XTREME

(X) Cross-Lingual Transfer Evaluation of Multilingual Encoders

A comprehensive benchmark for cross-lingual transfer learning on a diverse set of languages and tasks.

Overview

Recent progress in applications of machine learning models to NLP has been driven by benchmarks that evaluate models across a wide variety of tasks. However, these broad-coverage benchmarks have been mostly limited to English, and despite an increasing interest in multilingual models, a benchmark that enables the comprehensive evaluation of such methods on a diverse range of languages and tasks is still missing.

To encourage more research on multilingual transfer learning, we introduce the Cross-lingual TRansfer Evaluation of Multilingual Encoders (XTREME) benchmark. XTREME covers 40 typologically diverse languages spanning 12 language families and includes 9 tasks that require reasoning about different levels of syntax or semantics.

The languages in XTREME are selected to maximize language diversity, coverage in existing tasks, and availability of training data. The languages in XTREME are selected to maximize language diversity, coverage in existing tasks, and availability of training data. Among these are many under-studied languages, such as the Dravidian languages Tamil (spoken in southern India, Sri Lanka, and Singapore), Telugu and Malayalam (spoken mainly in southern India), and the Niger-Congo languages Swahili and Yoruba, spoken in Africa.

For a full description of the benchmark, languages and tasks, please see XTREME: A Massively Multilingual Multi-task Benchmark for Evaluating Cross-lingual Generalization.

Leaderboard results

Rank	Model	Participant	Affiliation	Attempt Date	Avg	Sentence-pair Classification	Structured Prediction	Question Answering	Sentence Retrieval
0		Human	-	-	93.3	95.1	97.0	87.8	-
1	CoFe	HFL	iFLYTEK	Oct 26, 2021	84.1	90.1	81.4	75.0	94.2
2	Turing ULR v5	Alexander v-team	Microsoft	Sep 17, 2021	83.7	90.0	81.4	74.3	93.7
3	InfoXLM-XFT	Noah's Ark Lab	Huawei	Oct 5, 2021	82.2	89.3	75.5	75.2	92.4
4	VECO + HICTL	AliceMind + MT	Alibaba	Sep 21, 2021	82.0	89.0	76.7	73.4	93.3
5	Polyglot	MLNLC	ByteDance	Apr 29, 2021	81.7	88.3	80.6	71.9	90.8
6	Unicoder + ZCode	MSRA + Cognition	Microsoft	Apr 26, 2021	81.6	88.4	76.2	72.5	93.7
7	ERNIE-M	ERNIE Team	Baidu	Jan 1, 2021	80.9	87.9	75.6	72.3	91.9
8	HiCTL	DAMO MT Team	Alibaba	Mar 21, 2021	80.8	89.0	74.4	71.9	92.6
9	T-ULRv2 + StableTune	Turing	Microsoft	Oct 7, 2020	80.7	88.8	75.4	72.9	89.3
10	Anonymous3	Anonymous3	Anonymous3	Jan 3, 2021	79.9	88.2	74.6	71.7	89.0

Participate in Competition

Task and Language Details

The tasks included in XTREME cover a range of paradigms, including sentencetext classification, structured prediction, sentence retrieval and cross-lingual question answering. Consequently, in order for models to be successful on the XTREME benchmarks, they must learn representations that generalize to many standard cross-lingual transfer settings.

Each of the tasks covers a subset of the 40 languages. In order to obtain additional data in the low-resource languages that can be used for analyses, we automatically translate test sets of a natural language inference and question answering dataset to the remaining languages. We show that these can be used as a reasonable proxy for performance on gold standard test sets, with the caveat that they overestimate the performance of models that were trained on translations themselves.

Family	Languages
Afro-Asiatic	Arabic, Hebrew
Austro-Asiatic	Vietnamese
Austronesian	Indonesian, Javanese, Malay, Tagalog
Basque	Basque
Dravidian	Malayalam, Tamil, Telugu
Indo-European (Indo-Aryan)	Bengali, Marathi, Hindi, Urdu
Indo-European (Germanic)	Afrikaans, Dutch, English, German
Indo-European (Romance)	French, Italian, Portuguese, Spanish
Indo-European (Greek)	Greek
Indo-European (Iranian)	Persian
Japonic	Japanese