixed, 5c – related
- MULTIpLE compared to *transfer* baselines in Subfigures 5g – unrelated, 5h – mixed, 5i – related

Figure 6 depicts 20-class results for AwA dataset:

- *Non-linear* baselines and MULTIpLE, with feature setting as in Figure 6
- MULTIpLE compared to *no-transfer* baselines in Subfigures 6a – unrelated, 6b – mixed, 6c – related
- MULTIpLE compared to *transfer* baselines in Subfigures 6g – unrelated, 6h – mixed, 6i – related

Figure 7 depicts 50-class results for AwA dataset, where MULTIpLE is compared to transfer baselines and no-transfer settings with features as in Figure 6.

References

[1] G. Cawley. Leave-one-out cross-validation based model selection criteria for weighted LS-SVMs. In *Proc. IJCNN*, 2006. 1

[2] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Proc. ICCV*, 2009. 2

[3] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. CVPR*, pages 951–958. IEEE, 2009. 2

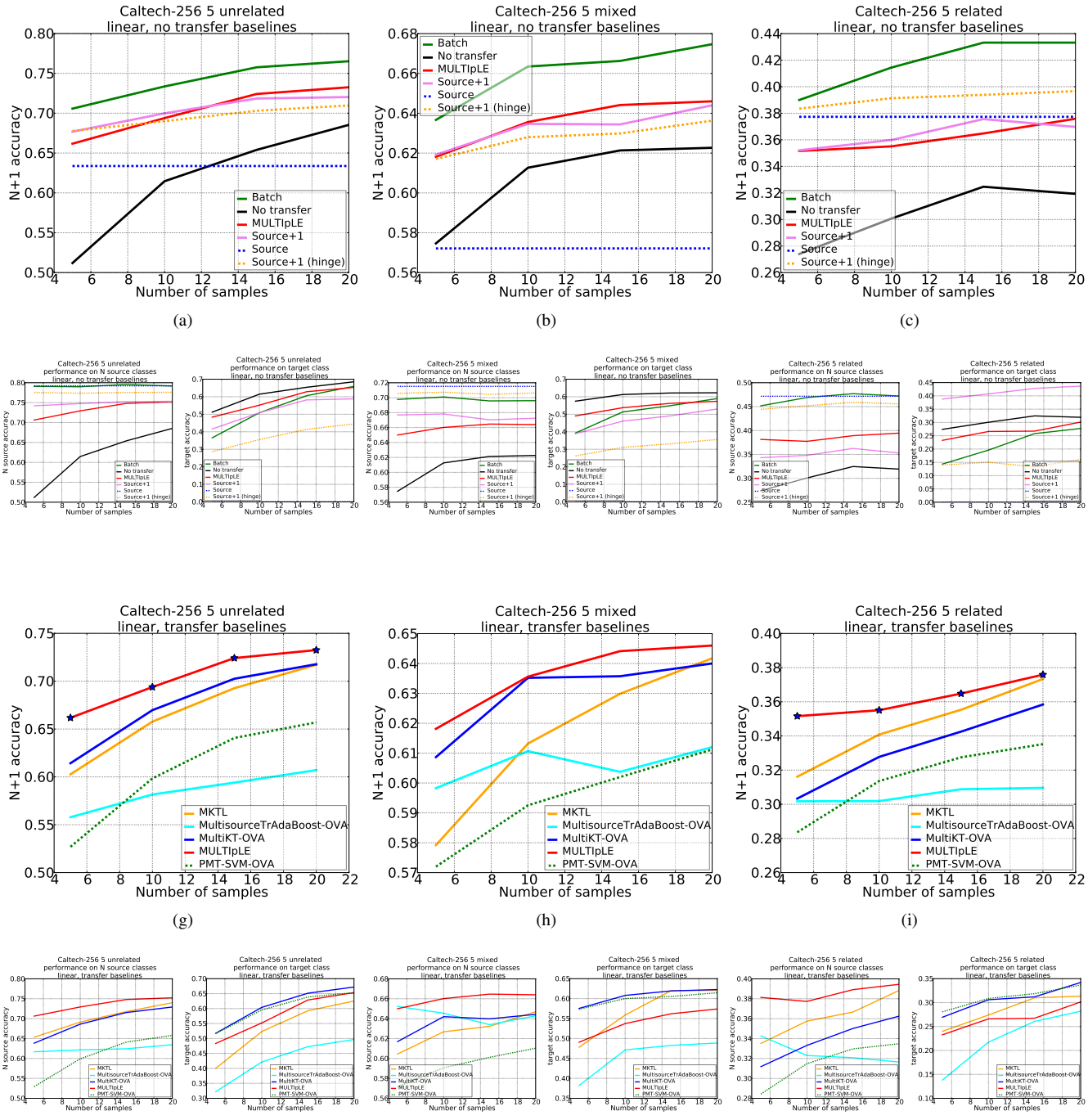

Figure 1: Caltech-256 5 classes linear, no-transfer and transfer baselines

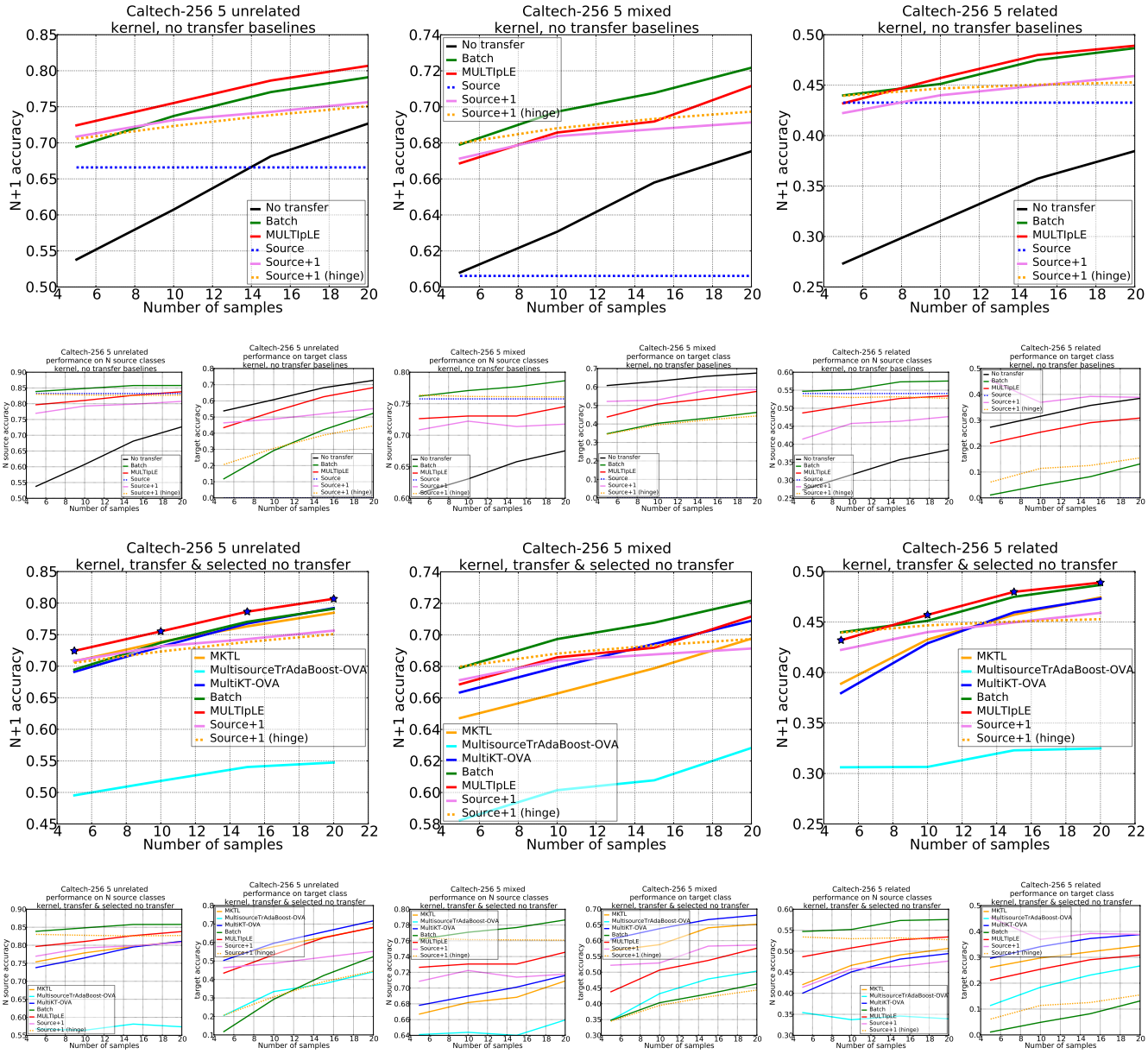

Figure 2: Caltech-256 5 classes, non-linear, no-transfer and transfer baselines

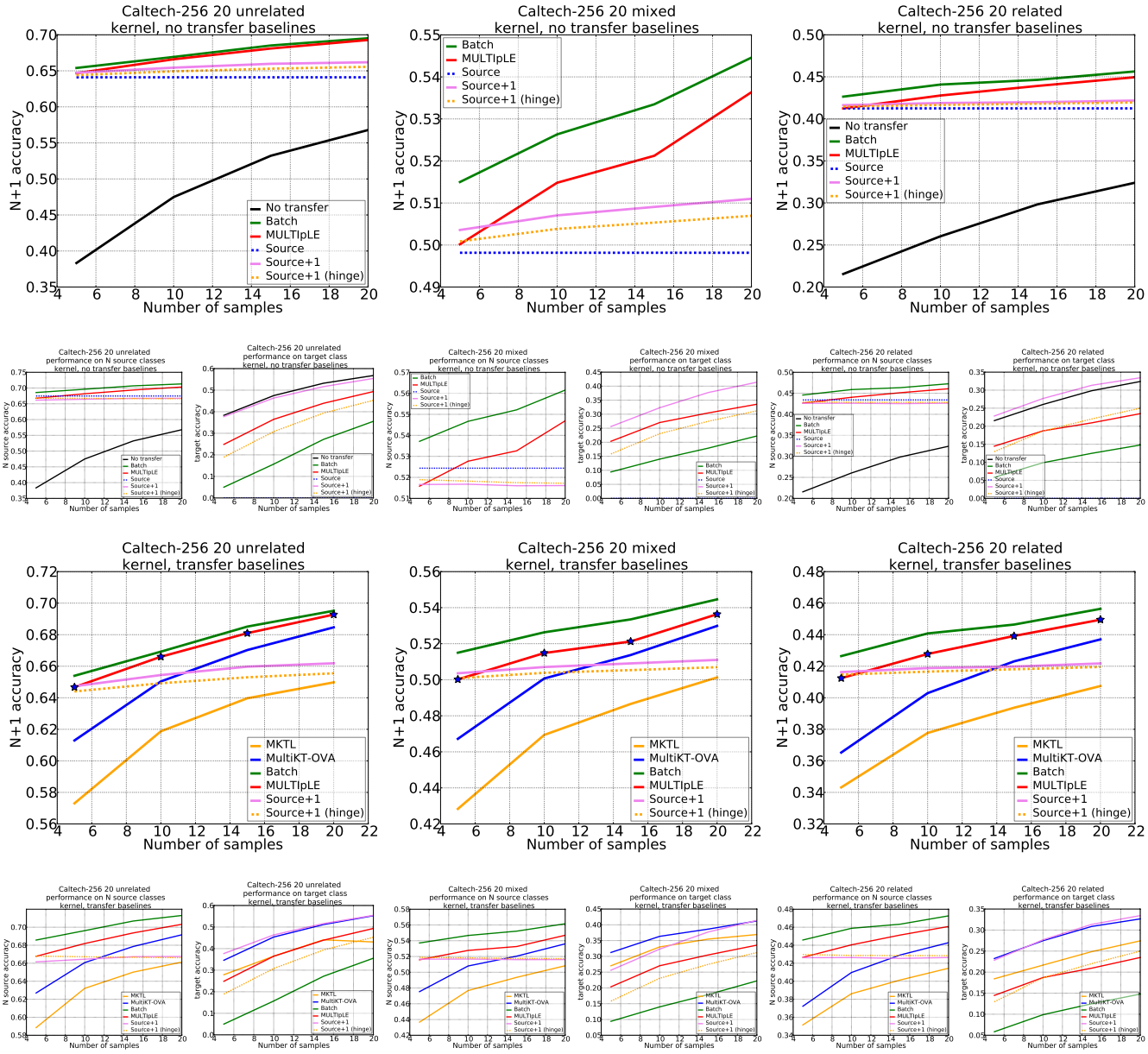

Figure 3: Caltech-256 20 classes, no-transfer and transfer baselines

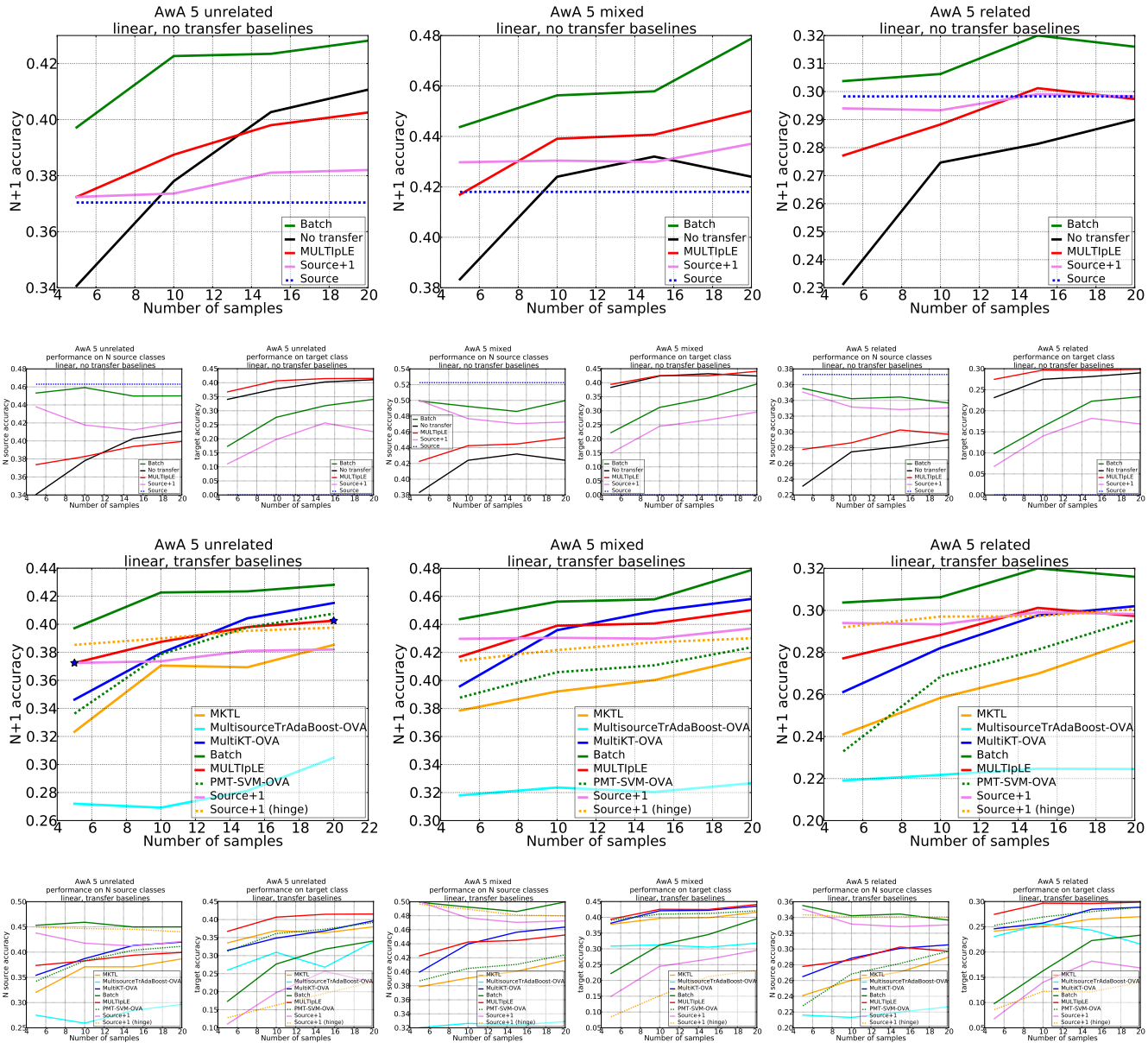

Figure 4: AwA 5 classes, linear, no-transfer and transfer baselines

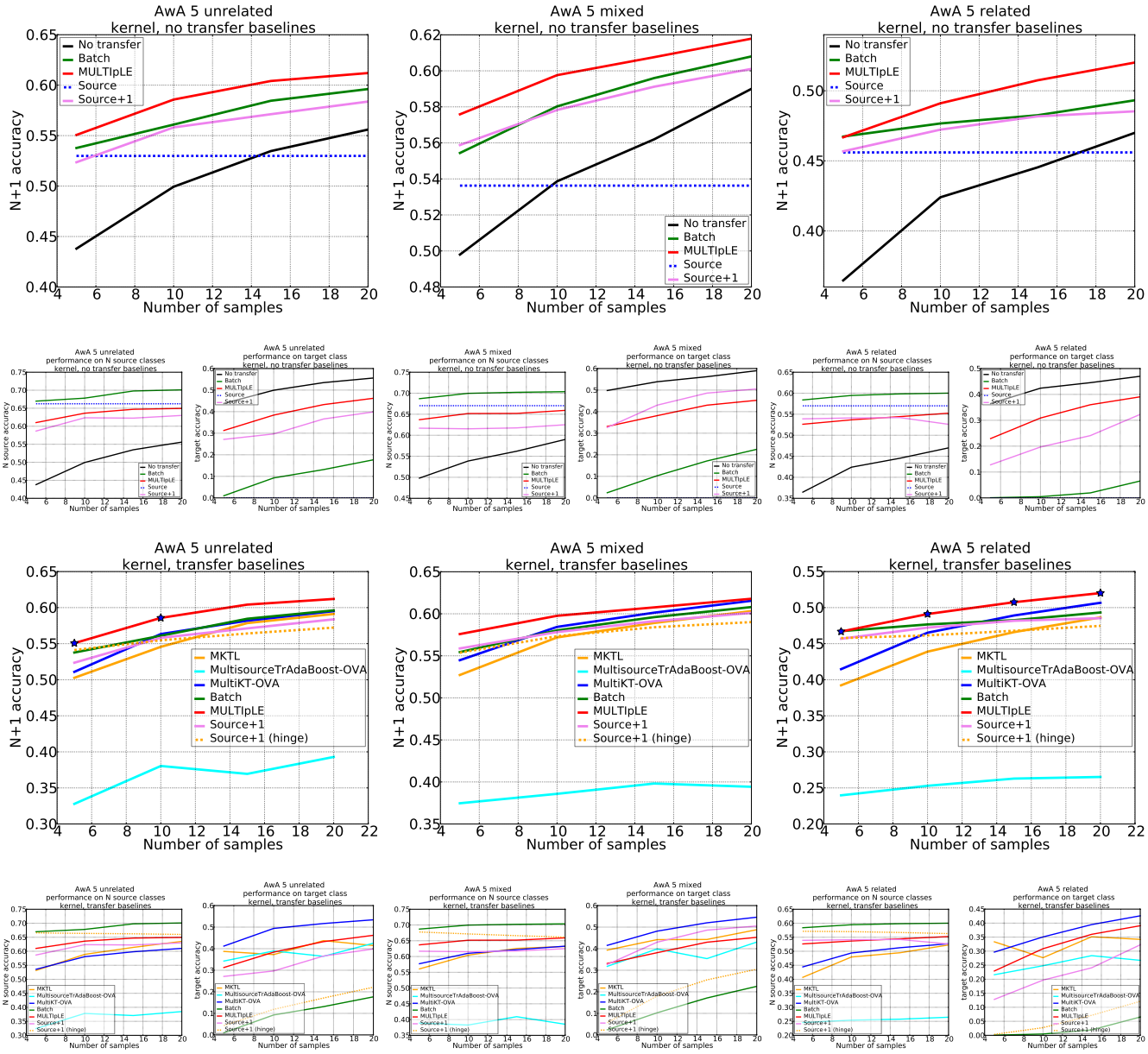

Figure 5: AwA 5 classes, non-linear, no-transfer and transfer baselines

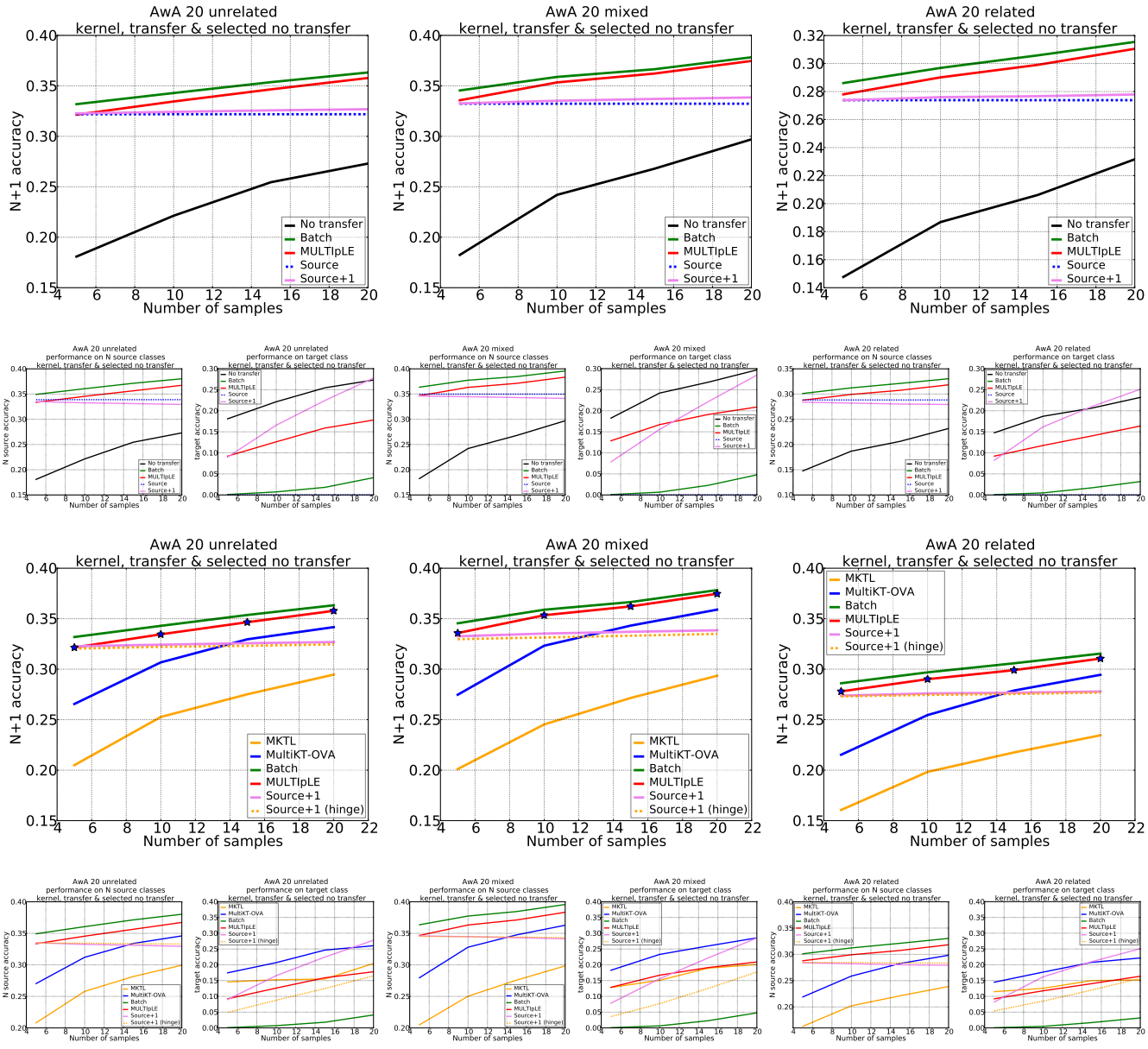

Figure 6: AwA 20 classes, non-linear, no-transfer and transfer baselines

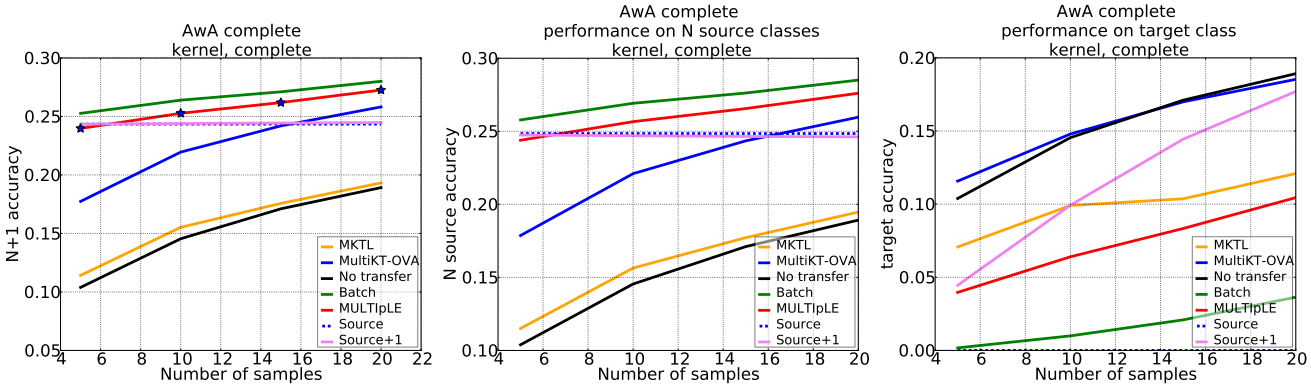

Figure 7: AwA 50 classes, non-linear, no-transfer and transfer baselines