Using BERT for NLU in production

What is BERT

Bidirectional **E**ncoder **R**epresentations from **T**ransformers

- Novel language representation model

- Fine-tuned BERT models achieved state-of-the-art results on multiple NLP tasks: SQUAD 2.0 question answering, GLUE multi task learning, Google natural questions, etc.:

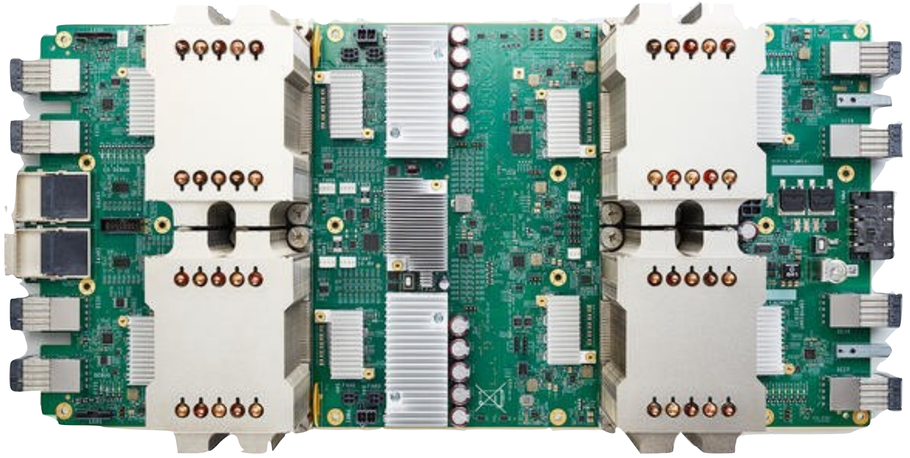
System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

GLUE Test results, scored by the evaluation server (https://gluebenchmark.com/leaderboard)

NLP PLATFORM

Pre-training BERT

- **Unsupervised**, performed on raw text data
- **Very compute-intensive**, takes days on TPUv2, months on GPU

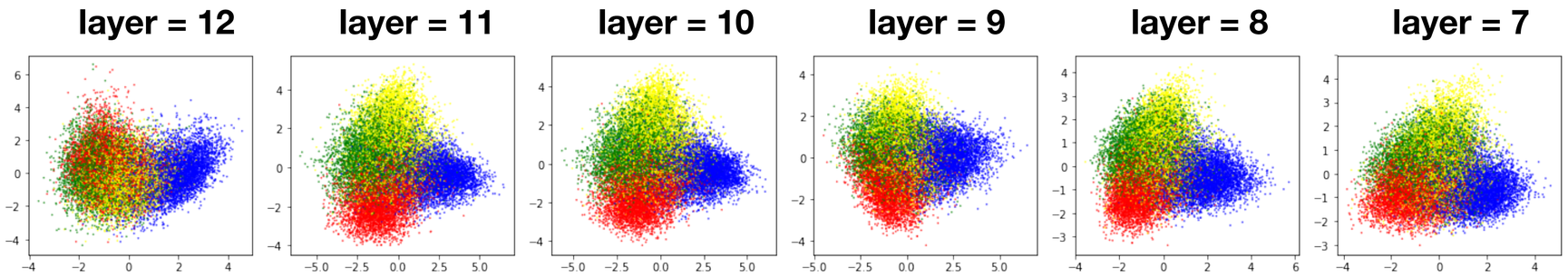


- Pre-trained models are available in open-source
- Guide to pre-training with cloud TPU is available on Medium

NLP PLATFORM

Text representation

- Even without fine-tuning, a pre-trained BERT model can be used to generate text representations (embeddings)
- These representations can be used for token-level or sentence-level tasks
- Encoder layer output shape == [encoder_dim, seq_len]
- Sentence embeddings are obtained by pooling encoder layer output



Mean-pooled representations for samples from UCI-news-aggregator dataset (4 classes)

NLP PLATFORM

Visualising embeddings

Even without fine-tuning, BERT can be used for efficient text representation

NLP PLATFORM

Fine-tuning with BERT

☆ **End-to-end**
The whole model is fine-tuned on a new task

- Superior results on large datasets
- Need **k** BERT models to solve **k** different tasks

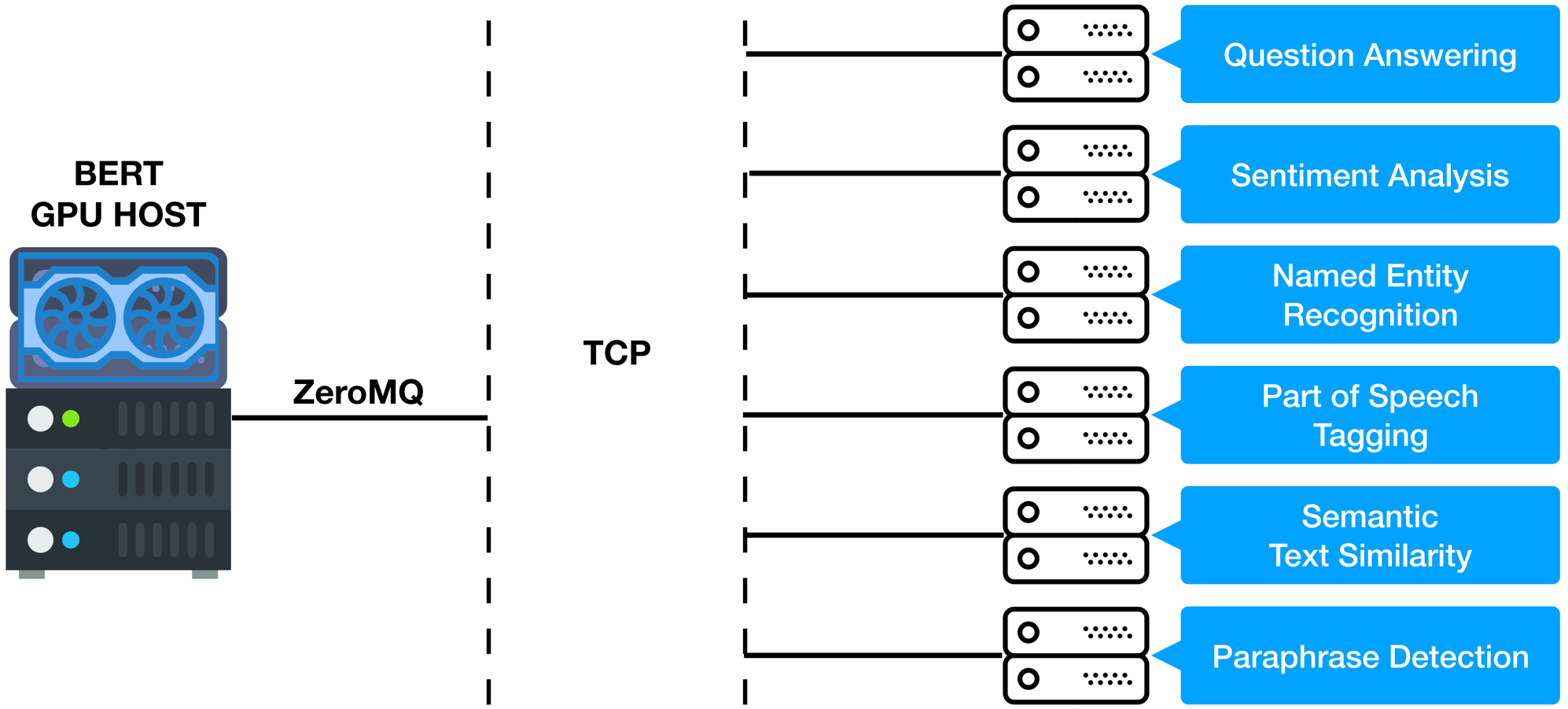
☆ **Frozen**
BERT model is fixed (frozen), encoder output used as features

- Good results with smaller datasets
- Need a single BERT + **k** lightweight models to solve **k** different tasks

NLP PLATFORM

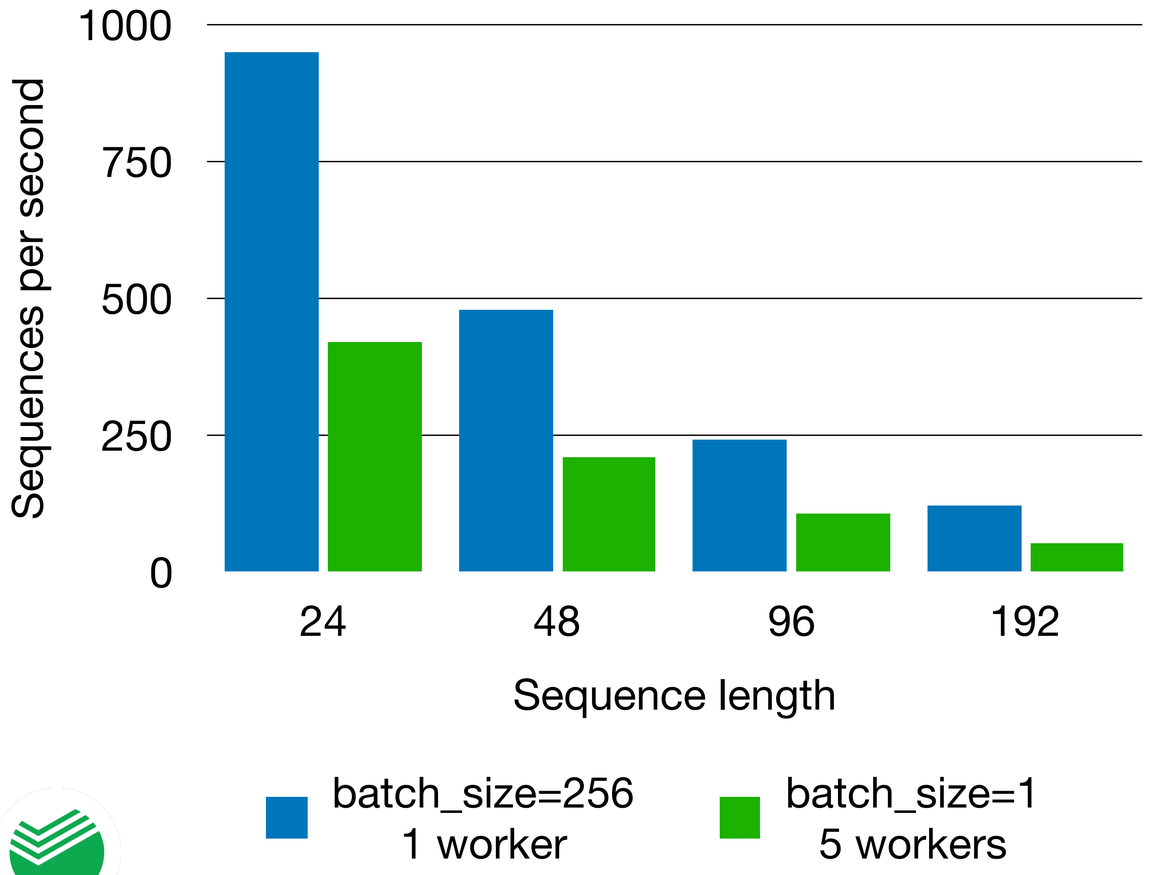
Serving architecture

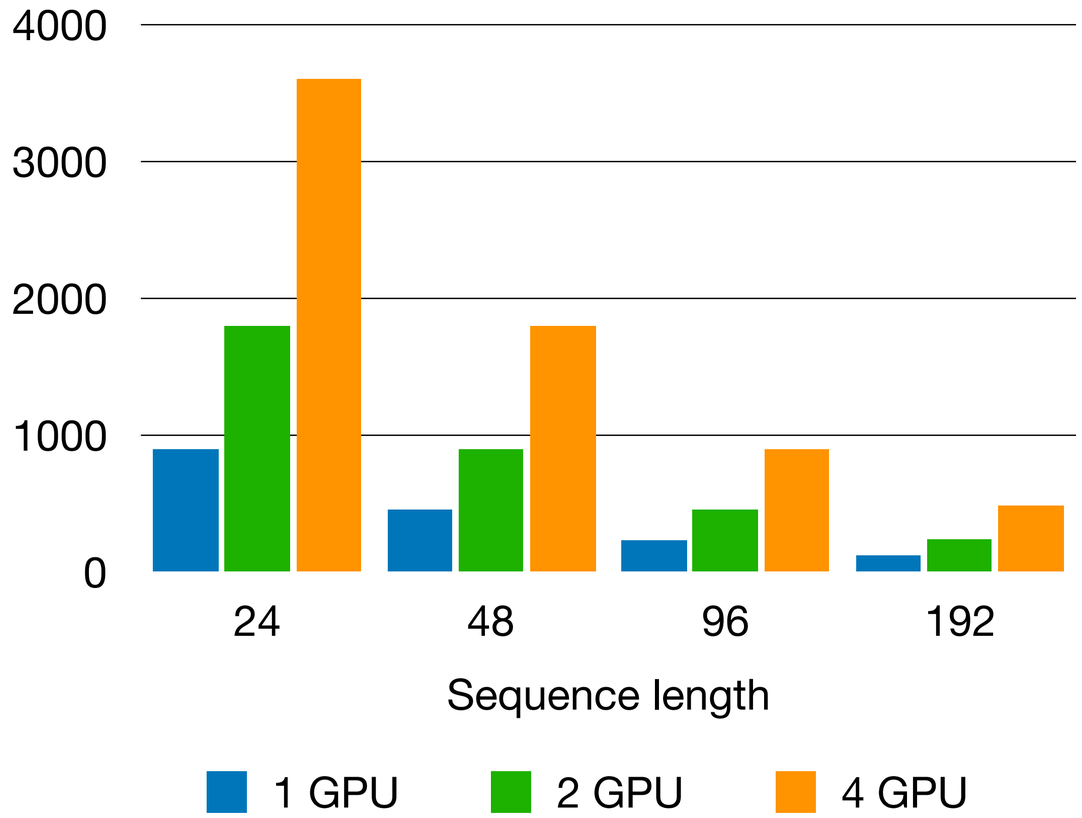
A single model can serve multiple clients solving a wide array of NLP tasks



NLP PLATFORM

BERT-base performance on Tesla V100





NLP PLATFORM

Service reliability

- Client-server infrastructure introduces additional failure modes
- A standalone solution is required for mission-critical systems
- Local model only requires TensorFlow
- Inference is performed in-process with tf.Estimator
- Model graph and weights are serialized and distributed in **Protobuf**

NLP PLATFORM

Info

- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
- Pre-training BERT from scratch with cloud TPU
- bert_experimental
- bert-as-a-service
- Embedding Projector
- Serving Google-BERT in Production using Tensorflow and ZeroMQ/

Denis Antyukhov
NLP Solutions, Sberbank

dantyukhov@icloud.com
Telegram: @aphex34