Save on compute power and make your models run faster

Deploy Model Compression

Any firm using machine learning should consider using Model Compression techniques on their usually very large models to save on compute power and make the models run faster. The ability to benchmark this focuses the firm on their priorities for decision-making (cost savings vs speed vs accuracy).

Machine learning techniques and architectures have led to models that have significantly improved accuracy rates over their predecessors. However, they rely on 'millions' to 'billions' of internal model parameters, trained over very long periods of time. As more complex models are developed, the desirable qualities of the deep learning models are:

- **Low Storage Requirements**
- **Computational Efficiency**
- **Fast Inference**

These qualities become critical factors when considering real-time applications, like high-frequency trading algorithms where speed is important or deployment on hardware or even mobile devices.

Your algorithm can be right 100% of the time, but, if the result is received past the point when the action should have been taken, then it was useless.

Addressing the Challenges

The main approach to tackling challenges in machine learning compression is to produce models which are significantly smaller in both memory and computational requirements without sacrificing the accuracy of the complete original model.

Compression categories include:

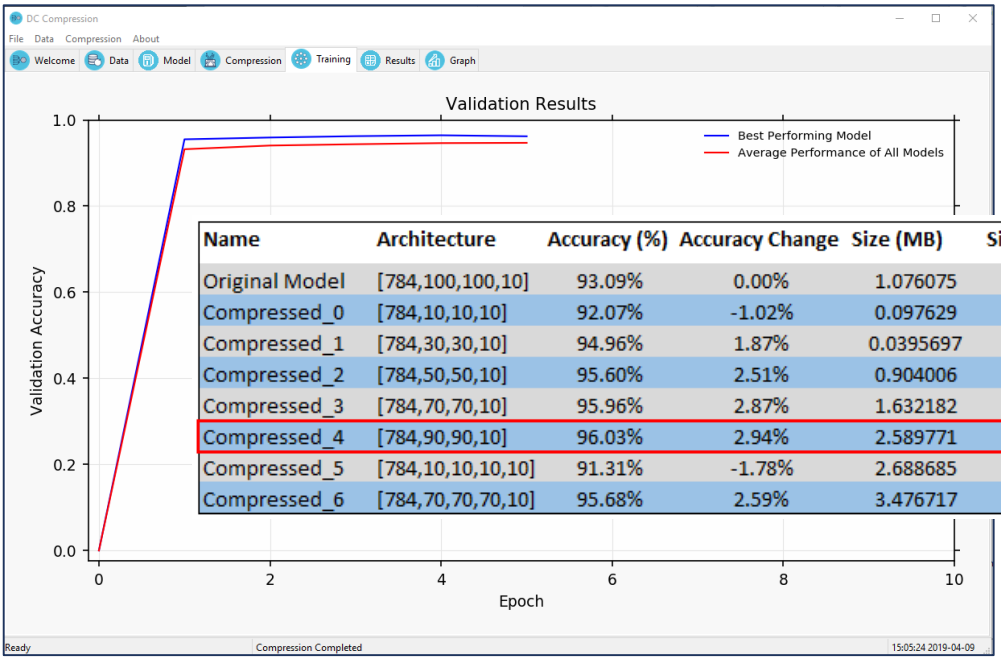
- **Parameter Pruning and Sharing**
- **Low-rank Factorization**
- **Knowledge Distillation**

An optimal model compression technique would allow you to reduce the space required to store your model, allowing you to deploy your models to smaller hardware devices or, where appropriate, mobile phones and wearable technology.

The ability to use more than one model compression techniques provides for the ultimate compressed model.

It is important to ensure the chosen optimal compressed model still have the 'same' model as the original model.

Using DC MINT, for AI Explainability, allows the user to check if the compressed model is still the 'same' model as the output is expected to be similar and a decision can be made if the differences are acceptable or not.



Name	Architecture	Accuracy (%)	Accuracy Change	Size (MB)	Size Reduction (%)	Time (ms)	Speed Up (%)
Original Model	[784,100,100,10]	93.09%	0.00%	1.076075	0.00%	0.008294	0.00%
Compressed_0	[784,10,10,10]	92.07%	-1.02%	0.097629	1002.21%	0.006479	28.01%
Compressed_1	[784,30,30,10]	94.96%	1.87%	0.0395697	171.94%	0.00751	10.44%
Compressed_2	[784,50,50,10]	95.60%	2.51%	0.904006	19.03%	0.007226	14.78%
Compressed_3	[784,70,70,10]	95.96%	2.87%	1.632182	-34.07%	0.007513	10.40%
Compressed_4	[784,90,90,10]	96.03%	2.94%	2.589771	58.45%	0.007961	4.18%
Compressed_5	[784,10,10,10,10]	91.31%	-1.78%	2.688685	-59.98%	0.006732	23.20%
Compressed_6	[784,70,70,70,10]	95.68%	2.59%	3.476717	-69.05%	0.007785	6.54%

For further information please contact:
capitalmarketsdelivery@deltacapita.com

DC COMPRESSION

Delta Capita

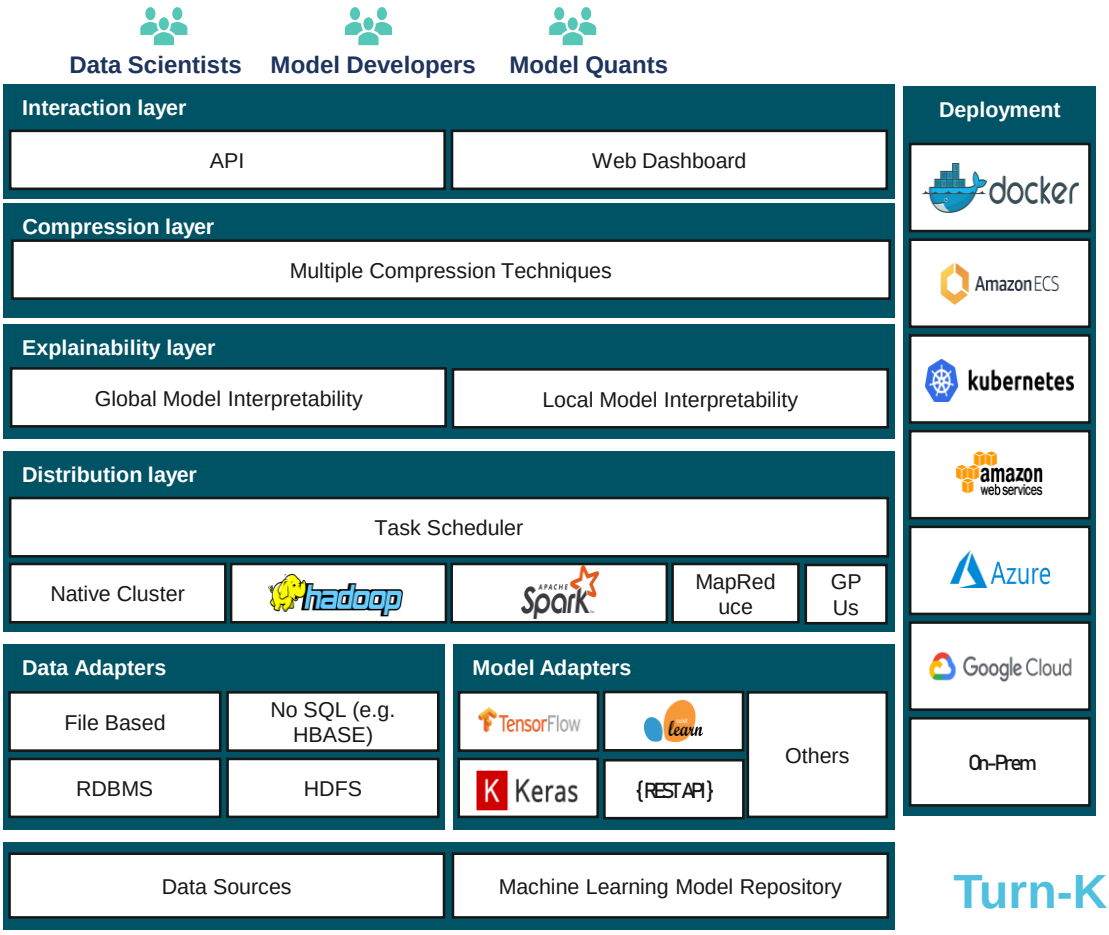
Use DC Compression and save money and make your models run faster

DC Compression allows each of the types of algorithm parameter, pruning and sharing, low-rank factorization, and knowledge distillation to be used separately and orthogonally.

DC Compression provides the user with ensemble techniques and our data scientists are available on hand to advise on suitable techniques or ensemble techniques to use.

DC Compression is a differentiator in reducing model complexity and to speed up models without a significant impact on the accuracy of the model.

"A genuine way of reducing model complexity and to speed up models without a significant impact on the accuracy of the model."



DC COMPRESSION

- ✓ **Easy to deploy**
- ✓ **Less Compute power Smaller models**
- ✓ **Ensemble methodologies**
- ✓ **Quick decision GUIs**
- ✓ **Use MINT to determine if the compressed model is fit-for-purpose**
- ✓ **Support, enhancement and upgrades**

Turn-Key Deployment

Please contact Delta Capita at *capitalmarketsdelivery@deltacapita.com* if you are interested in any of the following areas:

- **DC MINT - Model Interpretability**
- **DC COMPRESSION - Model Compression and Model Optimisation**
- **DC DOCS - Automated Model Documentation**
- **DC VOICE - Voice/NLP on large data sets and Wealth applications**
- **DC RISK - Contagion Risk Models**

For further information please contact:
capitalmarketsdelivery@deltacapita.com