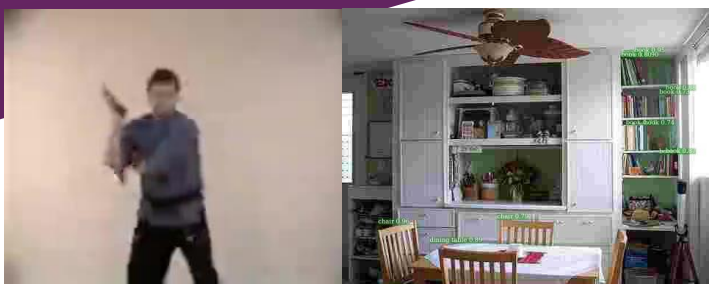
Lossy Image and Video Compression within Deep Neural Networks

Matt Poyser, Amir Atapour-Abarghouei, Toby Breckon

Department of Computer Science
Durham University



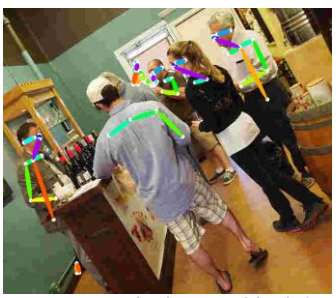
Frame at high compression from UCF101 Human Action Recognition dataset (left), Object Detection at JPEG compression level 10 (right)

Approach

- **How much does lossy compression impact performance of pre-trained networks?**
- **How much performance can we recover by retraining the networks on lossily compressed data?**

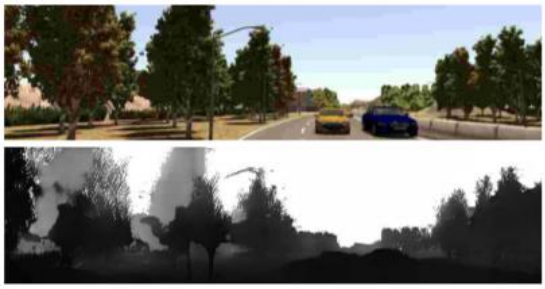
- 5 domains: Segmentation, Depth Estimation, Human Pose Estimation, Object Detection, and Human Action Recognition.
- 5 architectures: encoder-decoder (SegNet [4]), GAN [1], end-to-end CNN (OpenPose [5]), Region-based CNN (FasterRCNN [3]), and two-stream [2].

Pre-trained Network Performance



Human Pose Estimation at JPEG level 10

Segmentation at JPEG level 10



Original (top) and Depth-Estimation (bottom) with pre-trained GAN at JPEG level 10

Results

Results before (left) and after (right) re-training. Red indicates maximum safe compression rate.

Compression Rate	mAP
95	0.703
75	0.686
50	0.666
15	0.545
10	0.442
5	0.187

Compression Rate	mAP
95	0.703
75	0.694
50	0.692
15	0.647
10	0.627
5	0.559

Object Detection at various JPEG compression rates

Compression Rate	mAP
95	0.711
75	0.689
50	0.655
15	0.413
10	0.323
5	0.098

Compression Rate	mAP
95	0.711
75	0.708
50	0.678
15	0.654
10	0.597
5	0.454

Human Pose Estimation at various JPEG compression rates

Compression Rate	Abs. Rel.	Sq. Rel.	RMSE
95	0.0112	0.0039	0.0588
75	0.0116	0.0039	0.0589
50	0.0123	0.0038	0.0587
15	0.0146	0.0040	0.0599
10	0.0192	0.0042	0.0617
5	0.0283	0.0060	0.0749

Compression Rate	Abs. Rel.	Sq. Rel.	RMSE
95	0.0112	0.0039	0.0588
75	0.0113	0.0035	0.0560
50	0.0103	0.0029	0.0502
15	0.0121	0.0034	0.0556
10	0.0152	0.0031	0.0528
5	0.0159	0.0040	0.0599

Depth Estimation at various JPEG compression rates

Compression Rate	global ACC	mean ACC	mIoU
95	0.911	0.536	0.454
75	0.909	0.530	0.448
50	0.904	0.523	0.438
15	0.814	0.459	0.338
10	0.794	0.421	0.304
5	0.782	0.364	0.265

Compression Rate	global ACC	mean ACC	mIoU
95	0.911	0.536	0.454
75	0.910	0.522	0.446
50	0.908	0.503	0.431
15	0.902	0.494	0.420
10	0.895	0.477	0.405
5	0.879	0.445	0.374

Pixel-wise Segmentation at various JPEG compression rates

Compression Rate	Top-1 Spatial	Top-1 Motion	Top-1 Fusion
23	78.8736	70.1198	83.5485
25	78.7999	44.9225	73.6030
30	78.4563	37.3598	72.2329
40	74.5704	38.9565	70.8803
50	44.1977	15.3267	41.4777

Compression Rate	Top-1 Spatial	Top-1 Motion	Top-1 Fusion
23	78.8736	70.1198	83.5485
25	78.9056	39.7192	71.7616
30	78.5620	34.3161	70.5765
40	75.9450	9.2550	67.1227
50	62.5165	6.7300	56.2279

Human Action Recognition at various H.264 CRF values

Conclusions

- **Can afford to compress to 15% of the original size, across all domains**, without loss in performance
- SegNet, GAN and two-stream spatial stream are particularly resilient. We posit that this is because of the **up-sampling within the pooling layers of their decoder sub-network.**

[1] A. Atapour-Abarghouei and T. Breckon, "Real-time monocular depth estimation using synthetic data with domain adaptation," in Proc. Computer Vision and Pattern Recognition. IEEE, June 2018, pp. 1–8.
[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in Advances in Neural Information Processing Systems 27, 2014, pp. 568–576.
[3] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," Computing Research Repository, vol. abs/1506.01497, 2015.
[4] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," Computing Research Repository, vol. abs/1511.00561, 2015
[5] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in Conference on Computer Vision and Pattern Recognition, 2017, pp. 1302–1310