Report
on
Multi-Source Transfer Learning From Pre-Trained Networks

Ankit Grover

23rd January 2023

1 Paper Report

In this section of the report I will briefly summarize the paper [Lee et al., 2019] including it's various advantages, disadvantages,etc.

1.1 Motivation

The given paper tackles the problem of generalization using Multi-source transfer learning approach in which pre-trained models trained on multiple different source tasks are being leveraged to a completely different target task to be achieved.
The reasons for using a Multi-Source Transfer Learning approach with MCW algorithm is explained below:

- Our application/target task is resource scarce.
- We do not have direct access to the source training-data and,or resources for training using the same.
- Methods such as Multi-task learning,Model Agnostic Meta-learning procedures cannot be applied as we are dealing with a single task and cannot perform joint training.

1.2 Methodology

We essentially have a Source Network M which is trained on multiple different data sets each which are exclusive of each other i.e no intersection between class labels of source tasks and target tasks. We only have access to M which is a "black-box function" and target data sets. So given N models trained on N

source datasets (same task such as Binary Classification) we need to formulate a method to use our N models for our target task.

The authors work on a Image Classification task here and used multiple **datasets** such as $CIFAR-100$, $StanfordDogs$,etc. The authors chose a simple $LeNet$ architecture which consists of CNN,Pool,ReLu activations followed by 2 Fully Connected layers before the final FC layer. The network is divided into 2 blocks namely Net1() and Net2() both of which are combined for training on the source datasets. The authors use the Net1()'s feature map outputs with 84 activations vectors. Thus, we now have N diff. model's each trained on different source datasets with different(and disjont) labels(but same task) whose features can be combined together to make a new classifier for our target task. We use principles of Joint Probability and Expectation Maximization to calculate the correlation function g which maximizes the correlation σ given our featur-mapping function f^*. This g^* is essentially serves to map our feature outputs to labels. Thus, we are trying to find the correlation function g^* which can map the feature inputs to labels, and an associated maximal correlation σ which serves as the weights for each of the N source models.

2 Advantages & Disadvantages

The given formulation has some advantages and disadvantages which are discussed below:

Advantages

- **Explain ability:** Inspecting correlation weights σ of the source models can give an insight into the contribution of each source model/data set into the target task.
- **Low-resource scenario:** The method show generalization capabilities for *few-shot generalization* with less examples.
- **Federated Learning:** The method resembles *Federated Learning* strategies, thus can be used for decentralized scenarios

Disadvantages

- **Data reconstruction:** It is possible for adversaries to try and regenerate training data from the model features or through imitating the model outputs.
- **Zero-Shot Generalization:** The method resembles would fail when unlabelled data is present , hence we cannot find our classifier weights,correlations.
- **Data/Model-drift:** Relying on correlations, if the source training data and hence the model's output distributions differ widely from the target task then correlation weighting would fail.

3 Reproducing Results

In the given reproduction of results 3 random seed values of 3,42,142 were used and results were averaged over the same. Results for Single-Source models are not provided due to insufficient information for deciding best model.

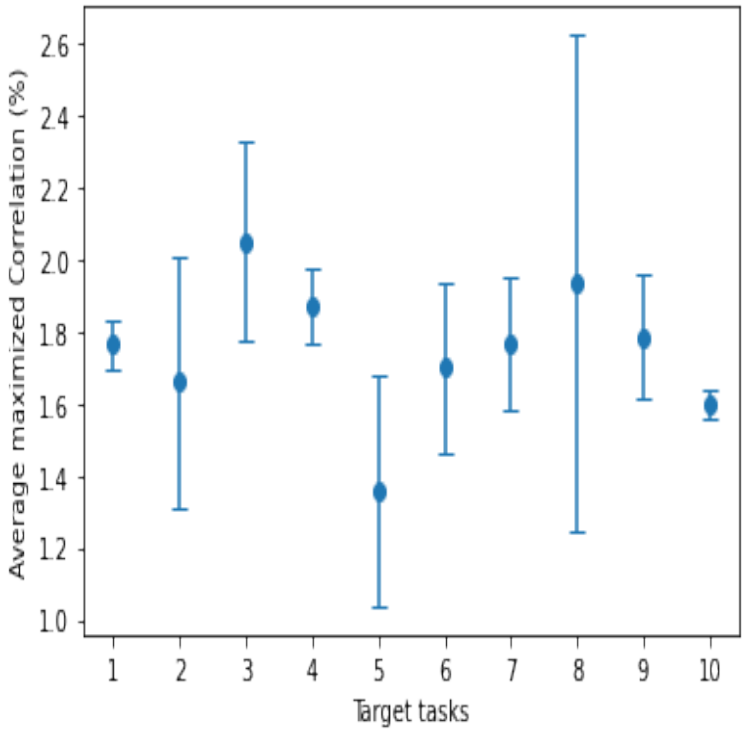
Compared to the results in the original paper, accuracy and std-dev vary due

Table 1: Generalization accuracy of Mult-Source models

Method	1-shot	5-shot	10-shot	20-shot acc
Multi-Source SVM	63.67+4.35%	56 + 8.6%	76 +8.3%	70.57 + 0.0032%
Multi-Source MCW	74+2%	75.67 +1.43%	78 +1.33%	76.33 +0.023%

to number of seeds used not being mentioned in the original paper. Moreover. *20-shot accuracy* results strangely less than 10-shot. This could be explained as either over-generalization or due to the seeds.One thing to note is how the deviation reduces with more samples denoting a better fit. Now, using 3 seeds we plot the σ corresponding to the associated maximal correlation weights of each source task.

Figure 1: Associated maximal correlation weights for each 10 source tasks.



Compared to the original results, we can see the correlation for 5th task is low which corresponds with the original plot, however the task 8 shows higher correlation compared to the 9th task . This can be explained with the fact that due to the results shown on only **3 seeds**, deviation is high as can be seen that error-bars of 8th task are a lot higher than those of 9th task.

As such, in the conventional SSMT pipeline, where all components are independent, these prosodic features cannot be transferred directly from the source speech signal and added back into the output speech signal. Thus, the prosodic features in the output speech signal have to be "learned" from large corpora and represented in the speech model of the TTS system.

References

[Lee et al., 2019] Lee, J., Sattigeri, P., and Wornell, G. (2019). Learning new tricks from old dogs: Multi-source transfer learning from pre-trained networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 4372–4382. Curran Associates, Inc.