International Journal for Modern Trends in Science and Technology, 8(06): 541-544, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0806092
Available online at: http://www.ijmtst.com/vol8issue06.html

Image Segmentation Evaluation of Data2Vec on Cityscapes and IDD

Mritunjay Musale | Sheetal Pereira

K. J. Somaiya College of Engineering, Somaiya University, Mumbai, India
*Corresponding Author : mritunjay.m@somaiya.edu

To Cite this Article
Mritunjay Musale and Sheetal Pereira. Image Segmentation Evaluation of Data2Vec on Cityscapes and IDD. International Journal for Modern Trends in Science and Technology 2022, 8(06), pp. 541-544. https://doi.org/10.46501/IJMTST0806092

Article Info
Received: 22 May 2022; Accepted: 18 June 2022; Published: 24 June 2022.

ABSTRACT

The advancements of Self-Supervised Learning have gained traction in recent years, across all modalities. Data2Vec is one of the approaches used in SSL which provides a unique stance where it can be finetuned on across multiple modalities without changing anything specific within the framework. In this work we finetune Data2Vec on autonomous driving cars dataset with respect to image segmentation task. We observe the effect of learning rate schedulers and half precision data types on the framework's performance.

KEYWORDS: ***Self-supervised Learning, Cityscapes, India Driving Dataset, Deep Learning***

1. INTRODUCTION

Self-supervised Learning (SSL) is an approach of extracting features from data with little to no supervision, with the help of recent advancements in Deep Learning. This is usually done through projection heads[1] or using momentum based encoders[2]. SSL itself relies on the idea of training using energy based functions, where corrected and incorrect information is treated like a manifold learning problem[3].

Data2Vec[4] is one such approach that uses transformers and is pre-trained in an SSL manner, which can be later fine tuned to a particular task. The authors of Data2Vec demonstrate that the framework can be used for speech, text and images without changing the internals of the framework and still get competitive results. We use Data2Vec for fine tuning on two selected autonomous cars dataset, the evaluation in this work is done with respect to performance on image segmentation task.

2. RELATED WORK

SSL has proven to be superior in performance in natural language processing (NLP) when applied using transformer based approaches[5], [6],[7]. Transformers have been made to work with image related tasks by modifying the input[8],[9], but those are trained in a supervised manner. Before transformers were applied to vision, SSL for vision was done using ConvNets based approaches, like [2] where a momentum based encoder was used for learning features by using modified dictionary based loss. Contrastive learning in SSL also achieves competitive results when training in unsupervised manner by means of using a projection head along with an initial set of convolutional layers

called as the encoder as demonstrated in [1].

iGPT[10] was one of the first implementations of images applied to transformers in an unsupervised manner. Since then there have been several improvements in training transformers in unsupervised manner, since then there have been several implementations of doing the same with improvements[11],[12]. Image segmentation tasks have also been applied to transformers for netting effective results[13],[14]. We use Data2Vec since it combines the idea of masked prediction while learning latent representation in an SSL manner, that can be evaluated later during supervised finetuning.

3. METHODOLOGY

In this work we use the HuggingFace[15] implementation of the Data2Vec model. Specifically, Data2VecVisionForSemanticSegmentation class is considered while working on segmentation tasks of the chosen dataset. We will be fine-tuning and evaluating the models on the Cityscapes[16] dataset and India Driving Dataset(IDD)[17]. The target outputs and specifications of the dataset are discussed in the section. We also discuss the training setup and experimentation in this section.

3.1 Pretrained Model

We use a pretrained Data2Vec model provided by the HuggingFace transformers library, similar to that defined in the Data2Vec[4]. The model has been pre-trained on Imagenet dataset[18], which consists of 1.2 million images with the resolution of 224x224. The HuggingFace implementation adds UPerNet[19] and fully convolutional layers at the end to make the model work with segmentation tasks.

3.2 Datasets

In this work we are interested in segmentation performance of the model since transformers recently have performance on par with their convolutional counterparts[20],[14]. In order to do this we will be utilizing the Cityscapes and IDD datasets for evaluating Data2Vec performance after fine-tuning, the specifications of these Datasets is discussed below.

The Cityscapes[16] dataset is a collection of high quality images taken from 50 cities across Germany. The images are taken in varying weather conditions during different times of day. The dataset consists of 30 classes, which are present in the pixel-level semantic labels. The dataset contains 5000 finely annotated images and 20000 coarsely annotated images with respect to segmentation tasks.

The India Driving Dataset (IDD)[17] is based on roads in India specifically in Hyderabad, Bangalore cities and their outskirts. This dataset shares similarity with Cityscapes with respect to the labeling approach where the label is a 2D image where each pixel is associated with a class. This dataset consists of 10,003 images with varying resolution with respect to segmentation and those masks are finely annotated with 34 classes.

Most labels share the same or similar class names between the two dataset, but they have a different ID assigned to them. Both the dataset have certain country specific classes in them such as auto-rickshaw, trailer, rectification border, etc. The datasets are divided into train, test and val for both the dataset, where train is what we fine-tune the model on, validation is used for validating the model's performance on unseen data and test is used when submitting results to a specific competition held by the dataset authors. We read the images in standard RGB format and resize them to 224, since the pretrained model itself was trained using 224 resolution, this also alleviates the issue of certain images in datasets being of different resolution. We also shuffle the images before giving as input to the model.

3.3 Fine-tuning and Hyperparameters

We train two models one for Cityscapes and other for IDD, for both the models we share a common set of hyperparameters. We use a single Nvidia Quadro RTX6000 for this work, we trained the models for 24 hours of wall time with mixed precision enabled[21] in pytorch[22] which leads to 229 epochs, with the batch size being 75. The initial learning rate was set to 1x10−4 which decreases down to 1x10−8 by the end of the fine-tuning process, using the learning rate scheduler ReduceOnPlateau available in PyTorch. The output of the model is N xW xH where N is the number of classes with respect to the dataset, W and H are the output dimensions of the model's predicted image which are 56 and 56 respectively. The image is then upscaled to the original input resolution which is 2242 using bilinear upscaler, we do this so that we can compare the prediction against the labels then back-propagate using a loss function. The segmentation is a pixel wise categorical classification task applied over an image, therefore we use cross-entropy loss.

4. RESULTS

The model's performance is evaluated based on pixel wise accuracy and loss. We also show how changing certain hyperparameters also affects the model's performance on selected metrics. We first train the model for 12 hours on IDD without using LRScheduler and a fixed learning rate of 0.001, we observed overfitting in validation set as shown in Figure 1. After adding LRScheduler we observe significant improvements in the model loss, but the improvements stop after 46th epoch, see Figure 2. We believe this is because the learning rate cannot go lower since the value is at the minimum value of fp16, which is enabled when using mixed precision.

Fig 1: Validation loss on IDD

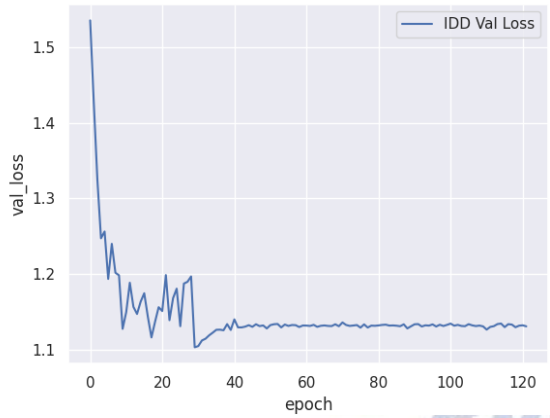

Fig 2 : Validation loss on IDD with LRScheduler

We notice a similar trend in Cityscapes as well where overfitting occurs when fixed learning rate is used, see Figure 3 and the learning goes stale when LRScheduler can't reduce the rate below the fp16 limit,see Figure 4.

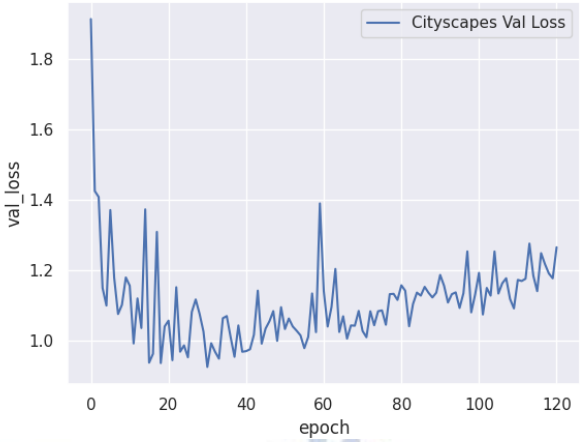

Fig 3 : Validation loss on Cityscapes

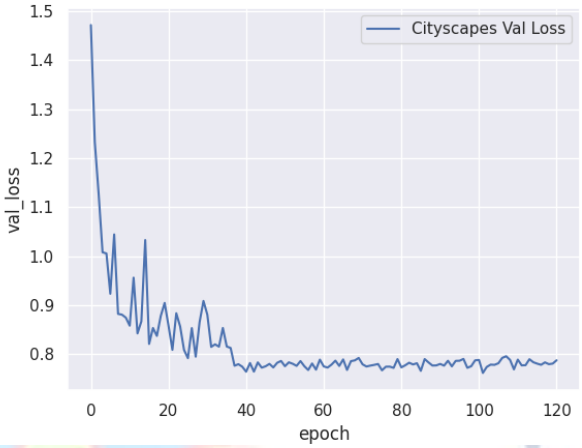

Fig 4 : Validation loss on Cityscapes with LRScheduler

5. CONCLUSION

Recent breakthroughs in Self-Supervised Learning (SSL) has helped in redefining how deep learning deals with massive amounts of data once again. The selected Data2Vec model is one such approach in current state-of-the-art methods of doing SSL, which has proven to be a single model that can achieve competitive results on speech, text and images. We take a pretrained model of Data2Vec and fine-tune it on Cityscapes and IDD datasets, in order to test its performance on a dataset that represents real world challenges in autonomous driving.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual

representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020.

[2] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9729–9738, 2020.

[3] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery.

[4] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. Data2vec: A general framework for self-supervised learning in speech, vision and language. arXiv preprint arXiv:2202.03555, 2022.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[7] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[9] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International Conference on Machine Learning, pages 10347–10357. PMLR, 2021.

[10] Mark Chen, Alec Radford, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In ICML, 2020.

[11] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 9650–9660, October 2021.

[12] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254, 2021.

[13] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6881–6890, 2021.

[14] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7262–7272, 2021.

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[16] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[17] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1743–1751. IEEE, 2019.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009.

[19] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In Proceedings of the European Conference on Computer Vision (ECCV), pages 418–434, 2018.

[20] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer, 2020.

[21] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F Diamos, and Erich Elsen. David garc'i a, boris ginsburg, michael houston, oleksii kuchaiev, ganesh venkatesh, and hao wu. 2018. mixed precision training. In Proceedings of the 6th International Conference on Learning Representations (ICLR). Vancouver, Canada, 2017.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Advances in Neural Information Processing Systems 32, pages 8024–8035. Curran Associates, Inc., 2019.