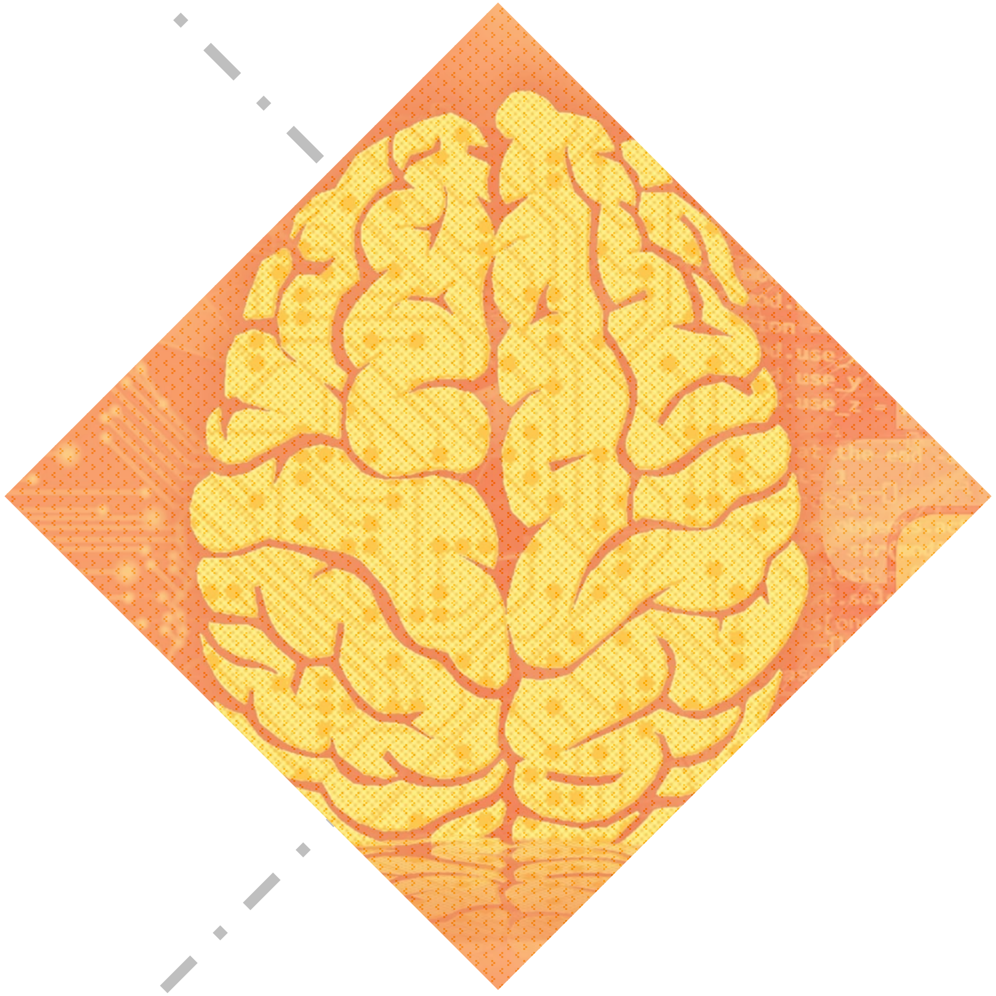
BERTVision

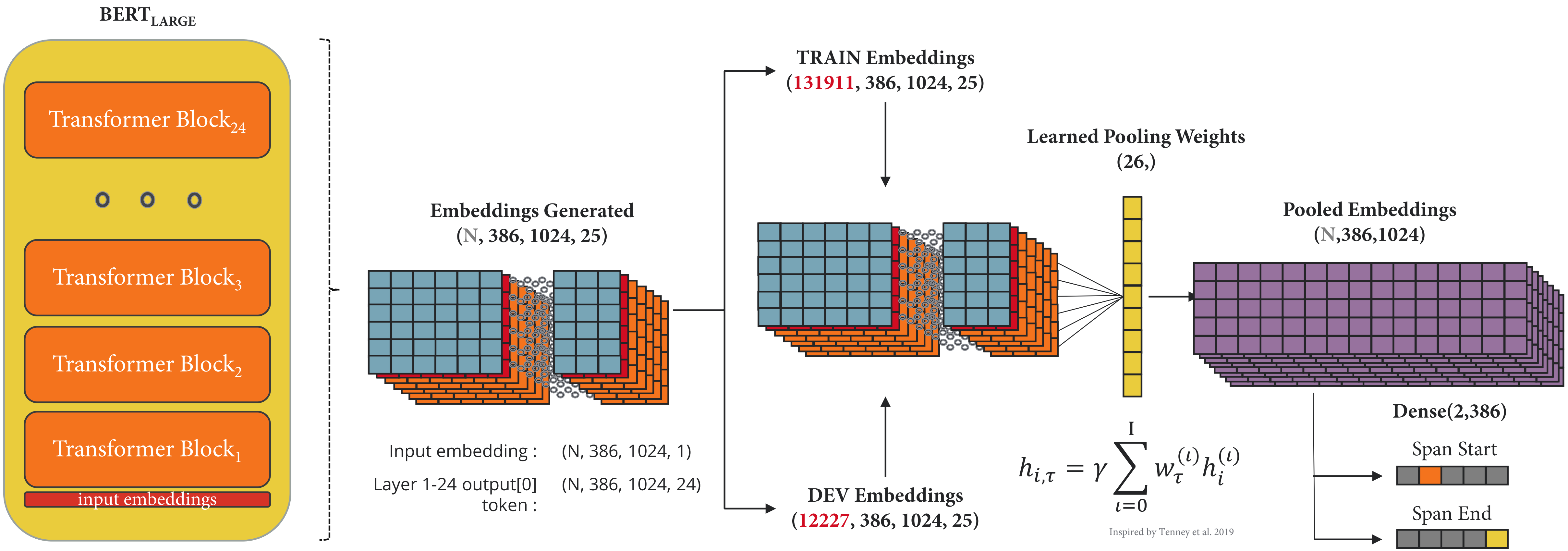
Improving performance on a wide range of **N**atural **L**anguage **P**rocessing tasks using parameter-efficient model architectures training on BERT’s hidden state activations



Impact & Feasibility

- Transformer models are powerful, but have a high parameter count and computational footprint.
- Other compression techniques prune parameters or reduce numeric precision, but at a cost.
- SOTA progress in NLP with transformers is on a collision course with absurdity (1T+ parameters).
- Our current method still requires minor fine-tuning of BERT, but we’re evaluating transfer learning with the smaller models in this project.
- If our solution ultimately requires BERT weights, we could offer inferencing as-a-service through REST API.

Introducing: BERTVision



Preliminary Work

SQuAD 2.0

BERT-Large and BERTVision on Q&A and binary classification

Cloud Infrastructure

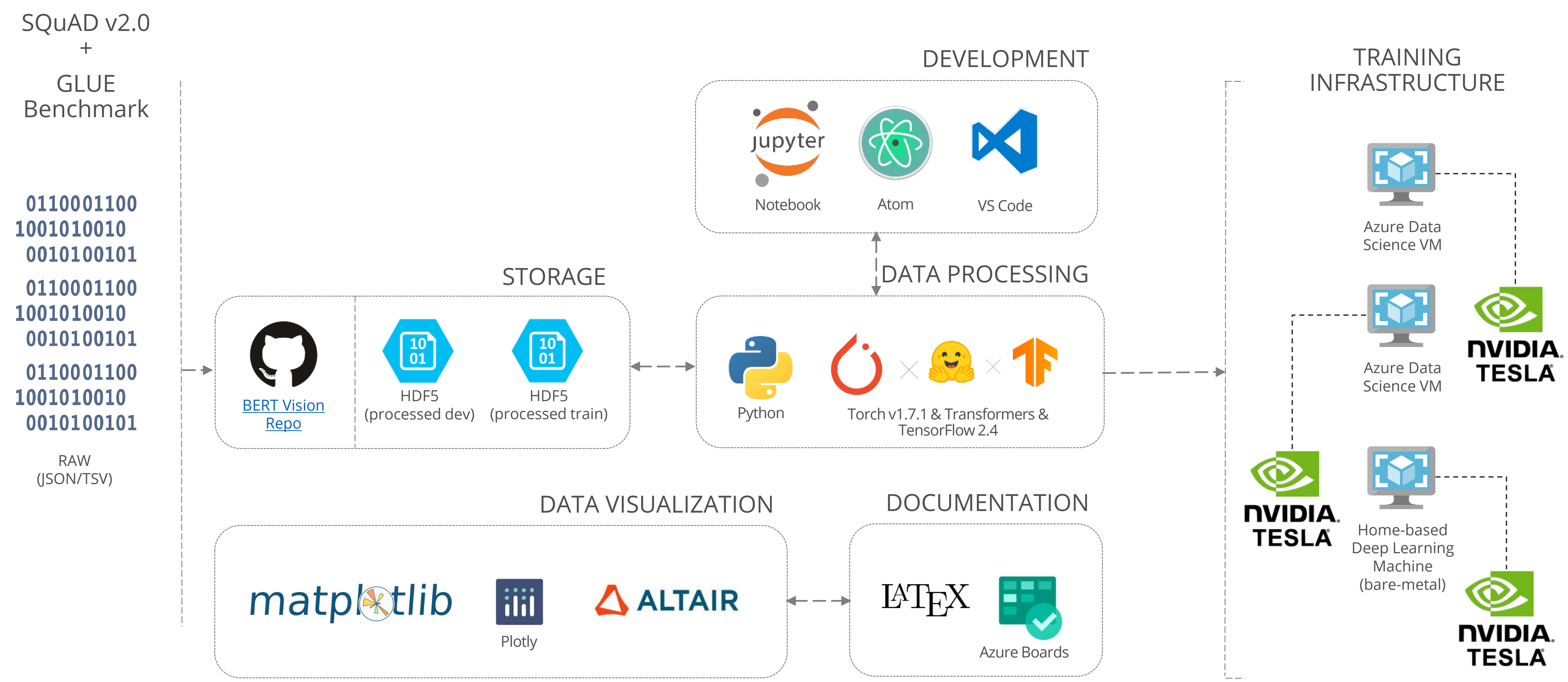
GPUs and VMs and Azure, oh my!

GLUE Benchmark

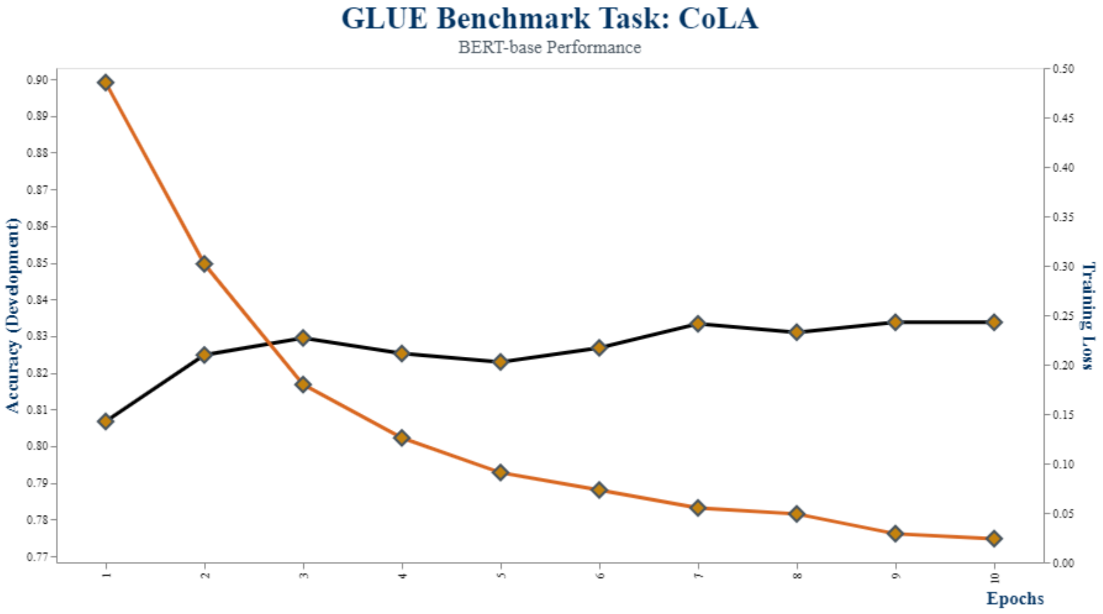
BERT-base benchmarks against all GLUE NLP tasks

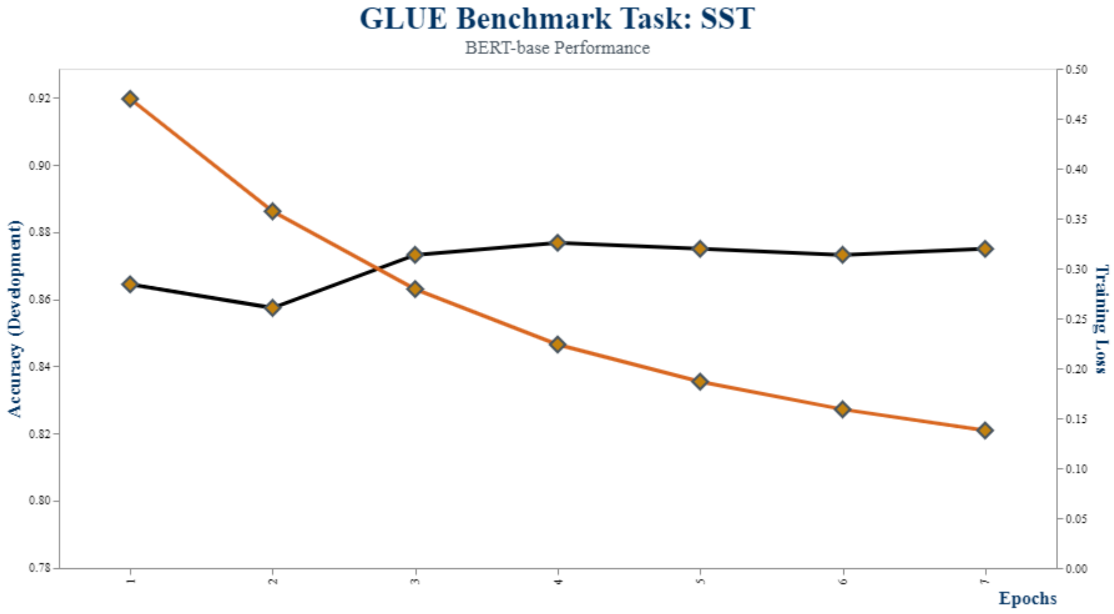
We have a robust pipeline of work completed to-date, including most of our infrastructure needs, a preliminary paper in Overleaf from our results in W266 against SQuAD 2.0, and up-to-date benchmark data on all GLUE NLP tasks for the BERT-base (110M parameter) transformer model

BERTVision Development Pipeline

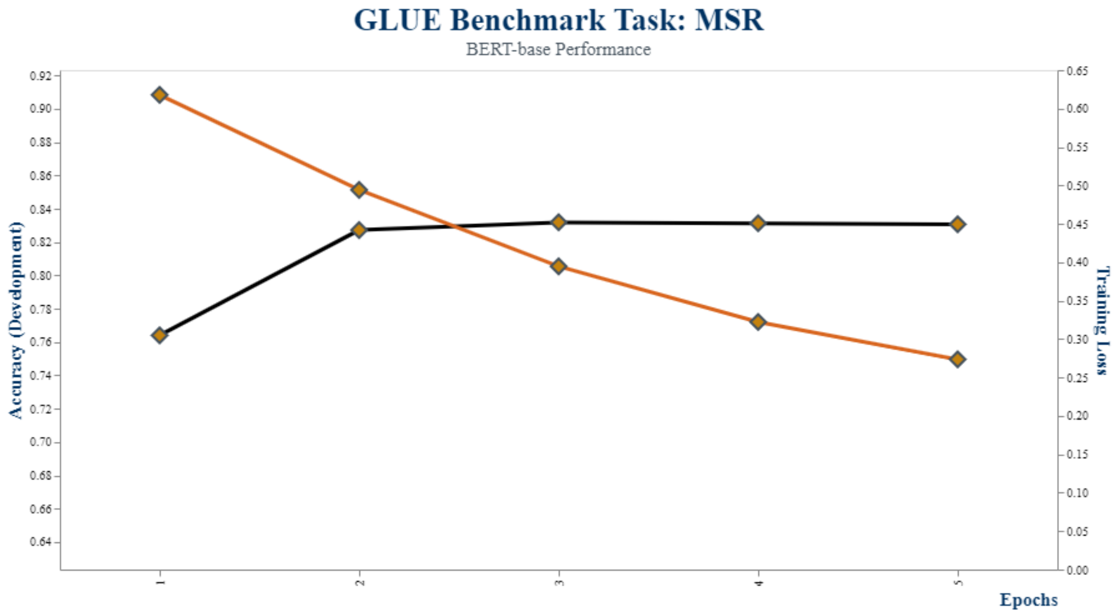


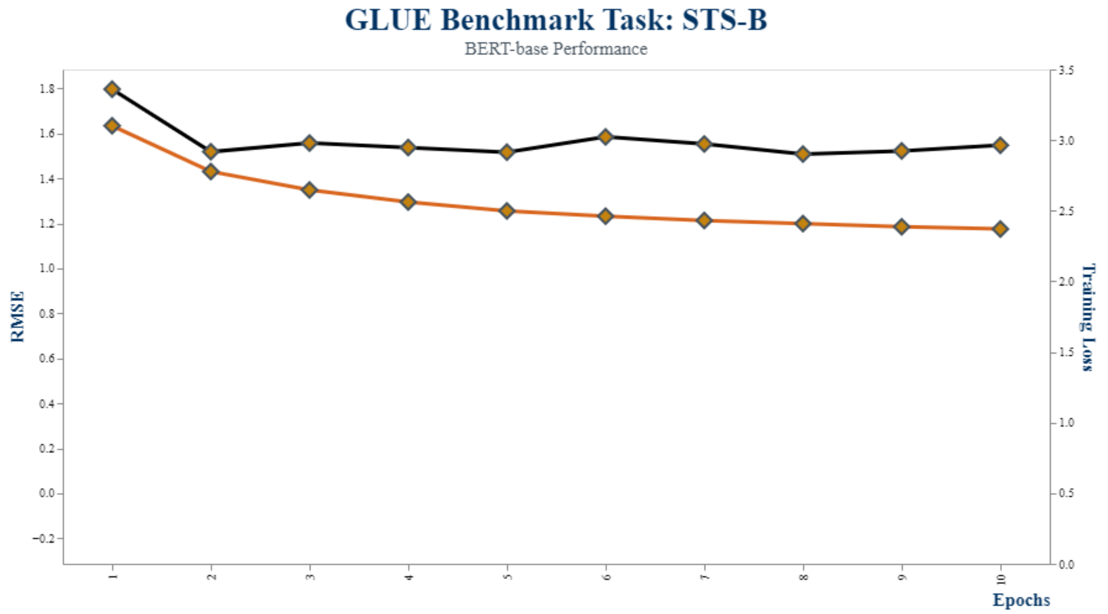
BERTVision GLUE Benchmarks*

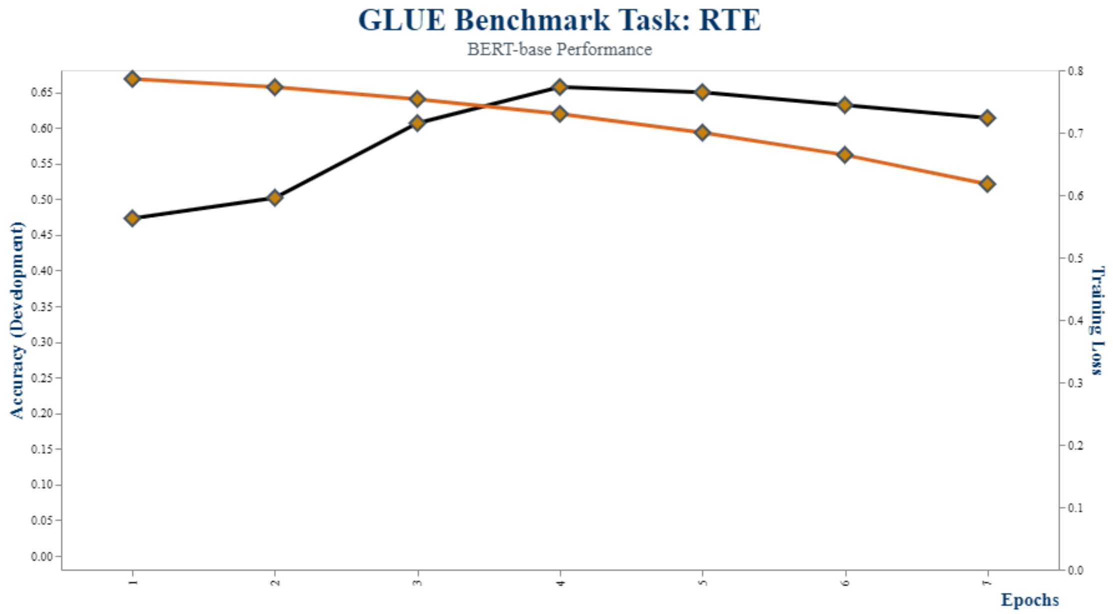


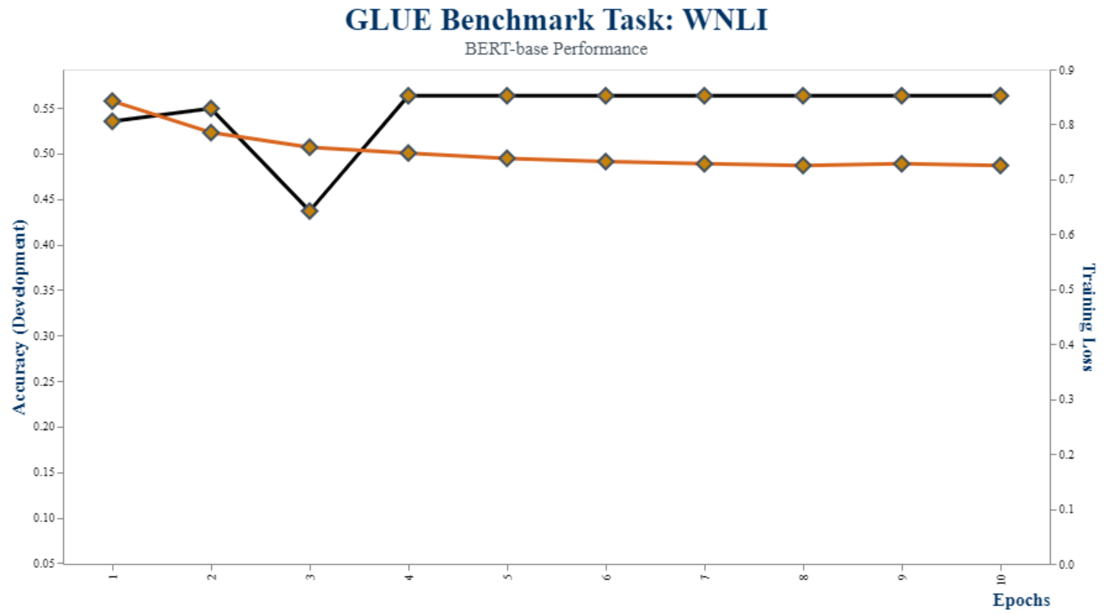


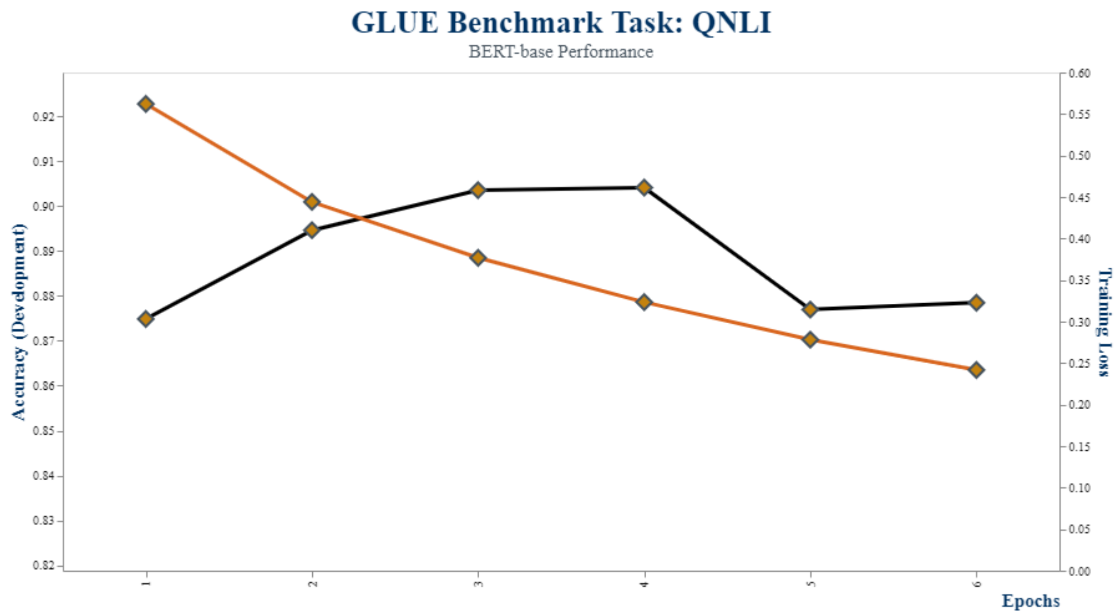
*BERT-base performance


GLUE Benchmark Task: MSR
BERT-base Performance
Accuracy (Development)
Training Loss
Epochs
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0.70
0.68
0.66
0.64
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
1
2
3
4
5

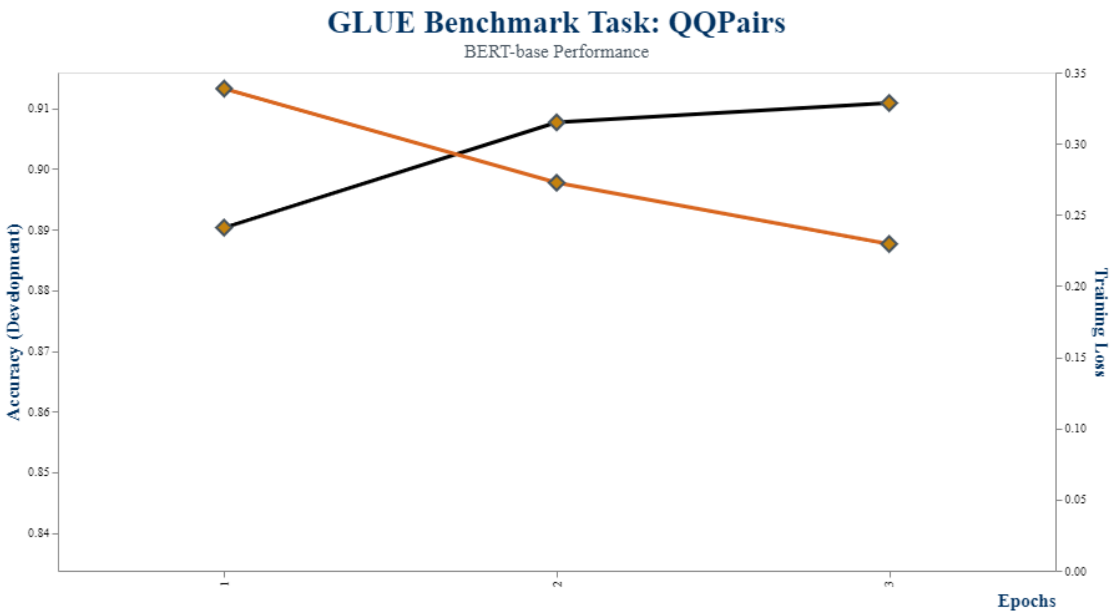

GLUE Benchmark Task: STS-B
BERT-base Performance
RMSE
Training Loss
Epochs
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6
7
8
9
10


GLUE Benchmark Task: RTE
BERT-base Performance
Accuracy (Development)
Training Loss
Epochs
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7


GLUE Benchmark Task: WNLI
BERT-base Performance
Accuracy (Development)
Training Loss
Epochs
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
1
2
3
4
5
6
7
8
9
10



GLUE Benchmark Task: QNLI
BERT-base Performance
Accuracy (Development)
Training Loss
Epochs



GLUE Benchmark Task: QQPairs
BERT-base Performance
Accuracy (Development)
Training Loss
Epochs

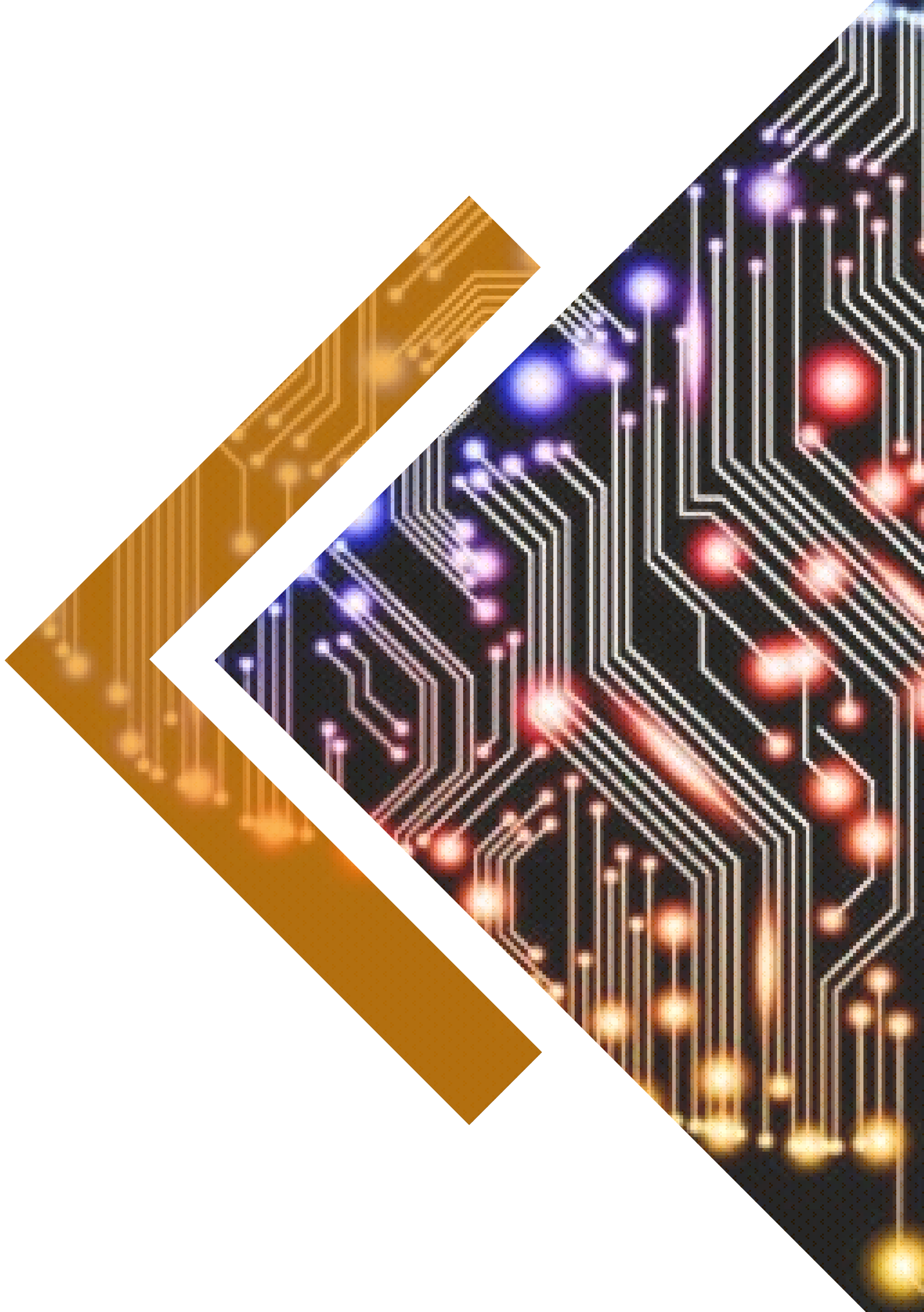
Semester Plan

We have essentially 9 weeks to complete baseline of dozens of large (340M+ parameter) models, evaluate transfer learning tasks between datasets and between task categories for BERTVision models, interpret results, and publish a high-quality paper aimed at the EMNLP 2021 conference [held in the sunny Dominican Republic on November 7th – 11th].

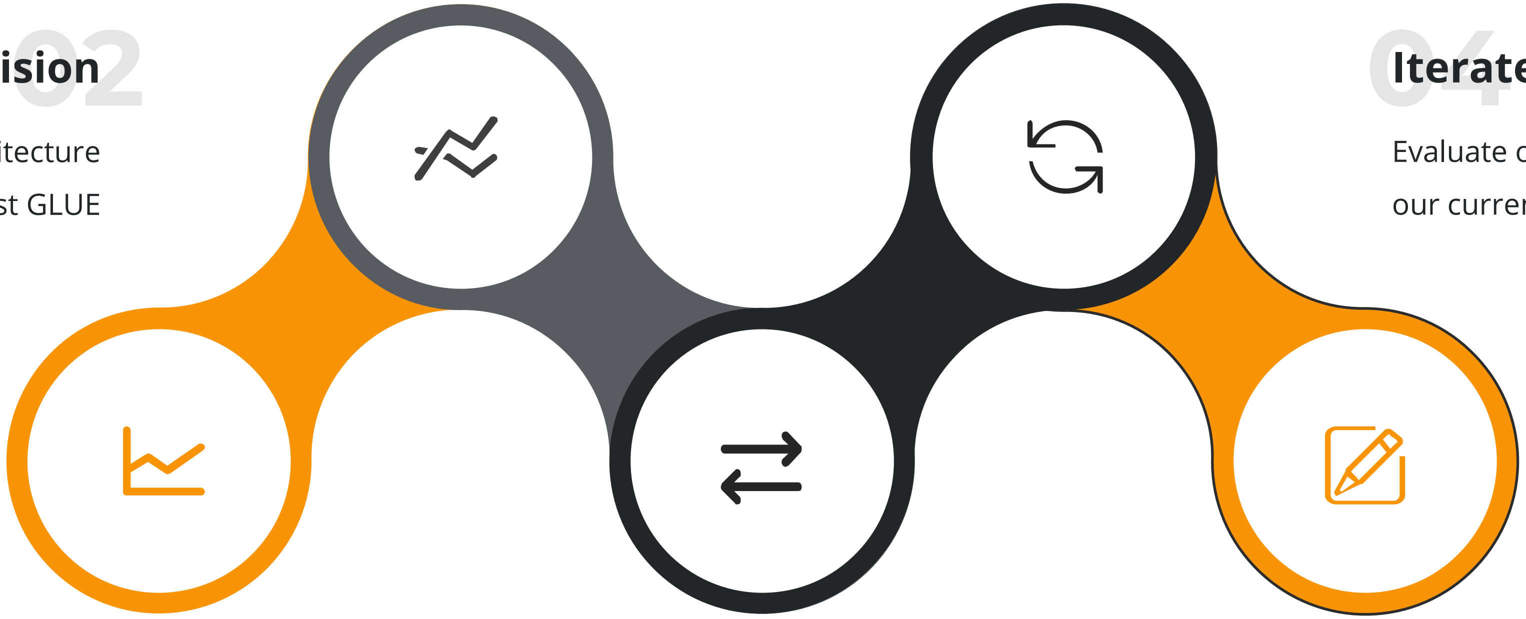
"The future is uncertain, and the end is always near." –

Jim Morrison, Roadhouse Blues (1970)

Our aim is to test a novel idea; the future of which is uncertain. if transfer learning is not as successful as we hope, we will need to shift focus to the value of extending BERT performance cheaply with our already demonstrated technique of embedding extraction and linear learning.



Remaining Research



01 Baseline BERT-large

Capture BERT-large baseline performance against GLUE tasks

02 Baseline BERTVision

Baseline the current architecture of BERTVision against GLUE

03 Transfer Learning

Evaluate BERTVision mini-models on transfer learning tasks

04 Iterate on Model Design

Evaluate other approaches beyond our current adapter-linear design

05 Publish

Work with Alberto, Puya, and Daniel to author a journal quality paper (targeting EMNLP 2021)

Success Evaluation

Success will depend greatly on our ability to live up to the vision of the project: to provide a novel approach to NLP tasks that requires a fraction of BERT’s computation cost and time by learning from the dormant hidden state activations within.

BERTVision Models **91%**

BERT-Large Models **71%**

* A fictionalized representation of our goal: to beat BERT-large across a spectrum of NLP tasks at a fraction of the size and cost.

Metrics for Success

BERT v BERTVision

We will evaluate our models with BERT-large for the NLP tasks

Think Small

Our success depends on shrinking the footprint of BERT

Reduced Time

Time to fine-tune and get results is a critical part of our value

Academic Contribution

Contribute to the NLP research community? Check!

THANKS

Questions?