SambaNova SYSTEMS

WHITE PAPER

Boost Accuracy in Domain-Specific NLP With SambaNova's Solutions

Modern industrial and research NLP models follow a cadence for their domain fine-tuning task training pipelines. They usually are initialized from off-the-shelf pre-trained weights — such as bert-large-uncased from HuggingFace's model zoo — and then are fine-tuned on task-specific and domain-specific datasets before deployment to serve massive online inference requests from application users.

Recent research has demonstrated that this industry standard pipeline does not capture domain-specific knowledge during language modeling, and can lead to suboptimal statistical metrics on any domain-specific fine-tuning task.[1] [2] [3]

Gururangan et al.[1] demonstrate how class-wise F1 scores on Hyper-Partisan Detection,[4] a task from the News Domain, can be increased from 86.6 to 88.2 by simply tuning the pre-training corpus to have more relevant domain exposure.

What's Inside

Introduction

One of the main reasons behind the lack of widespread adoption of domain-focused training is due to the enormous time, compute, and human capital investments necessary to implement this specialized training. SambaNova Systems, a full-stack hardware-software company, provides simple one-click solutions for both the hardware and software challenges involved in building these pipelines. This allows users to maximize downstream statistical accuracy for domain-specific NLP pipelines with minimal effort.

By enabling domain-specific pre-training, we demonstrate that such solutions can offer superior statistical performance compared to the standard off-the-shelf industry standard pipelines.

The Power of Domain Knowledge

Current industrial workflows initialize their fine-tuning pipelines from weights such as the bert-large-uncased weights from HuggingFace.[5] These weights do not have domain exposure as they tend to be from the general domain, often pre-trained on corpuses like Wikipedia English[6] and the Books Corpus.[7]

These corpuses have a limited exposure to domain-specific vocabulary. For instance, only about 0.004% of English Wikipedia's sentences contain the word "plaintiff," a word relevant to nearly all legal proceedings. On the other hand, if we consider a legal dataset such as CaseHOLD,[8] over 33% of that dataset's sentences contain the same word.

It is clear that a model exposed to CaseHOLD would have more legal domain emphasis than one trained on English Wikipedia. We showcase how this increased emphasis manifests itself by outperforming the standard industry pipeline on tasks from the legal, financial, and biomedical domains.

By enabling domain-specific pre-training, we demonstrate that such solutions can offer superior statistical performance

Domain Pipelines

In our solution, we implement two pipelines to enable domain exposure during pre-training and compare them to the standard pipeline utilized by many industrial cadences today, whereby bert-large-uncased from HuggingFace is used to fine-tune without any consideration to domain exposure.

The first pipeline pre-trains on Eleuther's PILE dataset,[9] a massive corpus with exposure to BioMedical and Legal domains. The second pipeline introduces an intermediate stage to the NLP cadence after the pre-training but before the fine-tuning stage, where we do language modeling on the fine-tuning dataset itself, as described in Gururangan et al.[1]

SambaNova's two pipelines, along with the current industry standard pipeline, are shown in Figure 1, with the differences highlighted in the dark rectangular boxes. We refer to these two pipelines as "Domain-Customized Pipeline 1" and "Domain-Customized Pipeline 2."

To emphasize the singularly important role of data in this process, we maintain the same vocabulary and model architectures across all three pipelines in this blog post. The vocabulary and model architectures can be further tuned with SambaNova's solutions for potentially even better performance.

Figure 1: Augmented parts of SambaNova's pipelines, compared to the industry standard, are shown in the dark rectangular boxes.

Out-Performance of Domain-Customized Pipelines

We demonstrate superior statistical performance on text classification on CaseHOLD (Legal), FiQA+PhraseBank[10] [11] (Finance), and ChemProt[12] (BioMedical) using either of the two Domain-Customized pipelines, compared to the current industry standard pipeline, without any changes to the model architecture or the vocabulary the model learns.

Legal Domain — CaseHOLD

We observe a 1.87% increase in seed-average test accuracy score when using Domain-Customized Pipeline 2 over the industry standard pipeline, and a 1.36% increase in seed-average test accuracy score when using Domain-Customized Pipeline 1 over the industry standard pipeline.

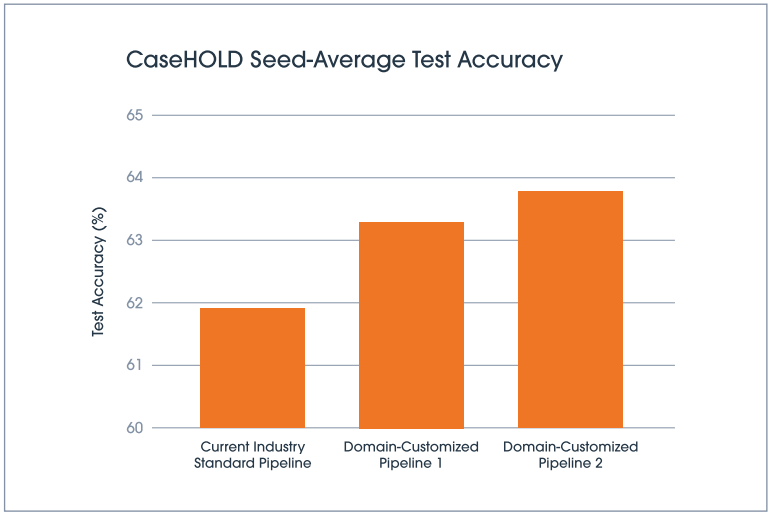


Figure 2: The seed-average test accuracy on CaseHOLD classification

BioMedical Domain — ChemProt

We observe a 1.37% increase in seed-average test accuracy when using Domain-Customized Pipeline 2 over the industry standard pipeline, and a 0.27% increase in seed-average test accuracy when using Domain-Customized Pipeline 1 over the industry standard pipeline.

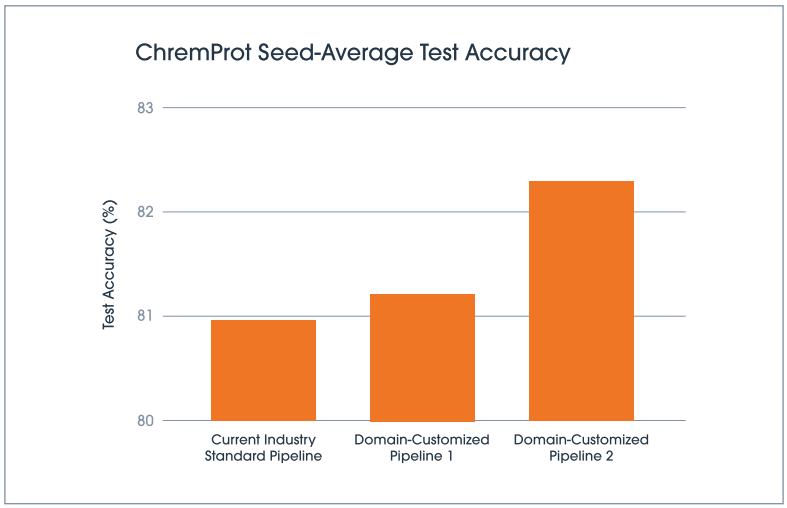


Figure 3: The seed-average test accuracy on ChemProt relation classification

Finance Domain — FiQA+PhraseBank

Due to the small size of FiQA and PhraseBank, we combined the datasets together before performing sentiment classification. We observe a 1.88% increase in seed-average test accuracy when using Domain-Customized Pipeline 2 over the industry standard pipeline. As the PILE dataset has no significant financial exposure, we do not report the performance of Domain-Customized Pipeline 1 as it is not relevant to the central message.

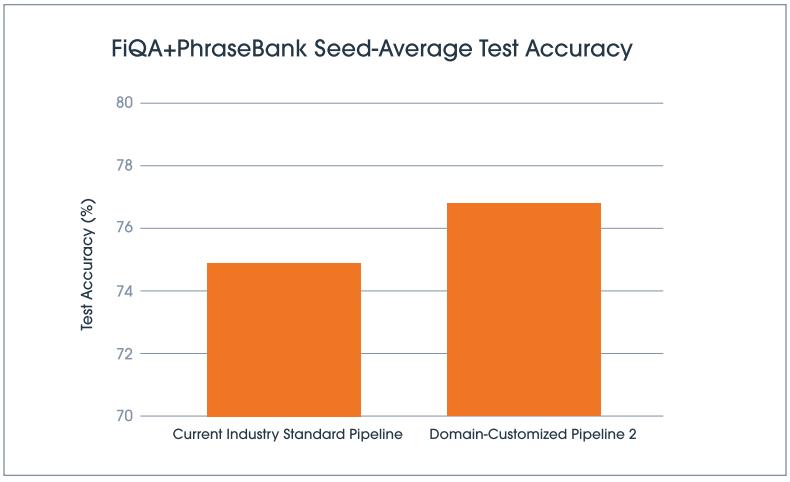


Figure 4: The seed-average test accuracy on FiQA+PhraseBank combined classification

What Do You Need to Do?

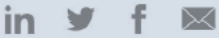
You should consider the following areas in preparation for NLP:

- Determine whether the business problem you are trying to solve is domain-specific or not. For example,are you trying to develop a general-purpose sentiment analysis solution or a domain-specific one?
- If you are trying to solve a domain-specific problem, like most SambaNova customers do, such as sentiment analysis in finance, you will need to ensure you have access to NLP models pre-trained based onpipelines with domain exposure.
- SambaNova offers a purpose-built hardware and software stack that makes it possible to substantially decrease the amount of time, compute, and human capital investment necessary to formulate and set up these domain-exposed pipelines for a given workflow.

To learn more, please visit our NLP solutions web page.

[1] https://arxiv.org/pdf/2004.10964.pdf
[2] https://arxiv.org/pdf/1908.10063.pdf
[3] https://arxiv.org/pdf/1903.10676.pdf
[4] https://pan.webis.de/semeval19/semeval19-web/
[5] https://huggingface.co/bert-large-uncased
[6] https://en.wikipedia.org/wiki/English_Wikipedia
[7] https://yknzhu.wixsite.com/mbweb
[8] https://arxiv.org/pdf/2104.08671.pdf
[9] https://pile.eleuther.ai
[10] https://sites.google.com/view/fiqa/
[11] https://arxiv.org/pdf/1307.5336.pdf
[12] https://pubmed.ncbi.nlm.nih.gov/20935044

Learn more at SambaNova.AI



SambaNova Systems is an AI innovation company that empowers organizations to deploy best-in-class solutions for computer vision, natural language processing, recommendation systems, and AI for science with confidence. SambaNova's flagship offering, Dataflow-as-a-Service, helps organizations rapidly deploy AI in days, unlocking new revenue and boosting operational efficiency. SambaNova's DataScale® is an integrated software and hardware system using Reconfigurable Dataflow Architecture (RDA), along with open standards and user interfaces. Headquartered in Palo Alto, California, SambaNova Systems was founded in 2017 by industry luminaries, hardware, and software design experts from Sun/Oracle and Stanford University. Investors include SoftBank Vision Fund 2, funds and accounts managed by BlackRock, Intel Capital, GV, Walden International, Temasek, GIC, Redline Capital, Atlantic Bridge Ventures, Celesta, and several others. For more information, please visit us at sambanova.ai or contact us at info@sambanova.ai. Follow SambaNova Systems on LinkedIn.