Artificial Intelligence and Machine Learning in Sustainable Energy

Stephen Gallant, Software Delivery Lead; Nishi Patel & Sara Zaheri, ML Engineers; Katalyst Data Management

This paper was prepared for presentation during the 18th International Congress of the Brazilian Geophysical Society held in Rio de Janeiro, Brazil, 16-19 October 2023.

Abstract

Artificial intelligence (AI) relies heavily on Machine Learning (ML). ML is akin to training a tiny model brain (an artificial neural-network) to do specific tasks extremely well and extremely fast. The result of which is a machine-learned model that can be put to use. When one or more of these models are brought together, aided by mathematical algorithms and procedural code, they become AI such as:

- Answering a question, like we see with ChatGPT
- Identifying a person, like we see with Google Photos
- Predicting shopping/reading preferences, like we see with Facebook

Introduction

E&P clients have masses of unstructured data they are uncertain of what they have or how to maintain it, otherwise known as the data swamp. Most often, data management teams (either internal or external vendors) deem these too costly to turn into economical data management projects. With the recent emphasis on sustainable energy projects, masses of legacy unstructured data are now being re-used to support low-carbon projects. Imagine the problem of manually picking out 10,000 raster well log images from a mixed bag of 30,000 images plus other file types. Then manually reading and re-typing well names to match and index into a well data repository. Even if undertaking such projects, data loading teams would be overwhelmed using traditional software tools known today. ML engineers and software developers can greatly improve such tools using AI/ML.

Challenges and Solutions

Service companies in the industry have set out to tackle these challenges:

1) Automatically separate raster well log image files from all other structured or unstructured files.

2) Automatically extract specific meta-data values from the raster well log images to satisfy an index into a well data repository.

3) Present the AI/ML choices to end-users whom can quickly and conveniently verify or make adjustments to index into a well data repository.

This has resulted in like-wise solutions:

1) A machine-learned image-classification model, which can (at a high-level of confidence) distinguish raster well log images from any other image file.

2) An enhanced Optical Character Recognition (OCR) system to extract meta-data from image files.

3) Natural Language Processing (NLP) algorithms to match extracted values against keyword lists labels (f.ex. “Well Name”) and data repository values (f.ex. “Bighorn Well 4506”).

4) A workflow user interface (UI) to bring these all together allowing end-users to break the barrier and accomplish much more with higher confidence and higher quality than before.

Choosing a Model

For the challenge of image-classification our ML engineers choose a convolutional neural network (CNN) to perform supervised training. CNNs are commonly used in solving problems related to computer vision and spatial data, such as images.

For efficiency, we chose to use pre-trained models (backbone models), which already have some intelligence (meaning pre-trained with a large amount of data). We chose a model pre-trained on Imagenet. Imagenet is an open-source labelled dataset containing over 14 million images across over 20,000 classes.

Such pre-trained models can still be fine-tuned and trained furthermore with our own dataset to meet more specific goals of well log classification.

We chose to compare two different pre-trained CNN models, VGG16 and InceptionResnetV2 (each further-more fine-tuned with our own training dataset) ultimately picking VGG16 to move forward with.

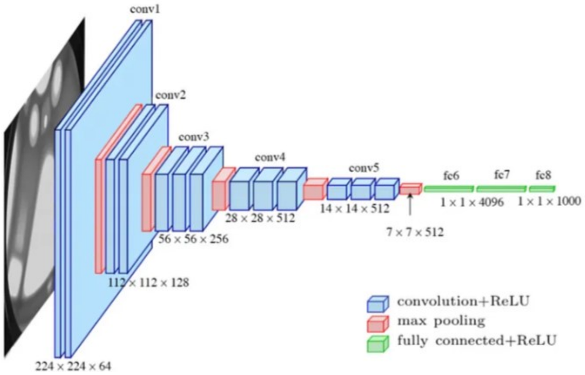


***Figure 1** - The architecture of VGG16* [1]

Accumulating a Training Dataset

In order to accomplish this, we accumulated our own training dataset of image files, which had to be accurately labeled meaning, we know exactly what they are. Supervised learning uses labeled input, which can be fed to the model and trained to recognize similar images when predicting new/unknown data.

Sourced from our existing Katalyst Data Management (KDM) data repositories, we had a head-start in the labelling process, but they still needed to be verified and corrected. We built tools, which allowed us to view these images and quickly left/right arrow key a labeled choice. We also incorporated clustering techniques to accelerate this process and focus training on those specific parts of well log images, which identify them as a well log. Those would be the chart measurement pages containing a graph with wavy lines running up and down the page. It is that specific type of image, which distinguishes a well log, not other report charts or header text, which commonly show up in other classes of image files.

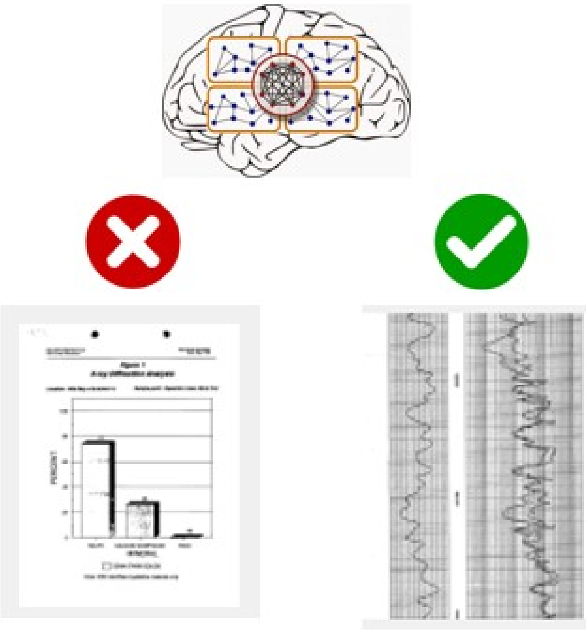


***Figure 2** – Comparison of a well log versus a non-well log*

We accumulated approximately 3 million image pages to compromise our training set. Only about 20% have been used for well log classification as we expect to add more classes in the near future. The training set must be balanced meaning, just as many non-well logs need to be used as well logs. Both log & non-log data sets must furthermore contain a balanced set of various styles following the proprietary nature of client data. Clustering techniques were again used here to accelerate this process.

A single image file contains many pages, so each file must be pre-processed to separate the pages and conform each one to have common attributes like dimension. This must be done not only for training data but also when consuming new/unknown data to predict. This entire pre-processing piece was coded and automated.

In the final version, our training data set was comprised of roughly 250,000 well log images and a balanced mix of 250,000 non-well log images. We also set aside roughly 25,000 images for validation and 25,000 images for testing.

Training the Model

Another challenge was obtaining high-end Linux machines, which not only have Central Processing Units (CPUs) but also Graphics Processing Units (GPUs). GPUs are responsible for rendering images by performing rapid mathematical calculations and in particular, performing multiple (millions of) calculations at the same time. Although GPUs are not required to perform model training, they greatly accelerate the process turning weeks or months of training time into hours or days. Using CPUs alone becomes almost impractical. Even with GPUs, training a model can take a significant amount of time. In our case, it took approximately 5 days to train (fine-tune) the VGG16 model using 2x 24G NVIDIA GeForce RTX 3090 GPUs.

We used pre-trained weights for most layers, unfroze some of the last layers (deleted those pre-trained weights) and trained them on our dataset. We also added more layers to improve training accuracy. It took five iterations of model training each time fine-tuning our dataset and model architecture to achieve acceptable results. In the final version, we ended up training with 20 epochs. Epochs represent the number of times your training dataset is passed through the model. There will be a certain number of epochs where the model has reached optimal learning.

In the following TensorBoard report, the blue line represents the validation dataset and the orange line represents the training dataset. The x-axis is the epochs and the y-axis is the accuracy. This type of output is carefully watched and evaluated as training proceeds and gives a good indication when to add or remove epochs.

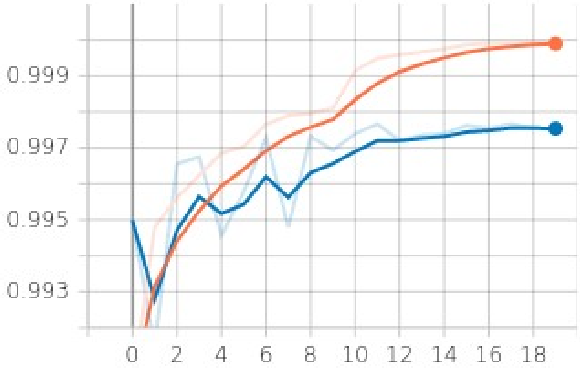
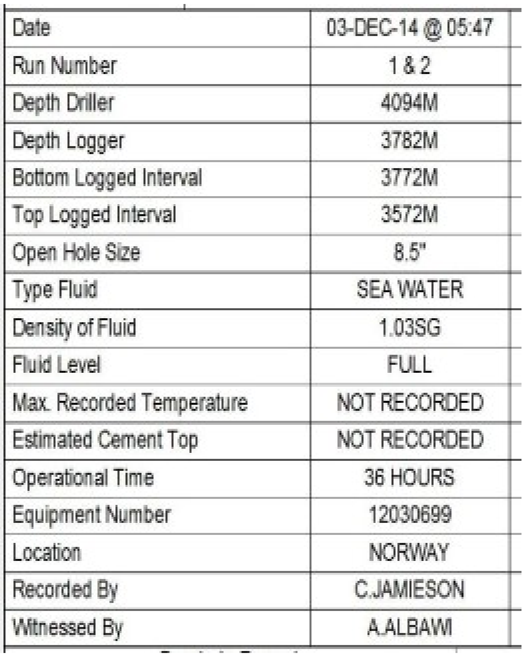


***Figure 3** – TensorBoard accuracy report*

We also attempted to fine-tune a different model architecture, Contrastive Language–Image Pre-training (CLIP), which is multi-modal combining NLP with computer vision. A main benefit is its ability to be trained and predict using text in addition to images. This is likely to be a future initiative. For CLIP, we did not unfreeze any layers (as it is very heavy) but we added more layers. Using images alone, it did not outperform VGG16 but we suspect with more training, CLIP may become a preference.

Meta-Data Extraction

For the challenge of OCR, we quickly realized that basic open-source software was not good enough to accurately extract text from tables and cells which is commonly found in textual header pages of well logs. Neither Tesseract nor PaddleOCR worked accurately with this style of data but we moved forward with PaddleOCR.



Date	03-DEC-14 @ 05:47
Run Number	1 & 2
Depth Driller	4094M
Depth Logger	3782M
Bottom Logged Interval	3772M
Top Logged Interval	3572M
Open Hole Size	8.5"
Type Fluid	SEA WATER
Density of Fluid	1.03SG
Fluid Level	FULL
Max. Recorded Temperature	NOT RECORDED
Estimated Cement Top	NOT RECORDED
Operational Time	36 HOURS
Equipment Number	12030699
Location	NORWAY
Recorded By	C.JAMIESON
Witnessed By	A.ALBAWI

Figure 4 – A typical well log header page containing meta-data values within tables/cells to be OCR extracted

Firstly, we attempted to use an object detection model called You Only Look Once (YOLO) [2] to filter out the boxes with inaccurate results. Then we fine-tuned a table detection model called CascadeTabNet [3] to be used in combination with an open-source line detection model called L-CNN [4] to significantly improve OCR extraction of values from tables and cells with great success.

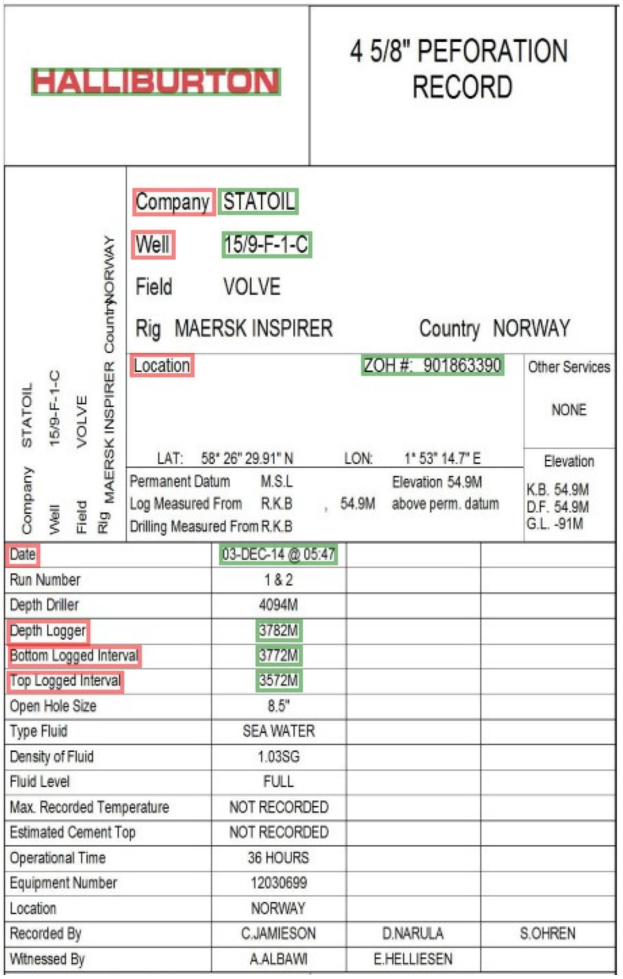
The next challenge was how to extract the specific values we need to populate a well log inventory. KDM already has a standard (LAS & DLIS) well log loading application, which extracts the following attributes:

UWI, Well Name, Well License, Acquired For Company, Logging Company, Log Date, Top/Bottom/Logger TD Depths & UOMs, and Sample Rate & UOM.

We aimed to extract the same attributes from raster well log image files. The business analyzed the data and established a list of keywords, which commonly exist as identifying labels for each of these values, label keywords. We also already had actual value listings for the well header data and company names, value keywords. Combined together, we were able to accurately extract label/value pairs to populate directly into the same well log loading application (this time using images).

Each piece of extracted text also includes the image pixel coordinates, so we were able to show the end-user exactly where each extracted value came from while they were loading the image files. This gave them the ability to quickly verify the AI/ML choices and even override the choice to click directly on the image and choose another value if necessary.

We used a combination of regular pattern matching and other NLP methods of both cosine similarity [5] and Natural Language Toolkit (NLTK) [6] edit distance algorithms to produce fuzzy match scores and rank them by various methods such as label + value scoring, label keyword priority, page number, location on the page, etc.



HALLIBURTON

4 5/8" PEFORATION RECORD

Company STATOIL
Well 15/9-F-1-C
Field VOLVE
Rig MAERSK INSPIRER Country NORWAY

Company STATOIL
Well 15/9-F-1-C
Field VOLVE
Rig MAERSK INSPIRER Country NORWAY

Location ZOH #: 901863390

LAT: 58° 26" 29.91" N LON: 1° 53" 14.7" E

Other Services NONE

Permanent Datum M.S.L Elevation 54.9M
Log Measured From R.K.B , 54.9M above perm. datum
Drilling Measured From R.K.B

Elevation
K.B. 54.9M
D.F. 54.9M
G.L. -91M

Date	03-DEC-14 @ 05:47		
Run Number	1 & 2		
Depth Driller	4094M		
Depth Logger	3782M		
Bottom Logged Interval	3772M		
Top Logged Interval	3572M		
Open Hole Size	8.5"		
Type Fluid	SEA WATER		
Density of Fluid	1.03SG		
Fluid Level	FULL		
Max. Recorded Temperature	NOT RECORDED		
Estimated Cement Top	NOT RECORDED		
Operational Time	36 HOURS		
Equipment Number	12030699		
Location	NORWAY		
Recorded By	C.JAMIESON	D.NARULA	S.OHREN
Witnessed By	A.ALBAWI	E.HELLIESEN	

Figure 5 – A typical well log header page showing where the AI/ML picked values were extracted

A future consideration for extracting these particular values is to use a Large Language Model (LLM). This is what ChatGPT is comprised of where the values are extracted in a question/answer type of implementation using prompt engineering. This involves the context of words in relation to other words in the full given text as a whole. The sustainable energy sector is looking at applications for this kind of text extraction in the areas of using well logs for reservoir characterization in hydrogen and carbon storage, and using well log data to support prediction of geomechanical risk in CO2 sequestration sites.[7]

Workflow UI Tools

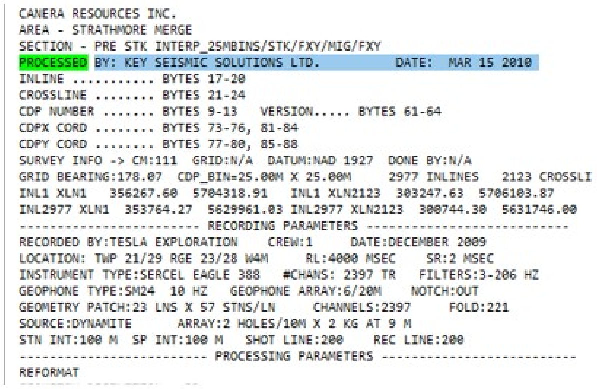
It is important to note none of these machine-learned models can stand up on their own. They require a lot of algorithms and procedural code to contain them and control their inputs and outputs. Much effort was put into building these workflow UI tools with business analysis directly from operations teams to purpose fit the KDM data ingestion process. The workflow tools are comprised of:

1) ML Service – headless-service hosting the machine-learned models constantly running on one or more high-spec Linux machines consuming data files and outputting results.

2) Kurator – user interface to interact with the ML Service, feed it new/unknown data files, compile the AI/ML outputs and pass along to downstream data loading applications.

3) Well Log Loader & KIT – existing data loading applications enhanced to present the AI/ML choices to data indexers.

We were able to apply the same NLP techniques to SEGY textual headers when loading seismic data. Here we focused on Acquired For, Created By, Created Date, Category (FIELD vs PROCESSED), Description and Coordinate System.



```
CANERA RESOURCES INC.
AREA - STRATHMORE MERGE
SECTION - PRE STK INTERP_25MBINS/STK/FXY/MIG/FXY
PROCESSED BY: KEY SEISMIC SOLUTIONS LTD.          DATE:  MAR 15 2010
INLINE ........... BYTES 17-20
CROSSLINE ........ BYTES 21-24
CDP NUMBER ....... BYTES 9-13   VERSION..... BYTES 61-64
CDPX CORD ........ BYTES 73-76, 81-84
CDPY CORD ........ BYTES 77-80, 85-88
SURVEY INFO -> CM:111  GRID:N/A  DATUM:NAD 1927  DONE BY:N/A
GRID BEARING:178.07  CDP_BIN=25.00M X 25.00M     2977 INLINES   2123 CROSSLI
INL1 XLN1   356267.60  5704318.91   INL1 XLN2123  303247.63  5706103.87
INL2977 XLN1  353764.27  5629961.03 INL2977 XLN2123  300744.30  5631746.00
------------------------ RECORDING PARAMETERS ---------------------------
RECORDED BY:TESLA EXPLORATION    CREW:1     DATE:DECEMBER 2009
LOCATION: TWP 21/29 RGE 23/28 W4M     RL:4000 MSEC    SR:2 MSEC
INSTRUMENT TYPE:SERCEL EAGLE 388   #CHANS: 2397 TR   FILTERS:3-206 HZ
GEOPHONE TYPE:SM24  10 HZ   GEOPHONE ARRAY:6/20M    NOTCH:OUT
GEOMETRY PATCH:23 LNS X 57 STNS/LN     CHANNELS:2397     FOLD:221
SOURCE:DYNAMITE      ARRAY:2 HOLES/10M X 2 KG AT 9 M
STN INT:100 M  SP INT:100 M   SHOT LINE:200    REC LINE:200
------------------------- PROCESSING PARAMETERS --------------------------
REFORMAT
```

***Figure 6** – A typical SEGY textual header showing where the AI/ML picked values were extracted*

Conclusions

The AI/ML technology and workflow tools described have been delivered into a production KDM environment currently undergoing a pilot project. Operations teams immediately see the benefit in using these tools giving them an enhanced ability to do more, better, faster. Raster/image well log loading can now be performed at the same efficiency as standard log (LAS & DLIS) loading. The expectation is this will lead to engagement in previously non-feasible data management projects as well as improve quality and quantity of regular/existing projects. The business is eager to add more data types and extracted data values in the near future.

References

[1] Source: VGG16 K. Simonyan and A. Zisserman from Oxford University proposed this model and published it in a paper called Very Deep Convolutional Networks for Large-Scale Image Recognition

[2] Source: YOLO

[3] Source: CascadeTabNet

[4] Source: L-CNN

[5] Source: cosine similarity

[6] Source: NLTK

[7] Source: What We Know ChatGPT Can Do for the Petroleum Industry, So Far, Ouadi, H, Journal of Petroleum Technology, April 14, 2023, Co-authors Carrie Christianson, Lonny Jacobson, James Sorensen, and Aimen Laalam