New Knowledge

Character-Level Convolutional Neural Networks for Semantic Classification

Paul Azunre & Numa Dhamani

What is Semantic Classification?

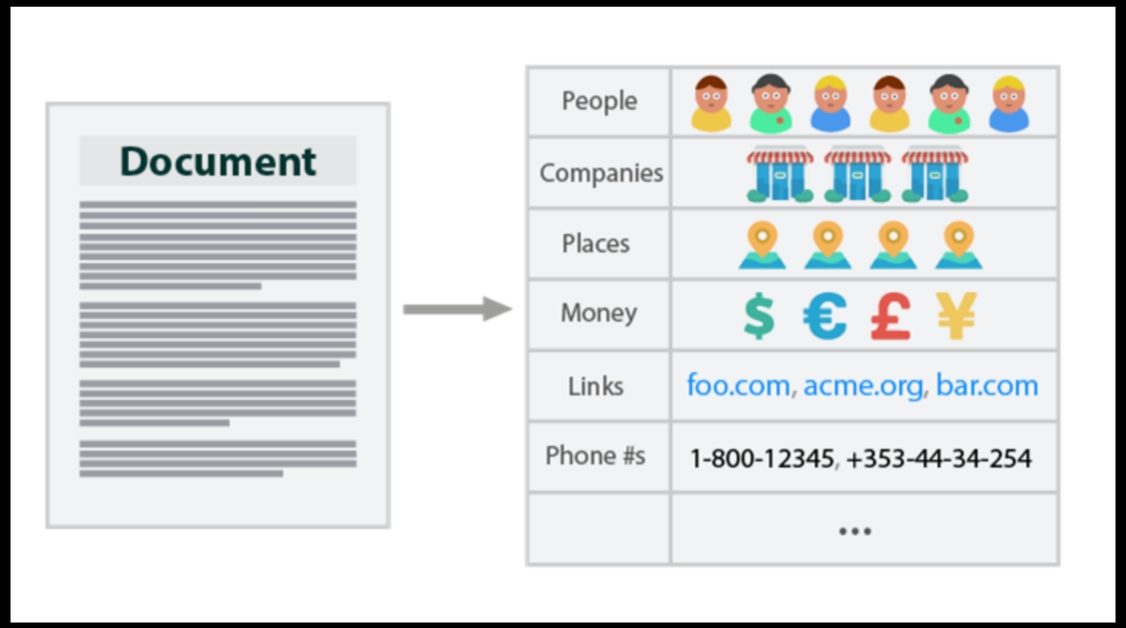


Image source: Aylien

Introducing **SIMON** - **S**emantic **I**nference for the **M**odeling of **On**tologies.

Key Highlights

Open-Source

SIMON is an open-source text classification tool.

Character-Level CNN

Semantically classifies columns in a tabular dataset using character-level convolutional neural networks (CNNs).

AutoML

Motivated by AutoML in DARPA's D3M program.

Transfer Learning

Relies on transfer learning for flexibility and reduced data/computing requirements.

Defined domain

Feature engineering

Contextual information

Language Engineering

Let’s talk about traditional text classification algorithms…

Knowledge repository

Generate features

Unable to handle misspellings and emoticons

Specific language style

Why Transfer Learning?

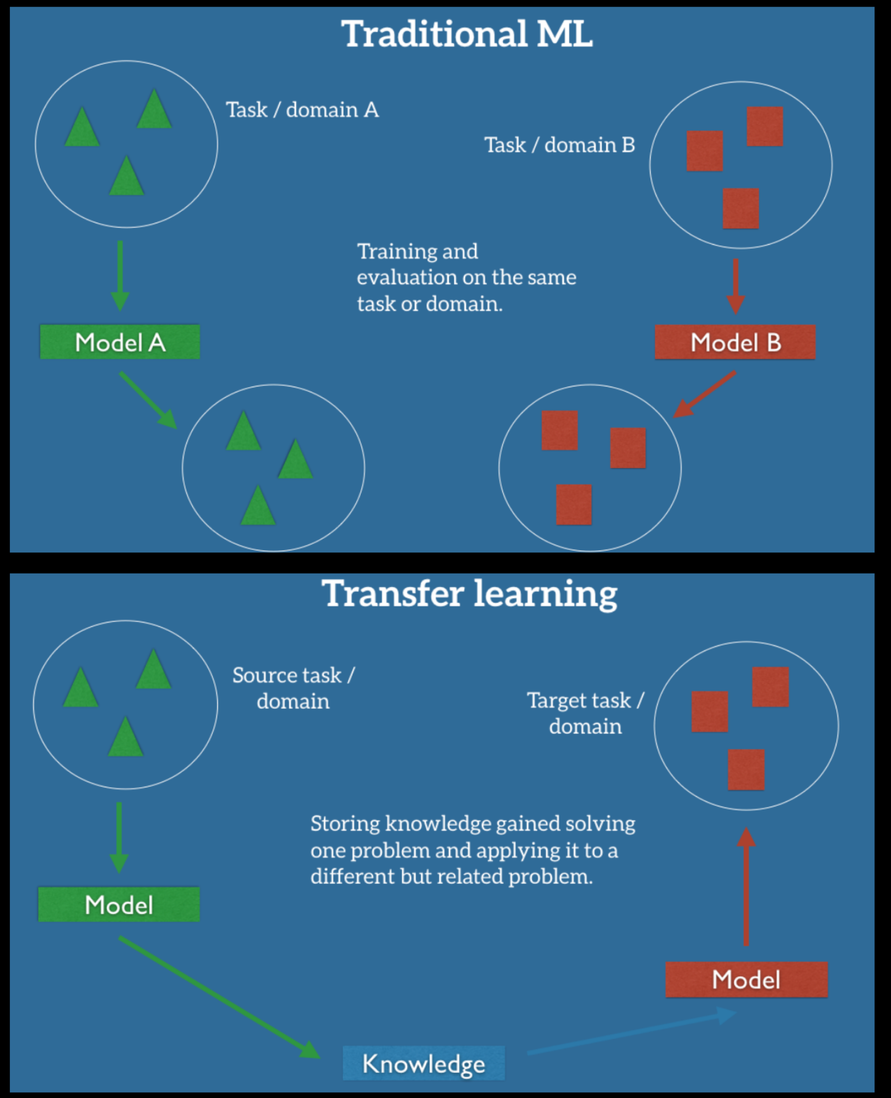


Image source: ruder.io

Multi-Class & Multi-Label Classification Problem Formulation

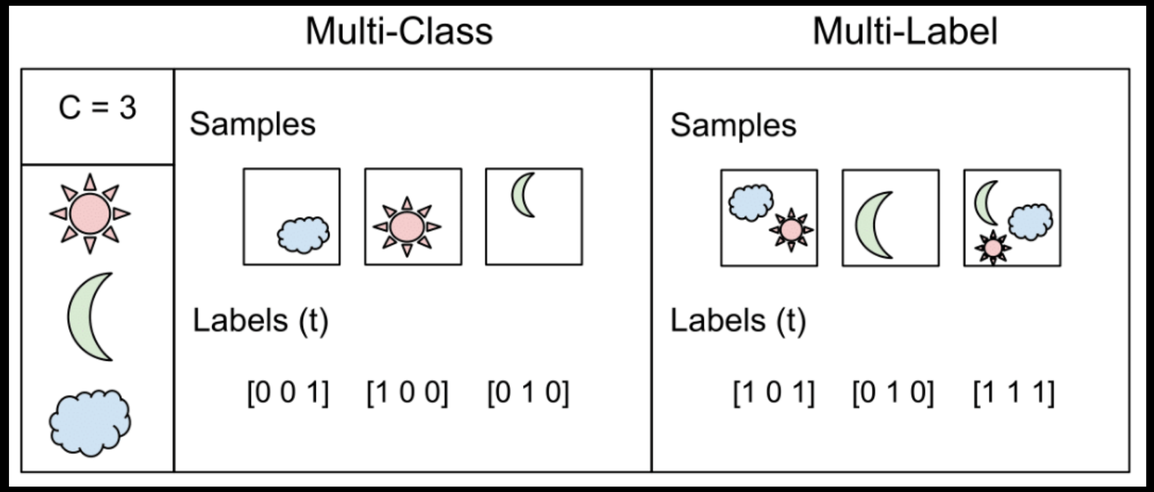


Image source: gombru.github.io

SIMON's **overall architecture** consists of two components.

Two Components

Network that encodes each individual sentence

Connects 13 layers – convolutional, max-pooling, fully-connected, etc.

Network that encodes the document as a whole

Connects 7 layers.

Character-based Convolutional Neural Networks

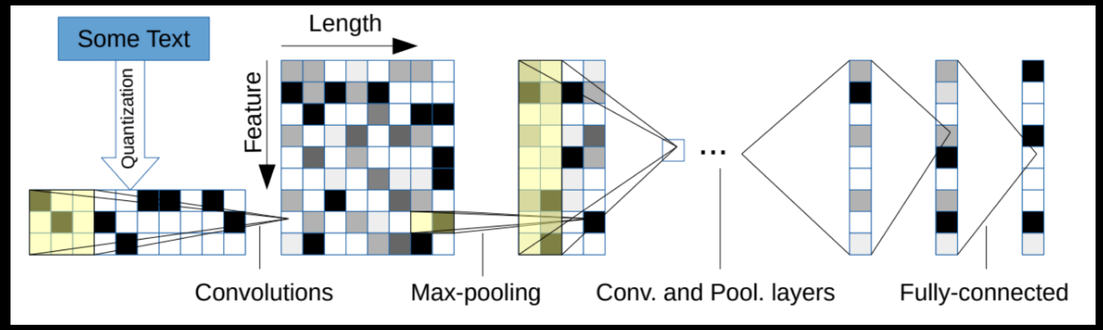


Image source: machinelearningmastery.com

Bidirectional Long-Short Term Memory Network

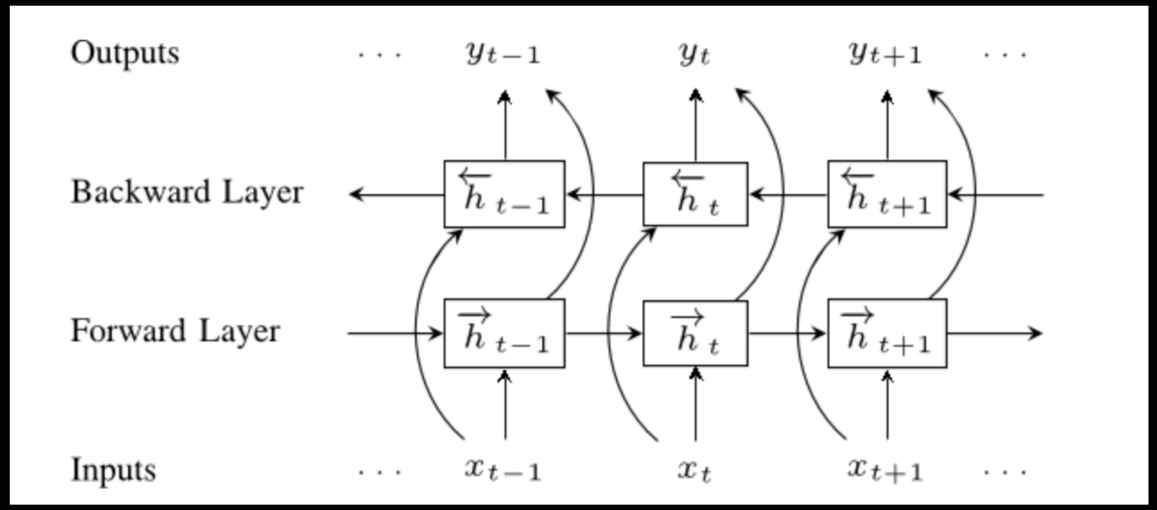
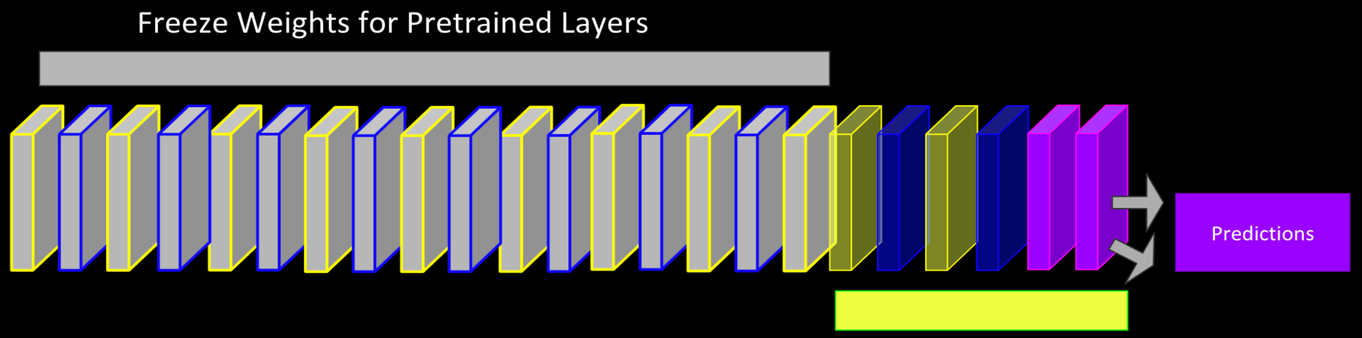


Image source: towardsdatascience.com



Transfer Learning Steps

Tabular Semantic Classification

98.4%

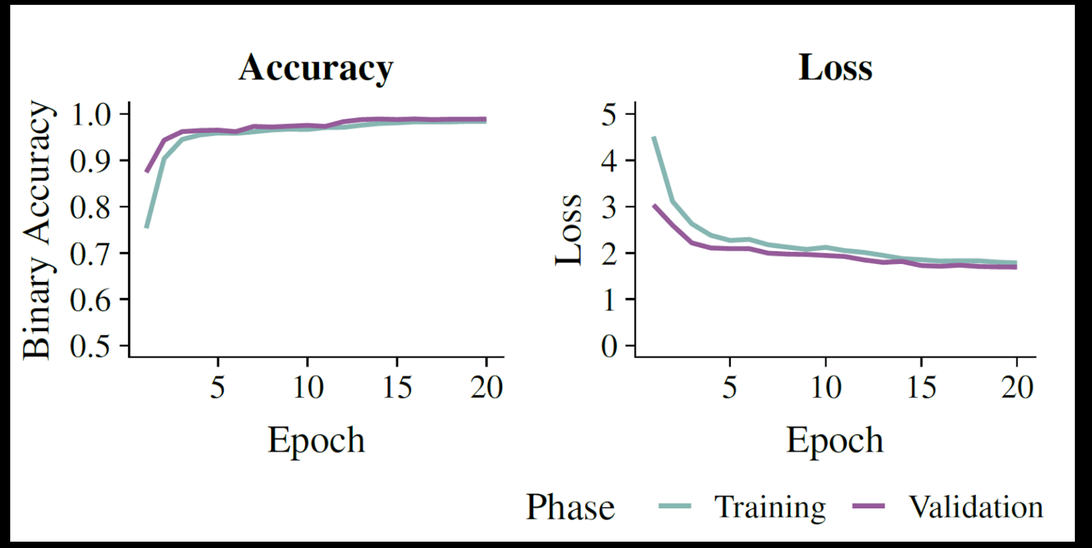
Training Binary Accuracy

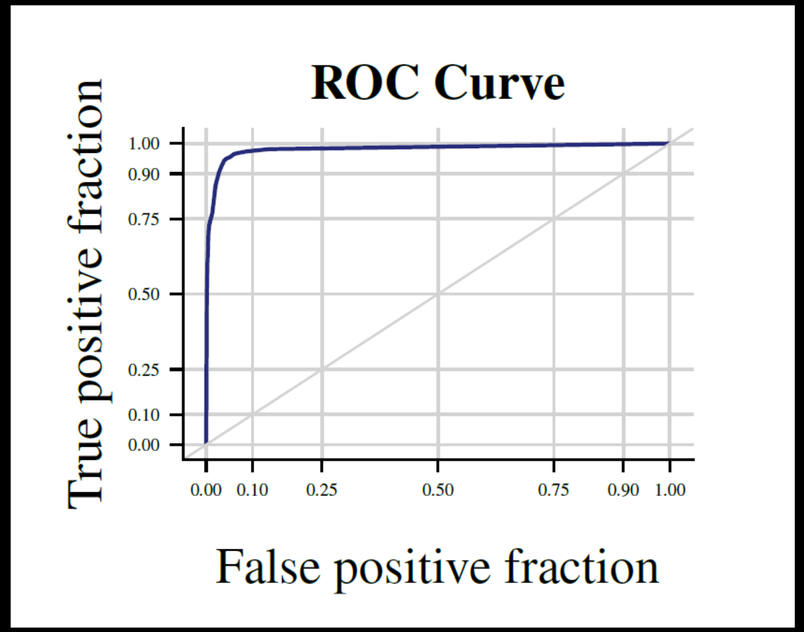
98.9%

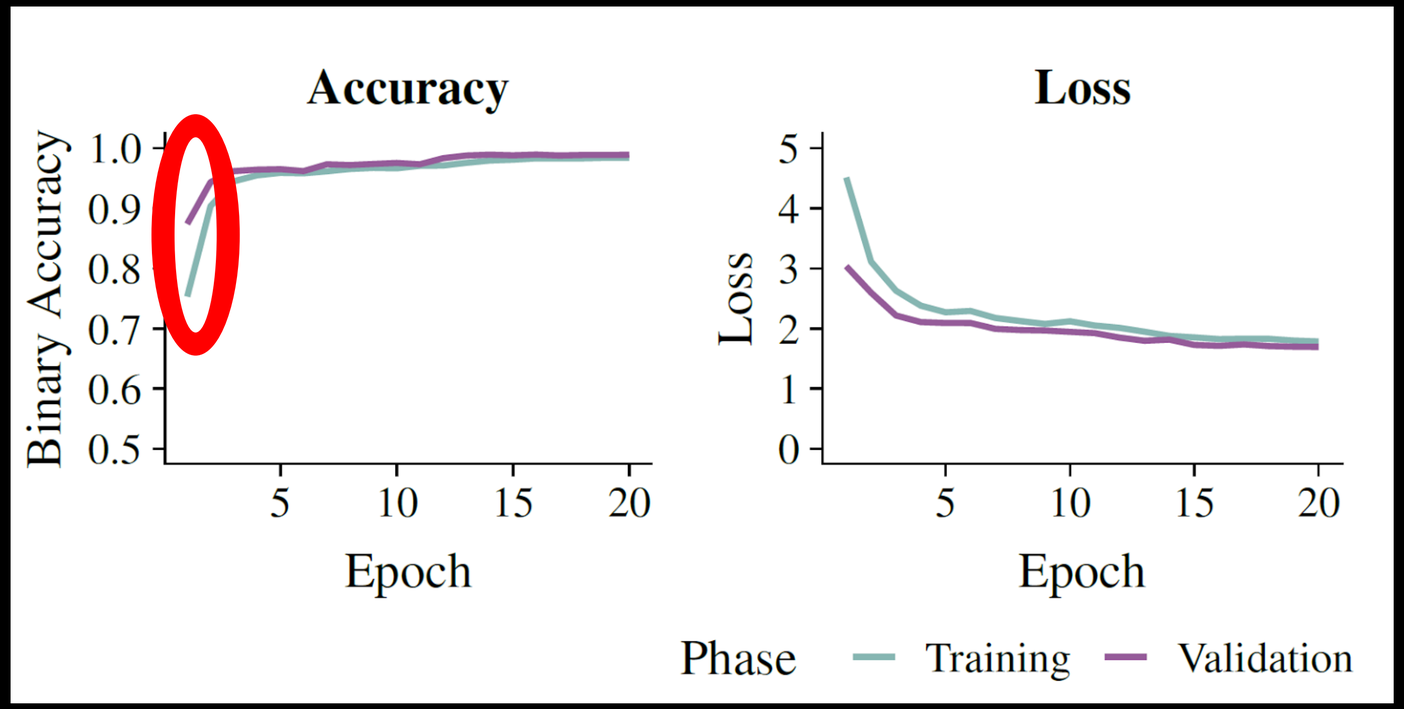
Validation Binary Accuracy

98.8%

Test Binary Accuracy







Tabular Semantic Classification

Tabular Semantic Classification Evaluation for **38 manually-annotated** D3M Datasets.

Similarity Score - percentage of labels, in which any of the annotations match the manual annotations.

92%

SIMON’s annotations compared to D3M datasets.

71%

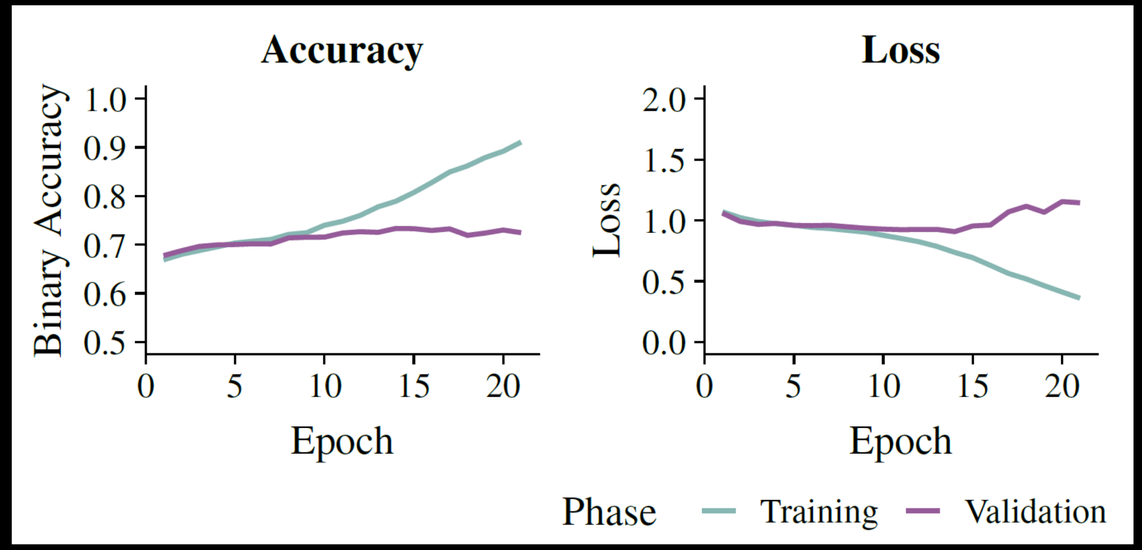
Pandas annotations compared to D3M datasets.

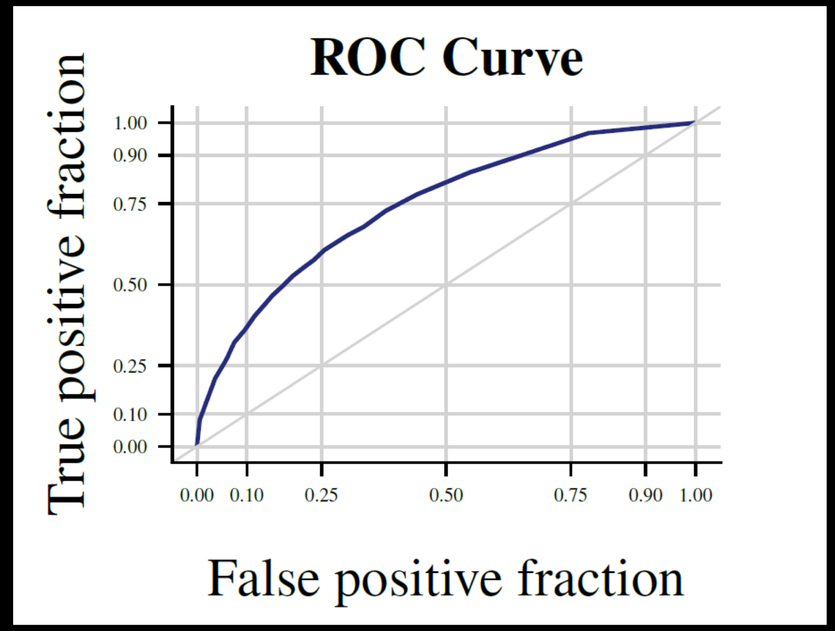
Twitter Age Prediction

79.0%
Training Binary Accuracy

73.4%
Validation Binary Accuracy

70.9%
Test Binary Accuracy





Email Spam Classification

98.7%

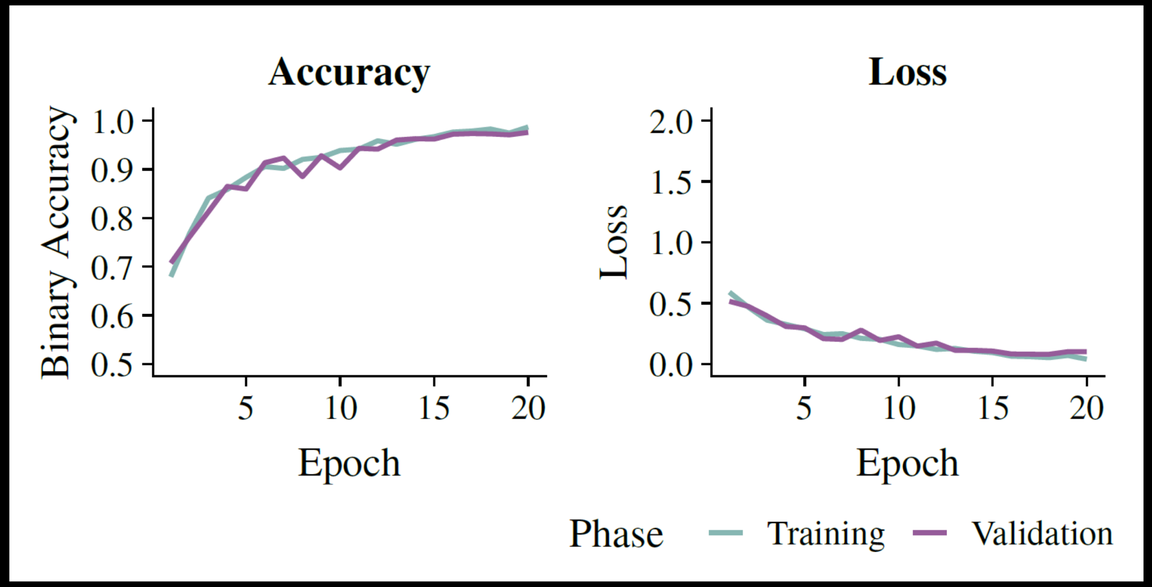
Training Binary Accuracy

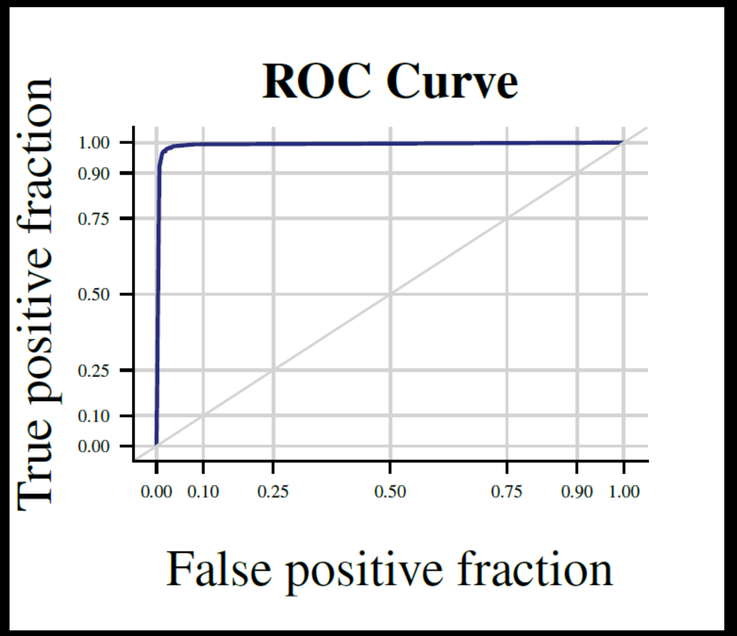
97.6%

Validation Binary Accuracy

97.9%

Test Binary Accuracy





https://github.com/NewKnowledge/simon

We would like to acknowledge the contributions of the following people.

Authors

Paul Azunre, Craig Corcoran, Numa Dhamani, Jeffrey Gleason, Garrett Honke, David Sullivan, Rebecca Ruppel, Sandeep Verma, Jonathon Morgan.

DARPA

- Work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract Number D3M (FA8750-17-C-0094).
- Views, opinions, and findings contained in this report are those of the authors and should not be construed as an official Department of Defense position, policy, or decision.

References

Please see https://arxiv.org/abs/1901.08456 for references.

We’re hiring!

www.newknowledge.com/careers

New Knowledge

Thank you

paul@newknowledge.io
numa@newknowledge.io
@NewKnowledgeAI