Cloud Infrastructure for Portable AI

Nathan Dolbir, Kennedy Fairey, Anirudh Oruganti

Problem Statement

- Portable devices don’t have hardware capabilities to host many ML systems
  - Big tech has solved this problem, but developers with limited resources rarely attempt
- Solution: Build an end-to-end cloud computing infrastructure for low powered, portable devices
  - Multiple Transformer models to prove robustness and flexibility
  - Sentiment classifier that is structured like a chatbot app

Technical Challenges

- Projects of this nature usually have many more resources
  - People
  - Money
  - Technology
- Cloud computing for ML is not a widely studied area for ameteur developers
- There are technologies that would make this project much easier, but they cost a lot of money
- Not as much documentation for this kind of project as there is for other areas

Related Work

- Exploring Transformers in Emotion Recognition: a comparison of BERT, DistillBERT, RoBERTa, XLNet and ELECTRA Diogo Cortiz Brazilian Network Information Center (NIC.br) Pontifical Catholic University of São Paulo (PUC-SP)
- • Learning rate: 5e-5
- • Batch Size: 16
- • Epochs: 4
- • Threshold: 0.30

Time to complete	Training	Evaluation
BERT	02:40:00	00:00:40
DistilBERT	00:33:58	00:00:09
RoBERTa	01:08:56	00:00:42
XLNet	01:33:49	00:01:06
ELECTRA	00:13:04	00:00:07

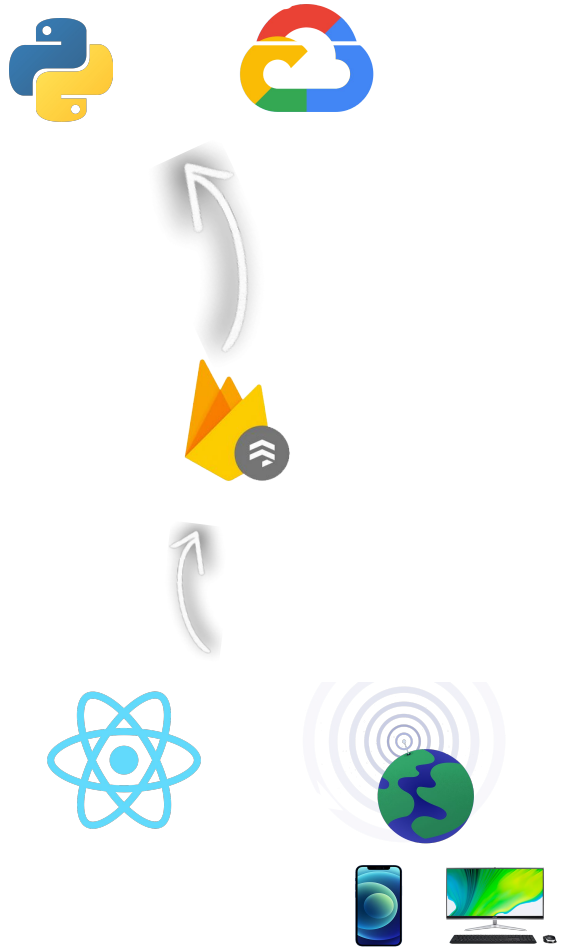
Table 5: Time to Complete for GoEmotion

Emotion	BERT	Distil BERT	RoBERTa	XLNet	Electra
admiration	0.65	0.71	0.73	0.73	0.71
amusement	0.80	0.79	0.79	0.78	0.79
anger	0.47	0.49	0.48	0.51	0.47
annoyance	0.34	0.40	0.38	0.37	0.29
approval	0.36	0.37	0.38	0.38	0.31
caring	0.39	0.43	0.47	0.48	0.00
confusion	0.37	0.43	0.43	0.44	0.34
curiosity	0.54	0.55	0.55	0.57	0.53
desire	0.49	0.53	0.58	0.56	0.00
disappointment	0.28	0.24	0.35	0.32	0.00
disapproval	0.39	0.39	0.43	0.41	0.35
disgust	0.45	0.47	0.49	0.47	0.00
embarrassment	0.43	0.54	0.57	0.55	0.00
excitement	0.34	0.33	0.34	0.32	0.00
fear	0.60	0.62	0.68	0.68	0.37
gratitude	0.86	0.90	0.90	0.90	0.90
grief	0.00	0.00	0.00	0.00	0.00
joy	0.51	0.57	0.57	0.56	0.53
love	0.78	0.77	0.79	0.78	0.78
nervousness	0.35	0.34	0.34	0.35	0.00
optimism	0.51	0.59	0.59	0.58	0.54
pride	0.36	0.22	0.00	0.00	0.00
realization	0.21	0.28	0.26	0.28	0.00
relief	0.15	0.00	0.00	0.00	0.00
remorse	0.66	0.71	0.70	0.77	0.63
sadness	0.49	0.55	0.55	0.53	0.48
surprise	0.50	0.53	0.58	0.56	0.47
neutral	0.68	0.66	0.66	0.65	0.68
macro avg	0.46	0.48	0.49	0.48	0.33
std	0.19	0.21	0.23	0.23	0.30

Table 4: Results of different models for GoEmotion Taxonomy

Approach: Implemented Infrastructure

- Use Google Cloud and Firebase for easy integration
- Google Cloud Triggers cause adjustments on UI and input for Cloud models



Approach: Models and Metrics

- Five Transformer models used
  - BERT
  - DistillBERT
  - RoBERTa
  - ELECTRA
  - MobileBERT
- Metrics used
  - Accuracy
  - F1 Macro score
  - Time to train
  - Evaluation

Model Parameters

- **Loss: Sparse Categorical Cross Entropy**
- **Optimizer: Adam**
- **Learning rate: 5e-5**
- **Batch Size: 16**
- **Epochs: 4**

F1 Macro Score= 2 * (Precision * Recall / Precision + Recall)

Results (Infrastructure Implementation)

- The app works and returns a response
- Limited funds = no constant server
  - Every time the cloud function runs, a new instance is created
    - This takes about 60 seconds
- Calculating response time was unnecessary because of this

Results (Transformer Models)

Emotion	Bert	DistilBert	Electra	Roberta	MobileBert
anger	0.894632	0.927007	0.925714	0.911972	0.906526
fear	0.887417	0.868778	0.887931	0.871671	0.908686
joy	0.944649	0.944803	0.943369	0.947589	0.950181
love	0.824798	0.795987	0.812698	0.804428	0.849398
sadness	0.962901	0.961571	0.962712	0.964041	0.963668
surprise	0.727273	0.744828	0.743802	0.751678	0.702703
accuracy	0.919	0.9205	0.924	0.9225	0.928
macro avg	0.873612	0.873829	0.879371	0.87523	0.880193
weighted avg	0.919962	0.920282	0.923378	0.921122	0.92727

Figure 7. Above are the f1-macro avg from [4] D. Cortiz results after fine tuning GoEmotion data.

Models	Training(min)	Evalution(min)	Evalution(msg/s)
Bert	**39.898933**	**3.108629**	**0.093259**
DistilBert	**20.787233**	**2.53336**	**0.076001**
Roberta	**40.637583**	**3.122908**	**0.08981**
Electra	**12.080952**	**2.993671**	**0.093687**
MobileBert	**30.145858**	**7.588735**	**0.227662**

Figure 8. Above are the elapsed time results for training and evaluation.

	Accuracy	Val Accuracy	Loss	Val Loss
DistillBERT	0.9265	0.9340	0.2069	0.1952
RoBERTa	0.9048	0.9300	0.2680	0.1903
BERT	0.9151	0.9364	0.2399	0.1547
ELECTRA	0.8887	0.9234	0.3096	0.2286
MobileBERT	0.8410	0.8731	0.4623	0.3513

Figure 5. Above are the metrics used to measure the performance of each of the five Transformer models. "Val" stands for validation and each value is the average over four epochs.

Conclusion and Broader Impact

- Proved the infrastructure can be refit for any ML system with custom models and UI with a NoSQL database
- Can be reproduced for research and industrial purposes
- High powered ML can be developed by anyone with a zero budget
- Intend to use and build upon this system going forward for future projects
- May publish work on how to use this infrastructure, reformatting final paper as full-fledged research paper