The Eight International Conference on Green Communications,
Computing and Technologies - GREEN23
Porto, Portugal, September 25-29, 2023

Comparable Machine Learning Efficiency: Balanced Metrics for Natural Language Processing

Daniel Schönle, Christoph Reich, Djaffar Ould Abdeslam

Presenter: Daniel Schönle (daniel.schoenle@hs-furtwangen.de)
Furtwangen University, Institute IDACUS
Université de Haute Alsace, IUT de Mulhouse, Institut IRIMAS

Daniel Schönle

daniel.schoenle@hs-furtwangen.de

Research Topics

- Blockchain, Cyber-physical Systems
- Machine Learning in e-Learning Environments
- Natural Language Processing, Feature Selection
- ML-Efficiency

Publications

- Linguistic Driven Feature Selection for Text Classification as Stop Word Replacement, (2023)
- Data-Driven Tutoring: challenges and prospects, (2021)
- Industry use cases on blockchain technology, (2021)
- Digital twin as a service (2021)

Machine Learning Efficiency Metric Motivation

Situation

- Decades ago: **Focus on computational efficiency** due to limited computing power.
- Shift in focus: Powerful computing resources, ML **emphasis on prediction quality**.
- Emergence of Large Language Models (LLMs) and increase of ML-Model sizes

Challenges

- Growing demand for efficiency alongside ML application expansion.
- Optimisation of resource-intensive solutions for sustainability.
- Need for economically and environmentally viable approaches.

- Goal: Improve efficiency in ML.
- Research Objective: Introduction of robust **metrics** for measuring ML model **efficiency**.

Machine Learning Efficiency Metric: Objectives

- Introduction of novel metrics to **measure ML procedure efficiency**.
- Metrics incorporate resource consumption, computational effort, and runtime considerations.
- Providing holistic perspective on true efficiency of ML procedure e.g. Training or Usage.

- Enable applicability to **all ML techniques**.
- Provide measurements for **common host setups**.
- **Balance efficiency effects** across diverse host setups.

Machine Learning Efficiency Metric: Efficiency & Dimensions

Solution Efficiency

- Trade-off between solution achievement and solution cost.
- Cost include solution parameters like efforts and resources consumed.
- Achievement is an optimisation of specific solution parameters (focus)

Efficiency Dimensions

Solution parameters can be seen as efficiency dimensions:

- Accuracy or Performance -> Quality q
- Computational Effort -> Work w
- Absolute Resource Utilisation -> Space s
- Relative Resource Utilisation -> Load l
- Training Duration -> Duration d

Compact Metric (CO)

Efficiency

- is a trade-off
- between **focused and unfocused** efficiency dimensions
- Where every dimension has **weights** to adapt their importance

Definition

- Compact Metric (CO) on Focus F for ML-Procedure M

$$[F]CO(M) = (F \times \alpha) \times U$$
$$[F]CO(M) = (q_M \times \beta_q) \times (w_M \times \beta_w) \times (s_M \times \beta_s) \times (l_M \times \beta_l) \times (d_M \times \beta_d) \times \psi$$

$$\text{where } D = \{r \in \mathbb{R} \,|\, r > 1\} \; and \; \{q, w, s, l, d \in D\}$$
$$F \subseteq D \quad \text{and} \quad U = D \backslash F$$
$$\psi = \text{Score-Compensation}$$

Quality focused Compact Metric (QCO): Definition

Quality focus

- Measures computational effort for prediction quality.
- Efficiency increases with less work, less resource consumption, less computational effort, less time, and higher prediction quality.

Quality focused Compact Metric (QCO)

- Quality focused efficiency of ML-Procedure = quality per work, space and time

$$QCO(M) = \frac{q^{\alpha * \beta_q}}{(w^{\beta_w} + s^{\beta_s} + d^{\beta_d})} * \psi$$

Quality focused Compact Metric (QCO): Instantiation

Instantiation Procedure

- Select appropriate Measurements
- Apply Data-Transformations
- Define Weight-Values

Results:

where $F1 =$ F1-Score, $BACC =$ Balanced Accuracy Score,
$FLOPS =$ Floating Point Ops., $MPF =$ Minor Page Faults,
$DS =$ Dataset-Size, $RSS =$ Resident Set Size,
$D =$ Duration, $TOC =$ Time on CPU

$$QCO_F(M) = \frac{((\frac{F1+BACC}{2})^6)}{(\log_{63P} FLOPS/DS[kB] + \log_{128G} RSS[MB] * log_{864M} D[s] + \log_{172T} TOC[ns]} * 10 \quad (4)$$

$$QCO_F(M) = \frac{((\frac{F1+BACC}{2})^6)}{(\log_{800M} MPF/DS[kB] + \log_{128G} RSS[MB] * log_{864M} D[s] + \log_{172T} TOC[ns]} * 10$$

Quality focused Compact Metric (QCO): Evaluation

Procedure

- Measure ML Procedures for each Config
- Expert Ranking of Results
- Rank Results by QCO-Score

Measurement X

ML Procedure Config X

Value
Datasetsize
RSS
FLOPs
Minor Page Faults
Time on CPU
Duration
Accuracy
F1

Datasets / Tasks

- SMS Spam Detection [1]
- IMDB Movie Reviews [2]

Vectorizer

- Word Frequency -> TFIDF
- Word Embedding -> Bert
- Finetuned Transformer -> DistilBERT

Classifier

- Shallow CLFs
- Finetuned Transformer -> DistilBERT

Host

- AMD Ryzen 7 5800U 32GB RAM

[1] Almeida et. al, 2011. Contributions to the study of sms spam filtering: New collection and results,
[2] A. L. Maas et. al, 2011.“Learning word vectors for sentiment analysis”

Evaluation: Ranking Comparison

- Expert Ranking almost fulfilled; Experts Rank Transformer higher

DAT	VECT	CLF	DUR	*EV Metric*					*QCO* Metrics		Rankings				
				QUA	$TIME$	SPA	WO_P	WO_F	QCO_P	QCO_F	EXP_1	QCO_{1P}	EXP_2	QCO_{2P}	QCO_{2F}
SMS	TFIDF	NB	00:00:34	0,968	0,407	0,260	1,043	0,691	1,260	1,726	1	1	1	1	1
SMS	TFIDF	GD	00:02:36	0,972	0,583	0,371	1,043	0,631	1,190	1,678	2	2	2	2	2
IMDB	TFIDF	GD	00:00:19	0,885	0,339	0,222	0,616	0,610	1,145	1,153	3	3	3	3	3
SMS	DIST-T	DIST-T	00:01:34	0,982	0,525	0,384	1,249		1,099		6	4			
SMS	BERT	SVM	00:11:29	0,972	0,755	0,510	1,143	1,465	1,018	0,852	4	5	5	4	4
IMDB	DIST-T	DIST-T	02:38:32	0,982	1,058	0,795	1,058		0,970		5	6			
SMS	BERT	GD	02:04:00	0,978	1,030	0,696	1,140	1,637	0,956	0,752	8	7	4	5	6
IMDB	DISTIL	DISTIL	11:31:52	0,978	1,228	0,923	1,136		0,848		7	8			
IMDB	TFIDF	NB	00:01:18	0,845	0,504	0,330	0,618	0,581	0,766	0,797	9	9	6	6	5
SMS	BERT	NB	01:58:30	0,932	1,024	0,692	1,141	1,638	0,715	0,563	11	10	8	7	8
SMS	DISTIL	DISTIL	01:39:58	0,983	1,005	0,750	1,845		0,697		12	11			
SMS	BERT	RF	01:09:50	0,916	0,963	0,651	1,140	1,639	0,660	0,516	10	12	7	8	9
IMDB	TFIDF	RF	00:00:58	0,809	0,469	0,309	0,617	0,646	0,604	0,586	15	13	9	9	7
SMS	TFIDF	RF	00:03:02	0,787	0,601	0,376	1,043	0,931	0,335	0,363	16	14	12	10	10
IMDB	TFIDF	SVM	00:20:20	0,714	0,821	0,554	0,638	0,812	0,223	0,194	13	15	11	11	11
SMS	TFIDF	SVM	00:00:35	0,652	0,409	0,264	1,048	1,092	0,117	0,113	14	16	10	12	12

Cloumns: Dataset, Vectorizer, Classifier, Duration, Quality, Time, Space, WO_F= Work (FLOPS), WO_P = Work (Minor Page Faults), QCO Metrics, Rankings by Domain EXPerts, or QCO, SMS = SMS Spam Dataset[32], IDB = IMDB Dataset[33], BERT = BERT word embedding, DIST-T = finetuned DistilBERT word embedding (PyTorch) & classification, DISTIL = finetuned DistilBERT word embedding (TensorFlow + keras) & classification, GD = Gradient Descent, SVM = Support Vector Machine, NB = Naïve Bayes

Evaluation: Best / Worst Duration & Quality

- Configurations per DAT/VECT: Fastest NB+GD; Best Quality: Transfomers
- Efficiency?

DAT	VECT	CLF	DUR	*EV Metric*					*QCO* Metrics		Rankings				
				QUA	$TIME$	SPA	WO_P	WO_F	QCO_P	QCO_F	EXP_1	QCO_{1P}	EXP_2	QCO_{2P}	QCO_{2F}
SMS	TFIDF	NB	00:00:34	0,968	0,407	0,260	1,043	0,691	1,260	1,726	1	1	1	1	1
SMS	TFIDF	GD	00:02:36	0,972	0,583	0,371	1,043	0,631	1,190	1,678	2	2	2	2	2
IMDB	TFIDF	GD	00:00:19	0,885	0,339	0,222	0,616	0,610	1,145	1,153	3	3	3	3	3
SMS	DIST-T	DIST-T	00:01:34	0,982	0,525	0,384	1,249		1,099		6	4			
SMS	BERT	SVM	00:11:29	0,972	0,755	0,510	1,143	1,465	1,018	0,852	4	5	5	4	4
IMDB	DIST-T	DIST-T	02:38:32	0,982	1,058	0,795	1,058		0,970		5	6			
SMS	BERT	GD	02:04:00	0,978	1,030	0,696	1,140	1,637	0,956	0,752	8	7	4	5	6
IMDB	DISTIL	DISTIL	11:31:52	0,978	1,228	0,923	1,136		0,848		7	8			
IMDB	TFIDF	NB	00:01:18	0,845	0,504	0,330	0,618	0,581	0,766	0,797	9	9	6	6	5
SMS	BERT	NB	01:58:30	0,932	1,024	0,692	1,141	1,638	0,715	0,563	11	10	8	7	8
SMS	DISTIL	DISTIL	01:39:58	0,983	1,005	0,750	1,845		0,697		12	11			
SMS	BERT	RF	01:09:50	0,916	0,963	0,651	1,140	1,639	0,660	0,516	10	12	7	8	9
IMDB	TFIDF	RF	00:00:58	0,809	0,469	0,309	0,617	0,646	0,604	0,586	15	13	9	9	7
SMS	TFIDF	RF	00:03:02	0,787	0,601	0,376	1,043	0,931	0,335	0,363	16	14	12	10	10
IMDB	TFIDF	SVM	00:20:20	0,714	0,821	0,554	0,638	0,812	0,223	0,194	13	15	11	11	11
SMS	TFIDF	SVM	00:00:35	0,652	0,409	0,264	1,048	1,092	0,117	0,113	14	16	10	12	12

Cloumns: Dataset, Vectorizer, Classifier, Duration, Quality, Time, Space, WO_F= Work (FLOPS), WO_P = Work (Minor Page Faults), QCO Metrics, Rankings by Domain EXPerts, or QCO, SMS = SMS Spam Dataset[32], IDB = IMDB Dataset[33], BERT = BERT word embedding, DIST-T = finetuned DistilBERT word embedding (PyTorch) & classification, DISTIL = finetuned DistilBERT word embedding (TensorFlow + keras) & classification, GD = Gradient Descent, SVM = Support Vector Machine, NB = Naïve Bayes

Evaluation: Classifiers

- Efficiency: Best Shallow: TFIDF NB+GD

DAT	VECT	CLF	DUR	*EV Metric*					*QCO* Metrics		Rankings				
				QUA	$TIME$	SPA	WO_P	WO_F	QCO_P	QCO_F	EXP_1	QCO_{1P}	EXP_2	QCO_{2P}	QCO_{2F}
SMS	TFIDF	NB	00:00:34	0,968	0,407	0,260	1,043	0,691	1,260	1,726	1	1	1	1	1
SMS	TFIDF	GD	00:02:36	0,972	0,583	0,371	1,043	0,631	1,190	1,678	2	2	2	2	2
IMDB	TFIDF	GD	00:00:19	0,885	0,339	0,222	0,616	0,610	1,145	1,153	3	3	3	3	3
SMS	DIST-T	DIST-T	00:01:34	0,982	0,525	0,384	1,249		1,099		6	4			
SMS	BERT	SVM	00:11:29	0,972	0,755	0,510	1,143	1,465	1,018	0,852	4	5	5	4	4
IMDB	DIST-T	DIST-T	02:38:32	0,982	1,058	0,795	1,058		0,970		5	6			
SMS	BERT	GD	02:04:00	0,978	1,030	0,696	1,140	1,637	0,956	0,752	8	7	4	5	6
IMDB	DISTIL	DISTIL	11:31:52	0,978	1,228	0,923	1,136		0,848		7	8			
IMDB	TFIDF	NB	00:01:18	0,845	0,504	0,330	0,618	0,581	0,766	0,797	9	9	6	6	5
SMS	BERT	NB	01:58:30	0,932	1,024	0,692	1,141	1,638	0,715	0,563	11	10	8	7	8
SMS	DISTIL	DISTIL	01:39:58	0,983	1,005	0,750	1,845		0,697		12	11			
SMS	BERT	RF	01:09:50	0,916	0,963	0,651	1,140	1,639	0,660	0,516	10	12	7	8	9
IMDB	TFIDF	RF	00:00:58	0,809	0,469	0,309	0,617	0,646	0,604	0,586	15	13	9	9	7
SMS	TFIDF	RF	00:03:02	0,787	0,601	0,376	1,043	0,931	0,335	0,363	16	14	12	10	10
IMDB	TFIDF	SVM	00:20:20	0,714	0,821	0,554	0,638	0,812	0,223	0,194	13	15	11	11	11
SMS	TFIDF	SVM	00:00:35	0,652	0,409	0,264	1,048	1,092	0,117	0,113	14	16	10	12	12

Cloumns: Dataset, Vectorizer, Classifier, Duration, Quality, Time, Space, WO_F= Work (FLOPS), WO_P = Work (Minor Page Faults), QCO Metrics, Rankings by Domain EXPerts, or QCO, SMS = SMS Spam Dataset[32], IDB = IMDB Dataset[33], BERT = BERT word embedding, DIST-T = finetuned DistilBERT word embedding (PyTorch) & classification, DISTIL = finetuned DistilBERT word embedding (TensorFlow + keras) & classification, GD = Gradient Descent, SVM = Support Vector Machine, NB = Naïve Bayes

Use Case: Hyperparameter optimization

Experiment

- Task: Search most efficient maximum sequence length for DistilBERT
- Dataset: SMS spam detection [1]

Results

- SL512 more efficient than SL256, although same quality and higher duration

	Measurements		Dimensions				Scores		Hidden Info	
SL	Duration	F1	Q	W	S	T	QCO_F	QCO_P	BACC	F1
128	09:08:50	0,76	0,64	3,02	0,45	0,78	0,159	0,164	0,502	0,775
256	00:27:02	0,78	0,64	2,86	0,45	0,71	0,171	0,172	0,502	0,779
512	00:50:13	0,78	0,65	2,94	0,47	0,72	0,183	0,174	0,517	0,784

Text Classification Efficiency with DistilBERT with different maximum Sequence Length (SL). Smoothed Dimensions: Quality, Work, Space and Time. Efficiency Scores Quality Focused based on FLOPS (QCO_F) and Minor Page Faults (QCO_P)

[1] Almeida et. Al, 2011, Contributions to the study of sms spam filtering: New collection and results,

Thank you.

Daniel Schönle

daniel.schoenle@hs-furtwangen.de