Self-Supervised Learning using Barlow Twins in Speech Recognition

Tarek Ayed
KTH
EECS, MSc Machine Learning
ayed@kth.se

Abstract

Self-supervised learning (SSL) is a common technique where synthetic tasks allow to reach high performance with low amounts of labelled data. In some of these methods, the task revolves around teaching the model representations that are invariant to distortions. Whereas most SSL techniques resort to implementation tricks to prevent representational collapse (to trivial constant solutions), Barlow Twins (BT) (Zbontar et al. [2021]) is a method that uses the cross-correlation matrix of the learned representation vector as an objective, in order to maximize invariance and minimize redundancy. The method was originally introduced for image data. In this project, we show that BT can be implemented for speech data as well, by taking the example of an ASR task on the TIMIT (Garofolo et al. [1993]) dataset. If applied to the entire model, BT allows to reach superior performance when training on a small number of labelled samples - 79.3 PER compared to 89.8 for the supervised baseline.

1 Introduction

Properly labelled speech recognition data has notoriously been difficult and expensive to acquire. However, supervised deep learning speech recognition models have followed the general trend of gaining in complexity and parameter count (Henighan et al. [2020]). This means that more labelled data is needed, if trained in a fully supervised manner.

Self-supervised learning is a technique consisting of using synthetic tasks to pre-train a model without any labelled data. One use of such techniques is to be able to learn tasks with fewer labelled examples, or to gain in overall performance. Popular techniques include SimCLR (Chen et al. [2020]), BYOL (Grill et al. [2020]) and SwAV (Caron et al. [2021]). SSL has also successfully been applied to speech data and ASR tasks (Baevski et al. [2020], Zhang et al. [2020]). Recently, Wav2vec (Baevski et al. [2020]), a contrastive SSL technique, achieved a new state-of-the-art on the TIMIT dataset (Garofolo et al. [1993]) and showed that good ASR performance can be achieved with as little as 10 minutes of labelled speech data.

In May 2021, Zbontar et al. [2021] introduced an extremely simple new Self-supervised learning method, called Barlow Twins. The idea is to randomly perturb a batch twice and then use the cross-correlation matrix of the projected vectors as an objective, by aiming at same-feature correlations of 1 and other correlations of 0. The intuition behind this objective is to learn meaningful representations while minimizing redundancy.

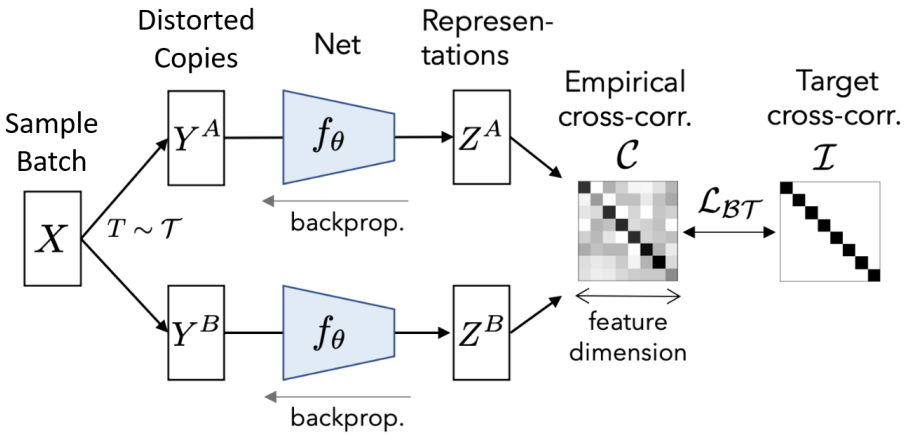

Figure 1: Overview of the Barlow Twins (Zbontar et al. [2021]) Self-supervised learning method. Adapted from the original paper.

2 Method

2.1 Barlow Twins Self-Supervised Learning

An overview of the method is given in figure 1. The general idea is to apply random distortions on a batch of samples. Two distorted copies of this batch are then fed into the main network. A projector is then applied to each of the two outputs to map them to fixed size representation vectors Z_A and Z_B. Finally, the two vectors are normalized across the batch to have zero mean and unit variance. An empirical cross-correlation matrix $\mathcal{C}$ is computed, i.e. $\mathcal{C} = Z_A^\top Z_B$.

The goal is to force the network to learn meaningful low-dimension representation of the data. This is enforced by targeting a 1 correlation between same-feature values. But we also want this representation to be non-redundant, as this would allow it to encode as much information as possible. This is the rationale behind wanting different components of the representation vectors to be uncorrelated.

$$\mathcal{L}_{\mathcal{BT}} = \sum_i (1 - \mathcal{C}_{ii})^2 + \lambda \sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2$$

The final loss function is, thus, a sum of an *invariance term* and a *redundancy reduction term*, as laid out in the original paper (Zbontar et al. [2021]). A natural trade-off arises between these two terms and a coefficient λ is added to make sure that the redundancy reduction term which contains a lot more terms in the sum does not dominate the invariance term, which is crucial for a successful training.

2.2 Adaptation to Speech Data

The Barlow Twins method was originally intended and tested for image data. This project attempts to reproduce the results for speech data. In order to do so, multiple choices and adaptations need to be made.

First of all, the original paper's experiments (Zbontar et al. [2021]) showed that the results, although impressive overall, are highly sensitive to the choice and richness of the distortions applied, which the authors identify as a key difference between Barlow Twins and other popular SSL methods. This is the first major choice to be made when adapting the method to Speech Recognition.

A second design choice is the level at which to project the encoded features of the sequence into a representation vector. This can either be done at the end of the sequence (using the output of the last time-step as input for the projector head) or at all time-steps (each sequential output is projected into a different representation vector, and the loss function is computed on each one of them).

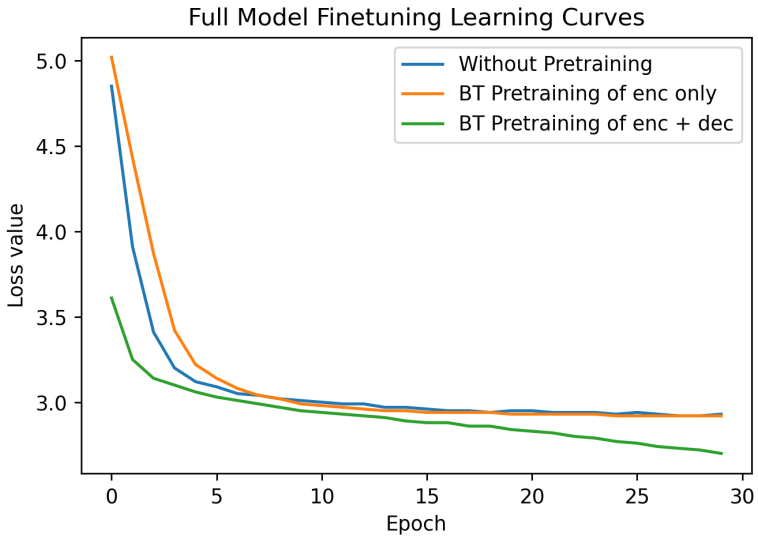

Figure 2: Loss curves comparison between with- and without Barlow Twins pre-training. The green plot corresponds to Linear Probing training (the encoder being frozen). The two others correspond to full-model fine-tuning.

The first method is easiest to implement and test, whereas the second method could be more effective at forcing the network to learn useful representation at every time-step (Zhang et al. [2020], Baevski et al. [2020]). It, however, implies time alignment issues as some of the distortions may induce time delays, which can be challenging to solve, and may hurt the overall training process.

3 Experiments

3.1 Dataset

For the purpose of this report, I chose to conduct my experiments on the TIMIT (Garofolo et al. [1993]) dataset, which consists of recordings of 630 speakers, each reading 10 phonetically-rich American English sentences. Recordings are labelled at the phoneme level. The only pre-processing applied is a Fast Fourier Transformation, along with a Mel scale.

As stated earlier, the choice of distortions applied to the data has a high impact on the end performance (Zbontar et al. [2021]). I chose to use two augmentation approaches simultaneously (i.e. applying one after the other on every batch). The first is SpecAugment (Park et al. [2019]), which consists of time-masking, frequency-masking and re-sampling. The second one was to apply standard environmental corruption (noise, reverb and babble), as implemented in the SpeechBrain Python library (Ravanelli et al. [2021]).

3.2 Choice of model

For experimentation purposes, the model I used is a combination of Convolutional layers, Recurrent layers and fully connected decoding layers. Implemented in the SpeechBrain library (Ravanelli et al. [2021]) with the "CRDNN" module, the model specifically applies 2 CNN layers to the Mel spectrum, followed by 2 RNN layers and 2 linear layers. These constitute the *encoder* block. The output of this block are phoneme-level probabilities that are fed into the *decoder* block which, in this experiment, is an Attentional RNN. The role of this layer is to interpret the previous output and curate the final predictions of the model. For Barlow Twins pre-training, the projector head used is constituted of two linear layers. The hidden layer has 2048 size and the output has a 512 size, which is the size of the representation vector used in the BT loss function.

Lastly, the pre-training step can either be applied on the encoder block alone, or on the encoder and the decoder combined.

3.3 Training

As a first step, I implemented the first method described in paragraph 2.2 of this report, which is to use the output of the last time-step as an input for the projector head. I reduced the size of the model to accommodate compute constraints. In total, the model has 2.1 million trainable parameters. I also reduced batch size to 64 for pre-training and to 8 for fine-tuning. Although batch size is critical for successful SSL training, the authors in Zbontar et al. [2021] suggest that Barlow Twins is relatively robust even at smaller batch sizes. I pre-trained for 50 epochs and fine-tunes for 30 more epochs. After experimenting with the LARS optimizer used by Zbontar et al. [2021], I switched for Adam, as it performed similarly while being significantly faster.

3.4 Evaluation

To assess the added value of SSL, the evaluation method is to fine-tune the pre-trained model on a small number of labelled samples (12% of the training data) and compare the performance to that of the same model without any pre-training. Fine-tuning was both conducted on the entire model, and by freezing the encoder's weights. At the fine-tuning and evaluation stage, a Cross-Temporal Classification loss was used at the phoneme level. Inference was based on Beam Search at testing. Phoneme Error Rate (PER) and validation loss are used as evaluation metrics.

4 Results

4.1 Pre-Training the Encoder

Empirical results when pre-training the encoder block of the model alone are shown in table 1 and the corresponding learning curves are in figure 2. There is no visible difference in these experiments between pre-trained models using the Barlow Twins method, and randomly initialized ones. It seems even that when, fine-tuning the entire model, pre-training hurts downstream performance.

The loss curves on figure 2 are almost identical, although it seems to be converging slightly quicker without SSL pre-training.

Evaluation Method	PER With BT Pretraining	PER Without Pretraining
Full Model Fine-tuning	97.3	89.8
Linear Probing	89.0	89.0

Table 1: Performance of Barlow Twins Pre-training of the encoder block, compared to same-model no pre-training baseline without pre-training, measured by Phoneme Error Rate (PER).

4.2 Pre-Training the Entire Network

Contrasting with the previous results, when pre-training the entire model (encoder block and decoder block), Self-Supervised pre-training using the Barlow Twins method seems to bring a significant performance improvement. The improvement is also visible in figure 2 where the green plot is lower than the other two and reaches lower terminal loss values. The full-model finetuned version of the Barlow Twins pre-trained model is lacking from table 2. This experiment was skipped because it was deemed less relevant, and because of time and computing resources constraints.

Evaluation Method	PER With BT Pretraining	PER Without Pretraining
Full Model Fine-tuning	–	89.8
Linear Probing	**79.3**	89.0

Table 2: Performance of Barlow Twins Pre-training of the encoder and decoder blocks, compared to same-model no pre-training baseline without pre-training, measured by Phoneme Error Rate (PER).

5 Discussion and Conclusions

This experiment clearly shows that Barlow Twins allows superior performance when training on a small number of labelled samples. The difference in final performance and in training trajectories

shows that the representations learned during Barlow Twins SSL pre-training are indeed meaningful and useful in an ASR task.

It is unclear whether this implies Barlow Twins would improve overall performance when training supervised models on all labelled data. More importantly, it is unclear whether the advantage of this SSL method pre-trained models would subsist when training on full-scale state-of-the-art ASR models. It is possible that this SSL technique's effectiveness is aided by the small size of the model.

However, it should also be said that the experimental results from table 2 are obtained after only 50 epochs of pre-training and 30 epochs of supervised fine-tuning. Moreover, in figure 2, the green curve seems to not have plateaued yet at the end of training. In the original Barlow Twins paper (Zbontar et al. [2021]), some experiments went as far as 1000 epochs in the pre-training part alone. Additional experimentation needs to confirm the findings of this project.

A second major takeaway is the difference between pre-training the encoder only, and pre-training everything except the final linear layers. This may be surprising and hard to grasp intuitively. However, other speech data SSL methods (Baevski et al. [2020]) rely on the approach of training everything but the final linear projector. This could be explained by the existence of low-level dependencies between the encoder and the decoder that can only be trained together, rendering an "encoder-only" features virtually useless in the fine-tuning phase.

References

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction, 2021.

J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus cdrom, 1993.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. Scaling laws for autoregressive generative modeling, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations, 2020.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments, 2021.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

Yu Zhang, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V. Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition, 2020.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, Sep 2019. doi: 10.21437/interspeech.2019-2680. URL `http://dx.doi.org/10.21437/Interspeech.2019-2680`.

Mirco Ravanelli, Titouan Parcollet, Aku Rouhe, Peter Plantinga, Elena Rastorgueva, Loren Lugosch, Nauman Dawalatabad, Chou Ju-Chieh, Abdel Heba, Francois Grondin, William Aris, Chien-Feng Liao, Samuele Cornell, Sung-Lin Yeh, Hwidong Na, Yan Gao, Szu-Wei Fu, Cem Subakan, Renato De Mori, and Yoshua Bengio. Speechbrain. `https://github.com/speechbrain/speechbrain`, 2021.