GLYNT.AI

Mixed Formal Learning:
A Path to Low Shot and Zero Shot Learning

Glints of latent variables from formal models mixed with specialized models guide learning

Introduction

Most technological improvements occur incrementally, but AI stands out for its discontinuous leaps ahead. Each advancement catches people off guard as the cutting edge is pushed quickly forward. The latest advancement, which we call Mixed Formal Learning, is one such leap ahead.

Today, most high performance machine learning requires huge amounts of labeled training data. It is not uncommon to see references to training sets of 15,000 data observations or more. For most applications, collecting and preparing the large training data sets is a significant choke point, and only a few companies have data assets large enough to train well-performing models. A better solution delivers AI results with very small amounts of labeled training data. One technique is Low Shot Learning, which requires very few data with labels, 1 - 10 observations. Zero Shot Learning refers to learning formerly supervised tasks without any supervision, e.g. learning that requires
no labels.

This paper presents Mixed Formal Learning, an architecture that learns models base on formal mathematical representations of the domain of interest that exposes latent variables. The second element in the architecture learns a particular skill, typically by using traditional prediction or classification mechanisms. Our key findings include that this architecture: (1) Enables Low Shot and Zero Shot training of machine learning without sacrificing accuracy or recall; (2) Is demonstrated for the extraction of phrase and numerical data from semi-structured documents; (3) Can enable other applications with Low Shot Learning in the document domain; (4) Can be applied to enable Low Shot and Zero Shot Learning in other domains.

Related Work

Though our work at GLYNT has gone a step further, recent work at other companies has parallels. Wengong Jin, et. al [3] used an autoencoder to find latent variables for their second model to automate the design of molecules based on specific chemical properties. The autoencoder, while not apparently based on a formal model, enables 100% validity of identified molecules, an effect consistent with our performance expectations of a well-designed formal model. R Devon Hjelm, et. al [2] explore mutual information, a concept that is related to how a formal model helps to guide the second model. The paper also uses adversarial techniques to computationally learn a model of a latent space of variables. This differs from our method of defining the formal model a-priori [2]. Jacob Devlin, et. al [1] create a language representation model to enable a number of of fine-tuning models for different applications. The fine-tuning models correspond to the non-formal aspects of our architecture.

The paper by Supasorn Suwajanakorn et. al [4] of Google follows our Mixed Formal Learning architecture. The paper identifies key points, which are spatial locations that stand out on an image and remain regardless of distortion. The authors describe training two neural networks. The first network uses the orthogonal Procrustes problem as a learning method to create a 3D representation from two sequential 2D images, such as photographs. In other words, imagine two successive still-frames of a Ferrari in motion. Plug these two flat images into Google's system and it can render a representation of a 3D image of the Ferrari.

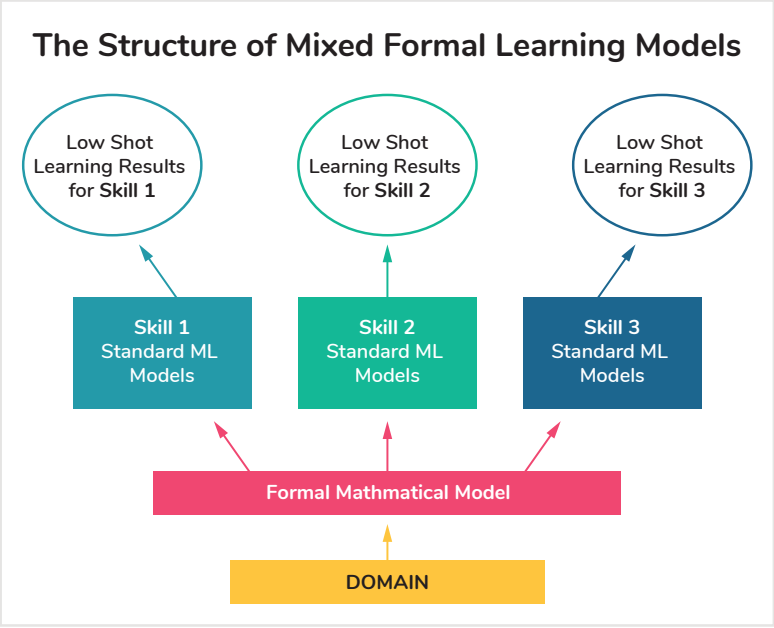
The 3D representation enables the next model to more easily find keypoints by exposing helpful latent variables. The second model's more typical neural net finds keypoints as accurately, or better, by using the exposed latent variables rather than solutions utilizing large training sets of labeled data. Because the method required no labels on the training data, the result is image keypoint identification with Zero Shot Learning.

At the conceptual level, GLYNT and Google used the same approach: an architecture of multiple models, one to model the domain and one to make predictions from the first model's outputs. The synergy between the two models often results in Zero or Low Shot training requirements — and incredible accuracy.

GLYNT's Mixed Formal Learning Implementation

Well-chosen formal representations learn models that expose key latent variables and enable subsequent models based on observed data to converge extremely fast. In addition, the models using the observed data can leverage the latent variables to learn without labeled data.

In GLYNT's application of the Mixed Formal Learning, the architecture learns a domain -specific mathematical model for semi -structured documents and then combines exposed latent variables with typical machine learning methods based on observed data. The result is an AI solution that requires dramatically less training data to achieve exceptionally accurate results. The GLYNT application of Mixed Formal Learning requires fewer than 10 examples for training to extract specific fields.



Mixed Formal Learning Enables Low Shot Results Across a Family of Models

While Mixed Formal Learning can support any number of models, the GLYNT implementation uses two primary models arranged in sequence. The first model learns latent variables based on a tightly constructed mathematical representation, tailored to the domain. Getting this representation right is key to enabling Low Shot Learning. The next model in the sequence combines the latent variables and the incoming labeled data and learns to find a phrase or numerical field and properly assign a label to it.

How does Mixed Formal Learning work synergistically to learn from such little data? Figuratively, the formal math-based model shines a glint of light on the area of the answer, allowing the second model to more quickly learn its skill. It was this glint effect that is the origin of the name of GLYNT.AI. More formally, the first model optimizes a set of latent variables that the second model exploits.

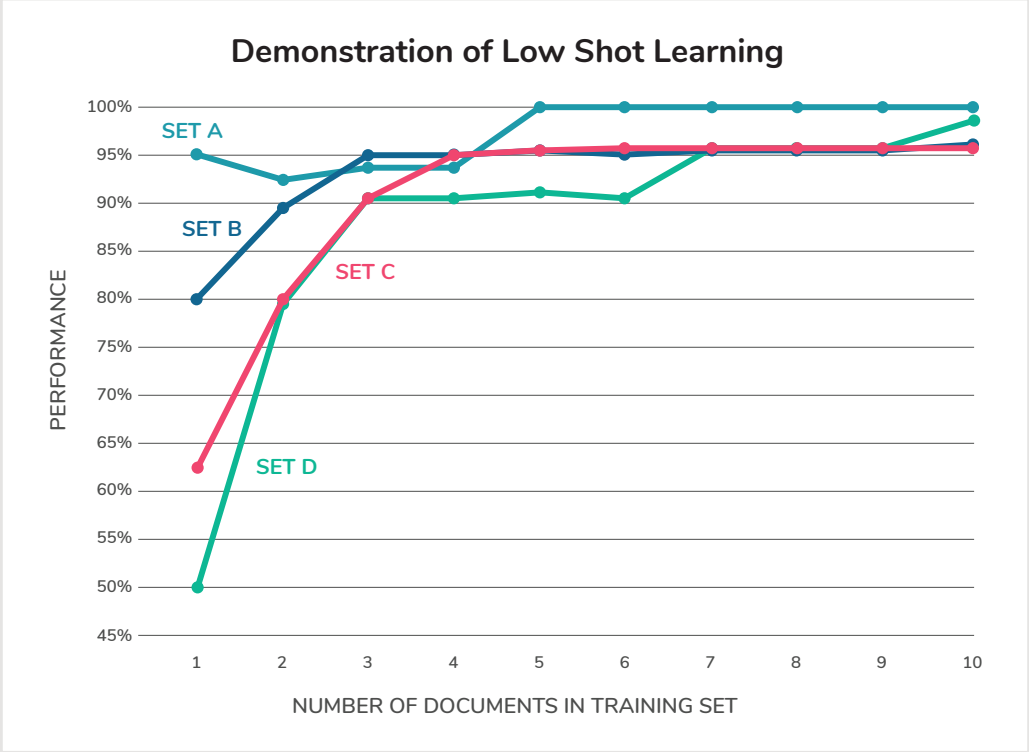
In the coming years there may be a library of domain-specific mathematical models, but there is also a sense that the formal model enables a set of capabilities shared by diverse domains. Each domain can leverage the formalism to create AI applications. GLYNT, for example, uses a domain-specific model to extract unstructured data, but the formal mathematical model can also be used in supply chain optimization and genomics data sets.

The second element in the architecture learns a particular skill, typically by using traditional prediction or classification mechanisms. While a fine-tuning approach, as is often used with embeddings and would strictly qualify as a component of this architecture, we expect our system and other applications will benefit from more ambitious models. The initial GLYNT application identifies fields according to customer-specified names.

It is notable that the formal model enables follow-on applications. Zero Shot applications should be expected for some cases. In the case of GLYNT, an application of interest would identify new fields that alert users to field-level changes in their set of of documents. We envision a Zero Shot implementation in this case. Another useful follow-on application would categorize the documents according to the publisher.

A Demonstration of Low Shot Learning

It is informative to review the results of Low Shot Learning with actual data. The graph below shows results from the extraction of unstructured data from four document sets, each with 50 documents. The documents are from the energy industry (2 different utility bills) and the healthcare industry (insurance cards and clinic visit summaries). 10 documents in each set were reserved for training and 40 documents in each set form a holdout set. The graph shows how well the data extraction performs as the size of the training document set increases (For more details see our related paper: Low Shot Learning in Action).



Extraction Results from Four Data Sets, 40 Documents Each

The F1 score is displayed on the graph for each document set. The key result is F1 score performance at 95% and above is obtained with less than 10 documents in the training set. As the graph indicates, the gain in performance levels off at about 7 documents. Most of the improvement in performance derives from improvements in recall, as accuracy remains fairly uniform. In previous work, we have documented the high performance of the GLYNT data extraction system. The results here show that with the addition of Low Shot Learning, GLYNT is able to achieve this level of performance on very little training data. This is in sharp contrast to other machine learning systems that require training data sets of 15,000 to 300,000 documents and months of work for similar tasks. GLYNT produces results in under an hour.

Conclusion

Low Shot Learning liberates unstructured data better, faster and cheaper than ever before. It also eliminates the advantage of large data sets in the AI ecosystem, creating a new landscape where advanced AI products and solutions are available to many researchers and data scientists, not just to the lucky few with huge data assets.

We expect to see three extensions of this work. First, the formal mathematical models of Mixed Formal Learning is tightly tailored to the domain at hand, but with mathematical insight, the same model can be used in seemingly disparate domains. As mentioned, the GLYNT formal model might well capture aspects of the domains of genomic data analysis and supply chain optimization. Second, other researchers are edging towards the same architecture, and enabling Low Shot Learning in structurally different applications. Third, the second model of Mixed Formal Learning provides automated skills. In the GLYNT application the skill is extraction of unstructured data. Additional skills can be used with GLYNT's formal model, such as new field identification or document classification. It is expected that the new skills will also be Zero or Low Shot and need very little training data.

Rapid advances within the AI field abound. Disrupting the need to use large training data sets changes who wins and loses by the adoption of AI. Mixed Formal Learning is a powerful new approach to the structure of machine learning models.

References

[1] BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
J Devlin, MW Chang, K Lee, K Toutanova
arXiv preprint arXiv:1810.04805

[2] Learning deep representations by mutual information estimation and maximization
RD Hjelm,A Fedorov, S Lavoie-Marchildon, K Grewal, A Trischler
arXiv preprint arXiv:1808.06670

[3] Junction Tree Variational Autoencoder for Molecular Graph Generation
W Jin, R Barzilay, T Jaakkola
arXiv preprint arXiv:1802.04364

[4] Discovery of latent 3D keypoints via end-to-end geometric reasoning
S Suwajanakorn, N Snavely, J Tompson, M Norouzi
arXiv preprint arXiv:1807.03146, 2018