Data Driven Sensing for Action Recognition Using Deep Convolutional Neural Networks

Ronak Gupta[1(✉)], Prashant Anand[1], Vinay Kaushik[1], Santanu Chaudhury[1,2], and Brejesh Lall[1]

[1] Department of Electrical Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India
ronakgupta143@gmail.com

[2] Indian Institute of Technology Jodhpur, Jodhpur 342037, India

Abstract. Tasks such as action recognition requires high quality features for accurate inference. But the use of high resolution and large volume of video data poses a significant challenge for inference in terms of storage and computational complexity. In addition, compressive sensing as a potential solution to the aforementioned problems has been shown to recover signals at higher compression ratios with loss in information. Hence, a framework is required that performs good quality action recognition on compressively sensed data. In this paper, we present data-driven sensing for spatial multiplexers trained with combined mean square error (MSE) and perceptual loss using Deep convolutional neural networks. We employ subpixel convolutional layers with the 2D Convolutional Encoder-Decoder model, that learns the downscaling filters to bring the input from higher dimension to lower dimension in encoder and learns the reverse, i.e. upscaling filters in the decoder. We stack this Encoder with Inflated 3D ConvNet and train the cascaded network with cross-entropy loss for Action recognition. After encoding data and undersampling it by over 100 times (10×10) from the input size, we obtain 75.05% accuracy on UCF-101 and 50.39% accuracy on HMDB-51 with our proposed architecture setting the baseline for reconstruction free action recognition with data-driven sensing using deep learning. We experimentally infer that the encoded information from such spatial multiplexers can directly be used for action recognition.

Keywords: Data driven compressive sensing (CS) · 3D Deep Convolutional Neural Networks (DCNN) · Perceptual compression · Reconstruction-free action recognition

1 Introduction

Action recognition is a fundamental task in computer vision community with widespread applications in video surveillance, unmanned aerial vehicles (UAV)

B. Deka et al. (Eds.): PReMI 2019, LNCS 11941, pp. 250–259, 2019.
https://doi.org/10.1007/978-3-030-34869-4_28

to name a few. In such applications, instead of using expensive methods for video compression and transmission implemented at transmitter end, compressive sensing (CS) can be applied to encode the data. The encoded data can then be decompressed and processed for high-level inference (action recognition). However, at higher compression ratio the reconstruction based action recognition framework will provide low quality results as they are not optimized to be used over compressed measurements or reconstructed data.

Recent works in deep learning have motivated us to perform data-driven CS for encoding the data and directly performing high-level inference on a large-scale dataset [1,2]. One of the key advantages data-driven encoder-decoder approach offers for CS based recognition is that it allows network to learn more complex patterns from data that may not be easily expressible in a model-based approach [3]. Thus, *this paper proposes a novel Data-driven sensing framework using Deep convolutional neural networks trained with joint MSE and perceptual loss in the context of reconstruction-free action recognition.*

One of the key challenges to perform reconstruction free action recognition using Deep convolutional neural networks(DCNN) is dimensionality reduction in convolutional layers. Usually, downsampling is performed in DCNN using max pooling, average pooling, stochastic pooling or spatial pyramid pooling [1]. Since the above methods are handcrafted which are designed for achieving translation invariance or fixed-length feature representation, they do not optimally preserve signal information while reducing dimensionality [1]. Therefore, learning dimensionality reduction from data itself is desired so as to preserve more information compared to handcrafted ways that do not use training data for downsampling. We have used sub-pixel convolutional layer for performing downsampling that uniformly distributes the samples into multiple dimensions thereby reducing their scale while increasing the number of channels of the data. This makes the data more suitable for dimensionality reduction using convolutional encoder where the higher dimension signal is brought to lower dimension by aggregating feature maps and then reducing the channels in the data. The signal can be recovered from undersampled measurements by using convolutional layers and a sub-pixel convolution layer which learns the upscaling filters and brings the signal to original dimension [4].

To train the network, we propose to use perceptual loss and mean squared error loss. The intuition of using perceptual loss is that the reconstructed output would be similar to the input image as it preserves structural information. In several works [5,6], it has also been shown that good quality images can be generated using perceptual loss functions based on differences between high-level percepts extracted from pre-trained convolutional neural networks, for classification, instead of based on per-pixel differences.

Outline of the paper is as follows: Sect. 2 revisits the existing work in this area. The proposed methodology is explained in Sect. 3. Section 4 presents experimental results to show the effectiveness of the framework and Sect. 5 concludes the paper.

2 Prior Work

Mousavi et al. [7] introduced first data driven sensing and recovery using deep learning framework. They applied stacked denoising autoencoder (SDA) as an unsupervised learning approach to capture statistical dependencies between the linear and non-linear measurements and improve the signal recovery compared to the conventional CS approach. The limitation of this work was that it consisted of a fully connected network. Kulkarni et al. [8] proposed a novel class of CNN architecture called ReconNet as decoder which takes in CS measurements of an image block as input and outputs reconstructed image block. Further the reconstructed image blocks are arranged appropriately and fed into an off-the-shelf denoiser to remove the artifacts. Mousavi et al. [1] proposed a deep convolutional neural network to learn a transformation from the original signals/images to a near-optimal number of undersampled measurements instead of using conventional random linear measurements obtained through a fixed sensing matrix and learns the inverse transformation for recovery from measurements to original signals/images. Learning undersampled measurements from original signal preserves more information. However their results were limited to 1D signals.

Early research focussed on non-deep learning architectures for recognition. Kulkarni et al. [9] presented a method for quantifying the geometric properties of high-dimensional video data in terms of recurrence textures for performing activity recognition at low data rates. In [10] Kulkarni et al. proposed a correlation-based framework in compressed domain and avoids reconstruction process. They showed that Action MACH (Maximum Average Correlation Height) correlation filters can be implemented in compressed domain to find correlation with compressed measurements using the concept of smashed filtering in the space-time domain. Recently, researchers have showed focus on deep learning based recognition on undersampled measurements. In [11] Adler et al. presented an end-to-end deep learning approach for Compressed Learning for classification of image in which the training jointly optimizes the sensing matrix and the inference operator. In [12] Lohit et al. shows that convolutional neural networks can be employed to extract discriminative non-linear features directly from data-independent random CS measurements. They project the CS measurements to the image space by a fixed projection matrix and later apply CNN to the intermediate projection to classify images. In [13], Zisselman et al. presented that optimizing the sensing matrix jointly with a nonlinear inference operator using neural networks, improved upon the methods which used a standard linear projection such as random sensing, PCA etc. In their experiments, the signals were reshaped to full dimension prior entering inference stage, since they used redesigned networks and added compression-decompression layers. In [3] Lohit et al. designed a three-stage training algorithm that allows learning the measurement operator and the reconstruction/inference network jointly such that the system can operate with adaptive measurement rates.

The proposed approach, shown in Fig. 1, does data-driven sensing and learns undersampled measurements with perceptual loss for action recognition. More-

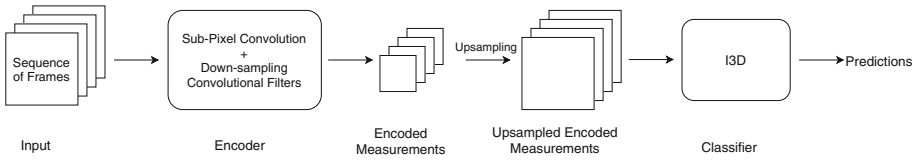

Fig. 1. Proposed framework for data-driven sensing and reconstruction free action recognition

over, the approach directly works on 2D signals which gives us good performance over compressively sensed UCF-101 and HMDB-51 action recognition datasets.

3 Methodology

3.1 Data-Driven Sensing Through Convolutional Autoencoder (CAE)

Our data-driven compression involves CAE (Fig. 2) with sub-pixel convolutional layers and optionally, VGG 16 pre-trained network. The input to the Encoder of CAE is the original image $x \epsilon \mathbb{R}^{H*W*3}$ which is multi-channel and 2-dimensional. The first layer is sub-pixel convolutional layer that learns to generate aggregated feature maps of reduced dimension. The value of reduced dimension is the height(H) and width(W) divided by r, which is the undersampling factor. So the output of first layer is $(H/r) \times (W/r) \times 3r^2$. It means the first sub-pixel convolutional layer divides the length of output feature map by a factor r^2 and increases the number of channels in output feature map by a factor of r^2. Mathematically, sub-pixel convolutional layer (inverse of pixel shuffle [4]) here can be described as:

$$\hat{x}(x,r)_{i,j,c} = x_{i*r+floor(\frac{mod(c,r^2)}{r}), j*r+mod(mod(c,r^2),r), floor(\frac{c}{r^2})} \quad (1)$$

The sub-pixel convolutional layer reduces the input image scale, however to undersample the input in Encoder, we employ several blocks of convolutional layers same as Inception module (Fig. 3) used in [14] to reduce the dimensionality by decreasing the total number of feature maps such that dimension of the output equals to $(H/r) \times (W/r) \times 3$. This reduces the total number of measurements by a factor of $r \times r$ and thus the undersampling ratio is $r^2(r \times r)$. Such sub-sampling ensures that the channels preserve the structural information of input image.

Once the undersampled measurements are obtained from Encoder, we now employ several blocks of convolutional layers to extract feature maps in Decoder. Hence in our architecture we learn to encode and reconstruct images directly to the measurement domain. Now, the output of encoder lies in $\mathbb{R}^{(H/r)*(W/r)*3}$ domain, however we want to recover the input signal which lies in $\mathbb{R}^{(H)*(W)*3}$ domain. Using the block of convolutional layers in Decoder would boost the dimensionality to $\mathbb{R}^{(H/r)*(W/r)*3r^2}$. In last we apply sub-pixel convolution

layer to rearrange the feature maps and generate the output, $\hat{x}$, which lies in $\mathbb{R}^{(H)*(W)*3}$.

Mathematically the last layer(pixel shuffle layer [4])can be described as:

$$\hat{x}(x,r)_{i,j,c} = x_{floor(i,r),floor(j,r),c*r^2+r*mod(i,r)+mod(j,r)} \tag{2}$$

We train our CAE with MSE, perceptual and joint loss, weighted combination of MSE and perceptual loss. The perceptual loss we used here is defined in [5], that measures high-level perceptual and semantic differences between images. Their perceptual loss is function of deep convolutional neural networks (VGG 16) pre-trained for image classification. Instead of per-pixel loss between the output image $\hat{x}$ and original image x, we use the similar feature representations as computed by the loss network ϕ. Let $\phi_j(x)$ be the j^{th} layer activations of network ϕ when processing the image x, if j is a convolutional layer then we get a feature map of size $C_j \times H_j \times W_j$ as $\phi_j(x)$. The perceptual loss is the sum of normalized Euclidean distance between feature representations of corresponding convolutional layers as showed in Eq. 3. That means the perceptual loss is minimum, when the classification output of pre-trained convolutional neural network (VGG 16 or VGG 19 [15]) for the reconstructed image would be same as that for the original image.

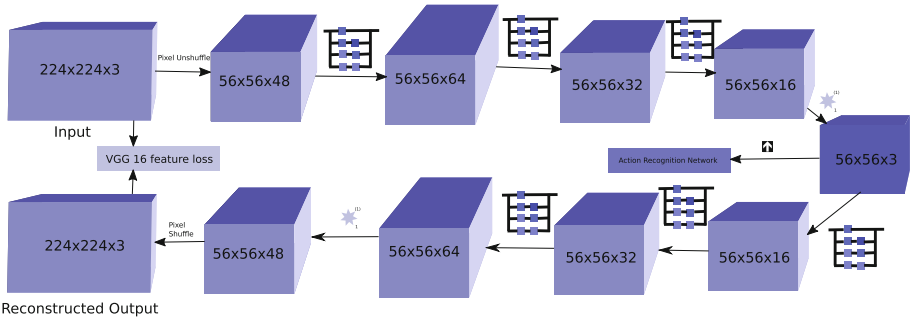

Fig. 2. Data driven Sensing framework for learning undersampled measurements (dimensions shown for undersampling ratio of 4×4)

$$l^{\phi,j}_{feat}(\hat{x},x) = \sum_j \frac{1}{C_j H_j W_j} ||\phi_j(\hat{x}) - \phi_j(x)||^2 \tag{3}$$

The image content and structural information is preserved but color, texture and exact shape are not preserved while using perceptual loss [5]. Hence we train our Autoencoder with joint loss (weighted) of MSE loss and perceptual loss as shown in Eq. 4.

$$L_{joint}(\hat{x},x) = \alpha \left(\frac{1}{s} \sum_{i=1}^{s} ||\hat{x} - x||_2^2 \right) + \beta \left(l^{\phi,j}_{feat}(\hat{x},x) \right) \tag{4}$$

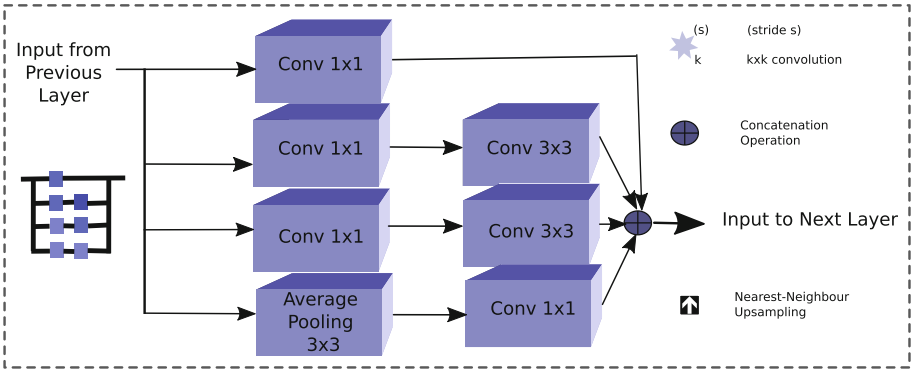

Fig. 3. Inception module in our CAE architecture

3.2 Learning Action Recognition in Compressed Domain

The convolutional encoder of the reconstruction network is now stacked with the Inflated 3D (I3D) convolutional neural network [16]. Since the structural information encoded in the Convolutional Encoder lies in $\mathbb{R}^{(H/r)*(W/r)*3}$, we upsample the encoded information by $r \times r$ times using a nearest neighbour upsampling layer. The upsampling of encoded feature maps is not equivalent to reconstruction. The encoder is initialized using learned weights from reconstruction network. The 3D convolutional network is initialized using pretrained weights from Imagenet and Kinetics dataset. The classification architecture is shown in Fig. 4. Further, this action recognition architecture is then trained with the cross-entropy classification loss using SGD optimizer.

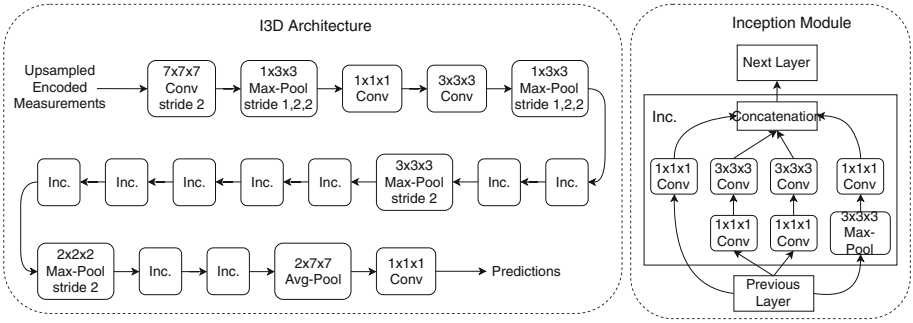

Fig. 4. Action recognition architecture

4 Experimental Results

In this section, we report the performance of our proposed framework at different undersampling ratios. Once our framework is end-to-end trained, the convolutional encoder and trained I3D network can be separated. The convolutional encoder does sensing on unseen data at transmitter end and generates undersampled measurements that are given to trained I3D network for action recognition at receiver end. Hence, our input size for action recognition is of lower dimension compared to the original signal. In Table 1, we compare action recognition results over UCF-101 [17] and HMDB-51 [18] datasets of different techniques. Here, STSF [10], Recon [19] + IDT [20] and our proposed approach are techniques that perform compressive sensing while C3D+ [21], RGB-I3D [16] are deep learning based action recognition over original signal. MB motion vectors+3DConvNet [22] shows action recognition results over macroblock motion vectors of H.264 compressed signal.

Table 1. Performances of different action recognition framework in compressed domain

Framework	Input size	Accuracy (%)	
		UCF-101	HMDB51
STSF [10]	FBI[a]	–	22.5
Recon [19]+IDT [20]	FBI	–	57.2
MB motion vectors	$24^2 \times 2 \times 160$	77.5	49.5
+3D ConvNet [22]			
C3D+ [21]	$112^2 \times 3 \times 16$	82.3	–
RGB-I3D [16]	$224^2 \times 3 \times 64$	95.6	74.8
Proposed (2 × 2)	$112^2 \times 3 \times 32$	91.25	66.66
Proposed (4 × 4)	$56^2 \times 3 \times 32$	89.53	65.88
Proposed (8 × 8)	$28^2 \times 3 \times 32$	80.84	56.33
Proposed (10 × 10)	$22^2 \times 3 \times 32$	75.05	50.39

[a]Full Blown Images

In Table 2, recognition results of our proposed approach with different undersampling ratios of original signal for different reconstruction loss has been presented. The accuracy for 2×2 and 4×4 undersampling ratio is similar, since there is not much change in Encoder parameters. In Table 2, we observe that the accuracy decreases in small amount at higher undersampling ratio and conclude that the information is correctly captured by the Encoder parameters. We present the same for both datasets displaying the efficiency of our proposed approach. Table 2 shows that when Encoder is trained with joint loss, the performance of action recognition pipeline increases as perceptual loss makes sure that the encoder captures more structural information. Table 3 compares the performance of sensing network with different number of Inception submodules

Table 2. Avg. Accuracy on UCF101 splits

Undersampling ratio	CAE RMSE loss	CAE perceptual loss	CAE joint loss	Encoder complexity (W)
4 (2×2)	87.89	90.85	**91.25**	3,651
16 (4×4)	89.40	88.79	**89.53**	14,894
64 (8×8)	80.78	76.50	**80.84**	179,408
100 (10×10)	73.62	73.96	**75.05**	404,278

in CAE, using reconstruction loss and action recognition accuracy as an evaluation metric over UCF-101 dataset splits at compression ratio of 4×4. The performance in Table 3 infers that optimal sensing is obtained with 3 Inception submodules in data-driven sensing encoder.

4.1 Implementation Details

For training CAE, we use ADAM optimizer with initial learning rate set to 10^{-3} which is reduced by 10^{-1} when validation loss gets saturated. While training CAE with joint loss function as shown in Eq 4, $\alpha = 1$ and $\beta = 0.04$ gives us the best results. To train the stacked network of Encoder and I3D for classification, we employ standard SGD with momentum set to 0.9 and initial learning rate set to 10^{-2}. All the networks were implemented in TensorFlow [23] and ran on nvidia-docker [24] for Tensorflow on NVIDIA DGX-1.

Table 3. Performance with respect to number of Inception modules in Proposed data-driven sensing encoder

No. of Inception submodules	Joint loss (CAE)	Accuracy on UCF-101 (%)	Encoder complexity (W)
1	7.5649	89.21	1,252
2	4.0129	85.94	4,322
3	4.2220	89.53	14,894
4	4.5798	87.84	53,766

5 Conclusion

A data-driven CS framework for reconstruction free action recognition is presented in the paper. Our proposed architecture preserves more structural information utilizing the joint loss function, based on perceptual loss and MSE loss, that provides better performance as compared to individual losses. Experimental

results on UCF-101 and HMDB-51 are presented to show the effectiveness of the framework at various undersampling ratios. For future work, our undersampling ratio specific architecture can be modified to a generic architecture which works with different undersampling ratios.

Acknowledgment. The NVIDIA DGX-1 for experiments was provided by CSIR-CEERI, Pilani, India

References

1. Mousavi, A., Dasarathy, G., Baraniuk, R.G.: DeepCodec: adaptive sensing and recovery via deep convolutional neural networks. arXiv preprint arXiv:1707.03386 (2017)
2. Xu, K., Ren, F.: CSVideoNet: a real-time end-to-end learning framework for high-frame-rate video compressive sensing. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1680–1688. IEEE (2018)
3. Lohit, S., Singh, R., Kulkarni, K., Turaga, P.: Rate-adaptive neural networks for spatial multiplexers. arXiv preprint arXiv:1809.02850 (2018)
4. Shi, W., et al.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1874–1883 (2016)
5. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
6. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T., Zelnik-Manor, L.: The 2018 PIRM challenge on perceptual image super-resolution. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11133, pp. 334–355. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11021-5_21
7. Mousavi, A., Patel, A.B., Baraniuk, R.G.: A deep learning approach to structured signal recovery. In: 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 1336–1343. IEEE (2015)
8. Kulkarni, K., Lohit, S., Turaga, P., Kerviche, R., Ashok, A.: Reconnet: non-iterative reconstruction of images from compressively sensed measurements. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 449–458 (2016)
9. Kulkarni, K., Turaga, P.: Recurrence textures for human activity recognition from compressive cameras. In: 2012 19th IEEE International Conference on Image Processing (ICIP), pp. 1417–1420. IEEE (2012)
10. Kulkarni, K., Turaga, P.: Reconstruction-free action inference from compressive imagers. IEEE Trans. Pattern Anal. Mach. Intell. **38**(4), 772–784 (2016)
11. Adler, A., Elad, M., Zibulevsky, M.: Compressed learning: a deep neural network approach. arXiv preprint arXiv:1610.09615 (2016)
12. Lohit, S., Kulkarni, K., Turaga, P.: Direct inference on compressive measurements using convolutional neural networks. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 1913–1917. IEEE (2016)
13. Zisselman, E., Adler, A., Elad, M.: Compressed learning for image classification: a deep neural network approach. Process. Anal. Learn. Images Shapes Forms **19**, 1 (2018)

14. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
15. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
16. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4724–4733. IEEE (2017)
17. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
18. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: a large video database for human motion recognition. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 2556–2563. IEEE (2011)
19. Needell, D., Tropp, J.A.: Cosamp: Iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**(3), 301–321 (2009)
20. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558 (2013)
21. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
22. Chadha, A., Abbas, A., Andreopoulos, Y.: Compressed-domain video classification with deep neural networks: there's way too much information to decode the matrix. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 1832–1836. IEEE (2017)
23. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous systems (2015). http://tensorflow.org/, software available from tensorflow.org
24. Nvidia gpu cloud tensorflow. nVIDIA offers GPU accelerated containers via NVIDIA GPU Cloud (NGC) for use on DGX systems. https://ngc.nvidia.com/catalog/containers/nvidia:tensorflow