Supplementary Material for: Learning More Universal Representations for Transfer-Learning

Youssef Tamaazousti, Hervé Le Borgne, Céline Hudelot, Mohamed-El-Amine Seddik and Mohamed Tamaazousti

Abstract—This supplementary material of our main paper contain the following: (i) a comparison of our methods to the state-of-the-art according all the universality evaluation-metrics of the literature (Sec. 1); (ii) an evaluation of the impact of more and different grouping SPVs used in our MulDiP-Net (Sec. 2); and (iii) the evaluation of MulDiP-Net with more training data and deeper architectures (Sec. 3).

1 Comparison to State-of-the-Art According Other Universality Metrics

In the Sec. 5 of the main paper, we proposed the mNRG universality evaluation metric and used it for the comparison of our methods with state-of-the-art. However, since our metric is novel, it is important to also perform the same comparison according the *metrics of the literature*, and compare the advantages and drawbacks of each metric. Such comparison is provided in Table 1 and should be analyzed with Table 3 of the main paper, which contain the detailed results on each benchmark. First of all, from the results we see that our MulDiP+FSFT method is the best universalizing method regardless the evaluation-metric. Moreover, still regardless the evaluation-metric, our FSFT is quite promising, since it significantly outperforms SPV_A^{spe}, SPV_G^{gen} and GrowBrain-WA at zero cost of annotation and without adding any additional parameter.

Regarding the comparison of the metrics, we can observe that our mNRG metric respects some of the properties highlighted in Sec. 5 of the main paper (*e.g.*, merit bonus, penalty for damage, penalty malus), which is not the case for the baseline Avg, the RG [1], the VDC [2], the BC and the aNRG (being our method with an average operator instead of the median). For instance, compared to Avg and RG that gives almost the same universality scores to GrowBrain-WA and FSFT, our mNRG is able to give more points to FSFT since it gives significantly better results than GrowBrain-WA on seven of the ten datasets. Moreover, beside the significantly better results on the seven benchmarks, mNRG gives only +0.5 compared to GrowBrain-WA, since it penalizes its loose of performance on the NWO dataset.

We also observe that the VDC do not penalize the methods that performs less than the reference on some benchmarks (*e.g.*, it gives to our MulDiP+FSFT almost *twice* the universality-score than MuCaLe-Net, while compared to the former, the latter never decreases performance in any of the benchmarks), while our mNRG is able to penalize it (MulDiP+FSFT outperforms MuCaLe-Net by only 2.1 points in terms of our mNRG metric). This is even more visible on the ISM method that should clearly give negative universality scores. Indeed, compared to the reference, ISM gives lower results on nine of the ten benchmarks (higher results only on the MIT67 benchmark), but since VDC do not perform *penalty for damage*, it gives 0.0 points on the the nine benchmarks and 0.9 points on the MIT67 one, which undesirably results in a positive universality-score. Moreover, VDC perform neither penalty for damage nor penalty malus, and as a consequence, is unable to say which method between AMECON and WhatMakes performs worse. A metric that has such ability (say method A is worse than B, even if they are both lower than the reference) could be interesting in a case, were for example, the methods A and B have some *practical advantages* compared to the reference and we would like to know which of these practically advantageous methods should be used as a reference for improving universality. Finally, regarding the VDC metric, we observe that compared to the scores (around 2000) reported in their paper, the scores reported in this experiment are much lower (around 100). It is important to note that, this is due to the fact that our evaluation scheme (transfer-learning: training the representation in the source-problem and evaluate on target-problem *unseen during training*) is much more challenging than theirs (end-to-end learning: training the representation in the source-problem and evaluate on the *test-set of the same source-problem*). Simply said, while we do *not* have access to the target-problems during the learning of the representation, they have, making it easier.

We also observe that compared to aNRG, our mNRG is able to decrease the SPV_A^{spe} to a similar universality score than SPV_G^{gen}, since the former seems to be well suited on some datasets like CA101 and CA256 (compared to other absolute improvements). Finally, while not visible here, by construction, the Avg do not provide coherent aggregation. For the same reason, our BC do not provide the same results according the comparison methods, making it not consistent with time. Note however that, as the best universality metric (our mNRG), BC has some good advantages like penalty for damage or independence to outliers.

2 MulDiP-Net with Different and More Grouping-SPVs

Our MulDiP-Net method is based on a grouping SPV using categorical-levels. Here, we assess what is the impact of using different grouping methods. In particular, we compared it to

- *Y. Tamaazousti, is at the CSAIL of MIT, USA. E-mail: ytamaaz@mit.edu*
- *H. Le Borgne, M.E.A. Seddik and M. Tamaazousti are at the CEA, LIST, France. E-mail: firstname.lastname@cea.fr*
- *C. Hudelot is at the MICS laboratory of CentraleSupélec (University of Paris-Saclay). E-mail: celine.hudelot@centralesupelec.fr*

Manuscript received September 2, 2018

Method	Avg	RG	VDC	BC	aNRG	mNRG
REFERENCE	49.2	0.0	0.0	50	0.0	0.0
SPV$^{spe}_{A}$ [3], [4], [5]	50.1	+0.9	18.3	62	+2.3	+1.5
SPV$^{gen}_{G}$ [6], [7], [8]	49.7	+0.5	6.7	56	+1.4	+1.4
AMECON [9]	40.1	-9.1	0.0	17	-20.2	-17.7
WhatMakes [10]	43.8	-5.4	0.0	22	-10.8	-7.5
ISM [11]	45.4	-3.8	0.9	32	-8.8	-4.3
GrowBrain-WA [12]	50.6	+1.4	20.1	71	+3.0	+3.5
GrowBrain-RWA [12]	51.7	+2.5	50.9	87	+5.6	+6.0
MuCaLe-Net [13]	52.3	+3.1	69.6	92	+7.0	+7.7
FSFT (Ours)	50.7	+1.5	36.7	76	+3.0	+4.0
MulDiP+FSFT (Ours)	**53.1**	**+3.9**	**136.9**	**103**	**+8.6**	**+9.8**

TABLE 1
Comparison to state-of-the-art, **according different universality evaluation metrics** (those mentioned in Sec. 5 of the main paper). Note that, for a set of 11 methods and 10 datasets, the best achievable BC score is 110, while the worse is 10.

Method	Network	VOC07 mAP	CA101 Acc.	CA256 Acc.	NWO mAP	MIT67 Acc.	stACT Acc.	CUB Acc.	FLOW Acc.	mNRG
Net-S (Ref.)	AlexNet	71.7	79.7	62.4	58.3	46.9	51.2	**36.3**	58.4	0.0
Net-G	AlexNet	71.5	77.4	60.4	57.8	42.8	49.3	19.5	52.4	-7.7
MulDiP-Net	AlexNet	**74.4**	**82.5**	**65.2**	**60.8**	**47.4**	**54.2**	36.1	**62.5**	+7.4
Net-S	VGG-16	86.1	88.8	78.0	71.8	66.7	73.5	69.8	78.9	+44.8
Net-G	VGG-16	85.7	87.6	76.9	70.3	65.8	72.2	67.0	75.0	+38.9
MulDiP-Net	VGG-16	**87.5**	**92.0**	**80.9**	**72.6**	**68.9**	**75.0**	**71.5**	**81.9**	**+55.3**
Net-S	DarkNet-20	82.7	91.0	78.4	70.5	64.8	72.2	59.5	80.0	+38.9
Net-G	DarkNet-20	83.2	91.5	78.1	73.2	64.4	72.6	52.5	78.9	+40.6
MulDiP-Net	DarkNet-20	**84.1**	**92.7**	**80.1**	**73.9**	**66.4**	**74.5**	**61.2**	**82.1**	+47.1

TABLE 2
MulDiP-Net performances with **different network architectures and more training data**. To compute the mNRG scores (last column in blue), we used the Net-S of AlexNet as reference. All the methods have been learned on the same initial SP (whole ILSVRC).

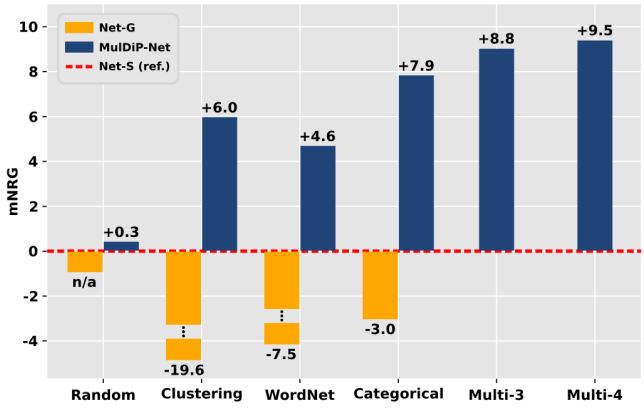

Fig. 1. Impact of the **different grouping SPV and more levels** considered in our MulDiP-Net. Net-S (red dashed line) is used as reference.

grouping based on hierarchical-levels [10] of WordNet, clustering ones [9] and also random. For every grouping method (**Random**, **Clustering**, **WordNet** and our **Categorical**), we also compare MulDiP-Net to each of its subnetworks alone – *i.e.*, the one trained on *specific* classes (**Net-S**) and the one trained on the *generic* ones (**Net-G**). In the main paper, we always used only two levels for fair comparisons, but as depicted in Sec. 4.4 of the main paper, our method could benefit from multiple levels. Thus, we implemented MulDiP-Net with more levels, namely **Multi-3**: initial specific SP and generic SPs obtained from from categorical and Wordnet grouping SPVs; and **Multi-4**: same as Multi-3 with an additional clustering-based grouping SPV. The results are presented in Fig. 1.

From the results, a first observation is that, whatever the grouping SPV, the Net-G is much less performing than Net-S, which contradicts the work of [10] (limited to few domains on target-datasets). Even if below than Net-S, our categorical one is the best grouping SPV, clearly highlighting the interest to introduce a grouping inspired by cognitive studies. Second, whatever the grouping, MulDiP-Net always performs better than its subnetworks (Net-G and especially the reference Net-S), which demonstrates the interest of combining specific and generic knowledge, in the way we do it. Third, in MulDiP-Net, while the best results are achieved with our categorical grouping (confirming its interest), it is worth noting that, the performance of random grouping is very close to Net-S, which highlights the utility of *semantic* grouping SPV. Finally, it is clearly observable that, the more levels we use in MulDiP-Net, the better performance we get.

3 MulDiP-Net with Deeper Networks and More Training-Data

Increasing network capacity (*wider* or *deeper* layers) can be a very efficient universalizing method, since it can learn to perceive more elements or configuration through its new features. However, it is important to note that, it is not easy to modify the architecture (many costly experiments are needed to set all the hyper-parameters as well as the architecture itself) and no certainty of convergence is promised. In all cases, our contribution is orthogonal to this domain, and our aim here is to demonstrate this orthogonality. To do so, we implemented the reference, as well as our MulDiP-Net method with three popular architectures, namely the basic AlexNet (5 convolutional and 2 fully-connected layers), the deep and wide VGG-16 (16 convolutional and 2 fully-connected layers) and the fast and very-deep DarkNet-20 (20 convolutional layers followed by average pooling). Another important question is whether our approach of learning from a fixed set of training data could benefit from more data if they are available (adding-data approach. Thus, in this experiment, instead of using ILSVRC* (containing half-million images and 483 categories) as the initial source-problem, we used the whole ILSVRC which contains 1.2M images and 1K categories. The results of these experiments are presented in Table 2.

Four observations can be made. First, even with twice more data than in Table 3, MulDiP-Net still significantly increases universality compared to the reference. This demonstrates the orthogonality of our approach with the works that adds more data (domains [2], [4], [14] or tasks [1]). Second, the deeper architecture do not learn the more universal representation (Net-S with VGG-16 is better than Net-S with DarkNet-20). This clearly highlights that, compared to diversifying the source-problem, naively increasing the capacity is not safe for improving universality. Third, we clearly observe that MulDiP-Net outperforms its subnetworks regardless the architecture, which demonstrates that our approach could benefit from the field of network architectures. Last but not least, we can observe that Net-G is always below Net-S, except for DarkNet. This is surprising since one could have the intuition that the finer categories we use for training, the better results we get. However, it seems that this depends on the architecture, or maybe on the ratio between the number of units in the representation and the number of classes used for training.

References

[1] S. Subramanian, A. Trischler, Y. Bengio, and C. J. Pal, "Learning general purpose distributed sentence representations via large scale multi-task learning," in *ICLR*, ser. ICLR, 2018.

[2] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Learning multiple visual domains with residual adapters," in *NIPS*, 2017.

[3] H. Azizpour, A. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *PAMI*, 2015.

[4] H. Bilen and A. Vedaldi, "Universal representations: The missing link between faces, text, planktons, and cat breeds," *arXiv:1701.07275*, 2017.

[5] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[7] P. Mettes, D. Koelma, and C. G. M. Snoek, "The imagenet shuffle: Reorganized pre-training for video event detection," in *ICMR*, 2016.

[8] Y. Tamaazousti, H. Le Borgne, A. Popescu, E. Gadeski, A. Ginsca, and C. Hudelot, "Vision-language integration using constrained local semantic features," *CVIU*, 2017.

[9] I. Chami, Y. Tamaazousti, and H. Le Borgne, "Amecon: Abstract meta concept features for text-illustration," in *International Conference on Multimedia Retrieval*, ser. ICMR, 2017.

[10] M. Huh, P. Agrawal, and A. A. Efros, "What makes imagenet good for transfer learning?" *arXiv:1608.08614*, 2016.

[11] Y. Wu, J. Li, Y. Kong, and Y. Fu, "Deep convolutional neural network with independent softmax for large scale face recognition," in *ACM*, 2016.

[12] Y.-X. Wang, D. Ramanan, and M. Hebert, "Growing a brain: Fine-tuning by increasing model capacity," in *CVPR*, 2017.

[13] Y. Tamaazousti, H. Le Borgne, and C. Hudelot, "Mucale-net: Multi categorical-level networks to generate more discriminating features," in *CVPR*, 2017.

[14] S.-A. Rebuffi, H. Bilen, and A. Vedaldi, "Efficient parametrization of multi-domain deep neural networks," in *CVPR*, ser. CVPR, 2018.