Supplementary Material for
Explicit Inductive Bias for Transfer Learning with Convolutional Networks

A. Effect of L^2-*SP* Regularization on Optimization

The effect of L^2 regularization can be analyzed by doing a quadratic approximation of the objective function around the optimum (see, e.g. Goodfellow et al., 2017, Section 7.1.1). This analysis shows that L^2 regularization rescales the parameters along the directions defined by the eigenvectors of the Hessian matrix. This scaling is equal to $\frac{\lambda_i}{\lambda_i+\alpha}$ for the i-th eigenvector of eigenvalue λ_i. A similar analysis can be used for the L^2-*SP* regularization.

We recall that $J(\boldsymbol{w})$ is the unregularized objective function, and $\tilde{J}(\boldsymbol{w}) = J(\boldsymbol{w}) + \alpha \left\|\boldsymbol{w} - \boldsymbol{w}^0\right\|_2^2$ is the regularized objective function. Let $\boldsymbol{w}^* = \operatorname{argmin}_{\boldsymbol{w}} J(\boldsymbol{w})$ and $\tilde{\boldsymbol{w}} = \operatorname{argmin}_{\boldsymbol{w}} \tilde{J}$ be their respective minima. The quadratic approximation of $J(\boldsymbol{w}^*)$ gives

$$\mathbf{H}(\tilde{\boldsymbol{w}} - \boldsymbol{w}^*) + \alpha(\tilde{\boldsymbol{w}} - \boldsymbol{w}^0) = 0 \ , \tag{1}$$

where $\mathbf{H}$ is the Hessian matrix of J w.r.t. $\boldsymbol{w}$, evaluated at $\boldsymbol{w}^*$. Since $\mathbf{H}$ is positive semidefinite, it can be decomposed as $\mathbf{H} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$. Applying the decomposition to Equation (1), we obtain the following relationship between $\tilde{\boldsymbol{w}}$ and $\boldsymbol{w}^*$:

$$\mathbf{Q}^T\tilde{\boldsymbol{w}} = (\boldsymbol{\Lambda} + \alpha\mathbf{I})^{-1}\boldsymbol{\Lambda}\mathbf{Q}^T\boldsymbol{w}^* + \alpha(\boldsymbol{\Lambda} + \alpha\mathbf{I})^{-1}\mathbf{Q}^T\boldsymbol{w}^0 \ . \tag{2}$$

We can see that with L^2-*SP* regularization, in the direction defined by the i-th eigenvector of $\mathbf{H}$, $\tilde{\boldsymbol{w}}$ is a convex combination of $\boldsymbol{w}^*$ and $\boldsymbol{w}^0$ in that direction since $\frac{\lambda_i}{\lambda_i+\alpha}$ and $\frac{\alpha}{\lambda_i+\alpha}$ sum to 1.

B. Matching the State of the Art in Image Classification

The main objective of this paper is to demonstrate that -*SP* regularization in general, and L^2-*SP* in particular, provides a baseline for transfer learning that is significantly superior to the standard fine-tuning technique. We do not aim at reaching the state of the art solely with this simple technique. However, as shown here, with some training tricks and post-processing methods, which have been proposed elsewhere but were not used in the paper, we can reach or even exceed the state of the art performances, simply by changing the regularizer to L^2-*SP*.

Aspect Ratio. During training, respecting or ignoring the aspect ratio of images will give different results, and usually it would be better to keep the original aspect ratio. In the paper, the classification experiments are all under the preprocessing of resizing all images to 256×256, i.e. ignoring the aspect ratio. Here we perform an ablation study to analyze the difference between keeping and ignoring the ratio. For simplicity, we use the same hyperparameters as before except that the aspect ratio is kept and images are resized with the shorter edge being 256.

Post-Processing for Image Classification. A common post-processing method for image classification is 10-crop testing (averaging the predictions of 10 cropped patches, the four corner patches and the center patch as well as their horizontal reflections).

We apply the aspect ratio and 10-crop testing techniques to improve our results, but we believe the performance can be improved but using additional tricks, such as random rotation or scaling during training, more crops, multi scales for test, etc. Table 1 shows our results. Caltech 256 - 30 outperforms the state of the art; our results in MIT Indoors 67 and Stanford Dogs 120 are very close to the state of the art, noting that the best performing approach (Ge & Yu, 2017) used many training examples from source domain to improve performance. On our side, we did not use any other examples and simply changed the regularization approach from L^2 to L^2-*SP*.

We add Foods 101 (Bossard et al., 2014) to supplement our experiments. Foods 101 is a database that collects photos of 101 food categories and is a much larger database than the three we already presented, yet rough in terms of image quality and class labels in the training set.

C. Application of L^2-*SP* to Semantic Image Segmentation

The paper compares different regularization approaches for transfer in image classification. In this section, we examine the versatility of L^2-*SP* by applying it to image segmentation. Although the image segmentation target task, which aims at labeling each pixel of an image with the category of the object it belongs to, differs from the image classification

Table 1. Average classification accuracies (in %) for L^2 and L^2-*SP* using the training tricks presented in Section B. The source database is Places 365 for MIT Indoors 67 and ImageNet for Caltech 256, Stanford Dogs and Foods. References for the state of the art are taken from Ge & Yu (2017), except for Foods-101 where it is taken from Martinel et al. (2016).

	Caltech 256 - 30	Caltech 256 - 60	MIT Indoors 67	Stanford Dogs 120	Foods 101
L^2	82.7±0.2	86.5±0.4	80.7±0.9	83.1±0.2	86.7±0.2
L^2-*SP*	84.9±0.1	87.9±0.2	85.2±0.3	89.8±0.2	87.1±0.1
Reference	83.8±0.5	89.1±0.2	85.8	90.3	90.3

Table 2. Mean IoU scores on Cityscapes validation set. Fine-tuning with L^2, Chen et al. (2017) obtained 66.6 and 70.4 for ResNet-101 and DeepLab respectively.

Method	L^2	L^2-*SP*
ResNet-101	68.1	**68.7**
DeepLab	72.0	**73.2**

source task, it still benefits from fine-tuning.

We evaluate the effect of fine-tuning with L^2-*SP* on Cityscapes (Cordts et al., 2016), a dataset with an evaluation benchmark for pixel-wise segmentation of real-world urban street scenes. It consists of 5000 images with high quality pixel-wise labeling, which are split into a training set (2975 images), a validation set (500 images) and a test set (1525 images), all with resolution 2048×1024 pixels. ImageNet (Deng et al., 2009) is used as source.

As for the networks, we consider two architectures of convolutional networks: the standard ResNet (He et al., 2016), which can be used for image segmentation by removing the global pooling layer, and DeepLab-V2 (Chen et al., 2017), which stayed top-ranked for some time on the Cityscapes benchmark and is one of the most favored structures. We reproduce them with L^2 and L^2-*SP* on Cityscapes under the same setting.

Most of the training tricks used for classification apply to segmentation, and we precise here the difference. Images are randomly cropped to 800×800, 2 examples are used in a batch, and batch normalization layers are frozen to keep pre-trained statistics. We use the polynomial learning rate policy as in Chen et al. (2017) and the base learning rate is set to 0.0005. For testing, we use the whole image.

Table 2 reports the results on Cityscapes validation set. We reproduce the experiments of ResNet and DeepLab that use the standard L^2 fine-tuning, and compare with L^2-*SP* fine-tuning, all other setup parameters being unchanged. We readily observe that fine-tuning with L^2-*SP* in place of L^2 consistently improves the performance in mean IoU score, for both networks.

References

Bossard, L., Guillaumin, M., and Van Gool, L. Food-101 – mining discriminative components with random forests. In *ECCV*, 2014.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., and Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., and Schiele, B. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Ge, W. and Yu, Y. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *CVPR*, 2017.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. Adaptive Computation and Machine Learning. MIT Press, 2017. URL `http://www.deeplearningbook.org`.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *CVPR*, 2016.

Martinel, N., Foresti, G. L., and Micheloni, C. Wide-slice residual networks for food recognition. *arXiv preprint arXiv:1612.06543*, 2016.