LANGUAGE MODEL KNOWLEDGE DISTILLATION FOR EFFICIENT QUESTION ANSWERING IN SPANISH

Adrián Bazaga, Pietro Liò & Gos Micklem
University of Cambridge
{ar989,pl219,gm263}@cam.ac.uk

ABSTRACT

Recent advances in the development of pre-trained Spanish language models has led to significant progress in many Natural Language Processing (NLP) tasks, such as question answering. However, the lack of efficient models imposes a barrier for the adoption of such models in resource-constrained environments. Therefore, smaller distilled models for the Spanish language could be proven to be highly scalable and facilitate their further adoption on a variety of tasks and scenarios. In this work, we take one step in this direction by developing SpanishTinyRoBERTa, a compressed language model based on RoBERTa for efficient question answering in Spanish. To achieve this, we employ knowledge distillation from a large model onto a lighter model that allows for a wider implementation, even in areas with limited computational resources, whilst attaining negligible performance sacrifice. Our experiments show that the dense distilled model can still preserve the performance of its larger counterpart, while significantly increasing inference speedup. This work serves as a starting point for further research and investigation of model compression efforts for Spanish language models across various NLP tasks.

1 INTRODUCTION

In the last years, recent advancements in the field of NLP have panned the way for progress on a variety of downstream tasks, primarily by fine-tuning pre-trained language models (PLMs) (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019) on large-scale task-specific datasets. These models are typically trained on large and high-quality annotated corpora, which are are usually scarce for languages other than English, posing a significant disadvantage for progressing multilingual NLP research. Therefore, resources for languages such as Spanish, the fourth most spoken language, remain underrepresented in terms of amount of training data available. Despite the availability of some Spanish language models (Gutiérrez-Fandiño et al., 2022; Cañete et al., 2020; de la Rosa et al., 2022; Pérez et al., 2022), their use in resource-limited environments remains challenging. To address this, we propose using knowledge distillation (Jiao et al., 2020; Boreshban et al., 2021) to condense a large Spanish RoBERTa model into a smaller, more efficient model, SpanishTinyRoBERTa, without significant loss of performance. Focusing on the question answering task (Rajpurkar et al., 2016; Carrino et al., 2019), our experiments demonstrate that SpanishTinyRoBERTa maintains comparable accuracy to the original model while significantly improving inference speed and reducing resource usage. Our work aims to facilitate the development of efficient Spanish language models, lowering computational barriers and encouraging the adoption of NLP technologies.

2 METHODOLOGY

We propose developing a Spanish TinyRoBERTa model using the knowledge distilled from a Spanish RoBERTa-large model in the SQuAD-es QA task (Carrino et al., 2019). In the context of language models, knowledge distillation can be modeled as penalizing the difference of feature representations between the teacher and the model, therefore aiming to minimize the following objective function:

$$\mathcal{L}_{\text{KD}} = \sum_{x\in\mathcal{M}} L\big(f^S(x), f^T(x)\big), \tag{1}$$

where $L(\cdot)$ is a loss function that computes the discrepancy between teacher and student models, x is the text input and $\mathcal{M}$ denotes the training dataset. In this work, we employ the knowledge distillation technique introduced in TinyBERT (Jiao et al., 2020) to reproduce the behavior of the larger model and leveraging the knowledge transferred from it. During model training, as the student model (SpanishTinyRoBERTa) contains much less layers than the teacher, we map layers in the student to the teacher by using a layer mapping function, so that the student learns from intermittent layers of the teacher model (more details in Appendix A.1).

3 Experimental Setup and Results

To demonstrate the effectiveness of our distillation for Spanish question answering (QA) tasks, we use the SQuAD-es dataset (Carrino et al., 2019), which is a Spanish version of SQuAD (Rajpurkar et al., 2016). To address questions answering as a learning problem, we treat the QA task as the problem of sequence labeling which predicts the possibility of each token as the start or end of answer span. We instantiate a tiny RoBERTa student model, with number of layers M=6, the hidden size as d_h=512, the feedforward size as d_{ff}=3072 and the head number as h=16), account for a total of 51.4M parameters. If not specified, this student model is referred to as the SpanishTinyRoBERTa. We use as teacher model a Spanish RoBERTa-large model pre-trained on a corpus from the National Library of Spain (Gutiérrez-Fandiño et al., 2022). The teacher model has layers N=12, a hidden size d_h=1024, feed-forward size d_{ff}=4096 and h=16 heads, accounting for a total of 355M parameters. We use four different models as baselines: the Multilingual BERT (mBERT), a Spanish BERT-base (Cañete et al., 2020), a Spanish RoBERTa-base (124M parameters) and a Spanish RoBERTa-large as teacher model. As the main focus of this work is to improve running time efficiency, we use these baseline models to measure both the inference speed and performance when compare with our lighter model. More details on the training hyperparameters in Appendix A.2.

Model	F1 (%)	EM (%)	Inference Speedup
Spanish RoBERTa-large (teacher)	87.50	78.30	1.0x
Multilingual BERT (mBERT)	77.60	61.80	3.0x
Spanish BERT-base	82.15	73.59	2.4x
Spanish RoBERTa-base	81.80	72.30	2.5x
SpanishTinyRoBERTa (ours)	80.52	71.23	4.2x

Table 1: Comparison of the different models on the SQuAD-es Spanish Question Answering task using the F1 and Exact Match (EM) metrics.

In terms of results, Table 1 shows the performance of the baseline models and our distilled model in the SQuAD-es dataset. The SpanishTinyRoBERTa achieved 80.51% and 71.23% for F1 score and Exact Match (EM), respectively, comparable to the performance attained by the teacher model, 87.50% and 78.30%, for F1 score and EM, respectively. These findings provide evidence that the SpanishTinyRoBERTa can achieve competitive results on the QA task while requiring much less computational resources. On this regard, we can observe in Appendix A.3 that SpanishTinyRoBERTa contains 6.9x less parameters (51M) than the teacher model (355M), and achieves 4.2x inference speedup (392ms vs 1683ms per query). This shows the benefit of using the distillation process to achieve a much lighter, faster and highly performant model.

4 Conclusion

In this work we present SpanishTinyRoBERTa, a compressed language model for efficient question answering in Spanish that meets similar performance results to its larger counterpart, with a significant reduction in terms of required computational resources. We showed that the model has the potential to contribute to the adoption of language model for Spanish question answering language-related tasks on resource-constrained environments, while preserving considerable levels of accuracy and robustness. As future work, we aim to produce compressed models for other NLP downstream tasks.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track.

References

Yasaman Boreshban, Seyed Morteza Mirbostani, Gholamreza Ghassem-Sani, Seyed Abolghasem Mirroshandel, and Shahin Amiriparian. Improving question answering performance using knowledge distillation and active learning, 2021.

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. Automatic spanish translation of the squad dataset for multilingual question answering, 2019.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.

Javier de la Rosa, Eduardo G. Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Marıa Grandury. Bertin: Efficient pre-training of a spanish language model using perplexity sampling, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68(0):39–60, 2022. ISSN 1989-7553. URL `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6405`.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.372. URL `https://aclanthology.org/2020.findings-emnlp.372`.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. RoBERTuito: a pre-trained language model for social media text in Spanish. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 7235–7243, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.785`.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

A Appendix

A.1 Details on the distillation process

In this section, we provide further details on the training distillation process. As mentioned in Section 2, we utilize the knowledge distillation process described in (Jiao et al., 2020) with an addition term to account for task-specific loss. More specifically, we assume that the teacher and student models have K and L Transformer layers, respectively, where $L \ll K$. To account for the discrepancy in terms of number of layers for the student and teacher models, we use a layer index mapping function, $g(l)$ = 3 x k, such that the student learns from every 3 layers of the teacher model. Therefore, the student acquires knowledge from the teacher by minimizing the following loss function:

$$\mathcal{L} = \sum_{x \in \mathcal{X}} \mathcal{L}_{\text{task}}(f^S(x), f^T(x)) + \sum_{k=0}^{K+1} \mathcal{L}_{\text{layer}}(f_k^S(x), f_{g(k)}^T(x)), \tag{2}$$

where $\mathcal{L}_{\text{layer}}$ is the loss function of a given model layer (e.g. a Transformer or embedding layer), $f_k(x)$ denotes the behavior function associated with the k-th layer of the model and $\mathcal{L}_{\text{task}}$ is the task-specific distillation loss function applied to the student and teacher models outputs. The layer loss, $\mathcal{L}_{\text{layer}}$, is calculated as the sum of two terms, $\mathcal{L}_{\text{attention}}$ and $\mathcal{L}_{\text{hidden}}$, for the attention scores and hidden representations, respectively. Specifically, $\mathcal{L}_{\text{attention}}$ is modelled as:

$$\mathcal{L}_{\text{attention}} = \frac{1}{h} \sum\nolimits_{i=1}^{h} \texttt{MSE}(\boldsymbol{A}_i^S, \boldsymbol{A}_i^T), \tag{3}$$

where h is the number of attention heads, $\boldsymbol{A}_i \in \mathbb{R}^{l \times l}$ is the i-th head attention matrix of teacher or student, l is the input text length and `MSE()` depicts the mean squared error loss function. On other side, $\mathcal{L}_{\text{hidden}}$ is defined as:

$$\mathcal{L}_{\text{hidden}} = \texttt{MSE}(\boldsymbol{H}^S \boldsymbol{W}_h, \boldsymbol{H}^T), \tag{4}$$

where the matrices $\boldsymbol{H}^S \in \mathbb{R}^{l \times d'}$ and $\boldsymbol{H}^T \in \mathbb{R}^{l \times d}$ refer to the hidden states of student and teacher networks respectively. The scalar values d and d' denote the hidden sizes of teacher and student models, and d' is often smaller than d to obtain a smaller student network. The matrix $\boldsymbol{W}_h \in \mathbb{R}^{d' \times d}$ is a learnable linear transformation, which projects the hidden states of the student network into the same dimensionality as the teacher feature space. Therefore, $\mathcal{L}_{\text{layer}}$ is expressed as:

$$\mathcal{L}_{\text{layer}} = \mathcal{L}_{\text{attention}} + \mathcal{L}_{\text{hidden}} \tag{5}$$

By following this distillation process, the smaller model is able to mimic the representation and attention of the teacher over the input sequences, while maximizing the downstream task performance.

A.2 Details on training hyperparameters and environment

Training was done using a single NVIDIA RTX A6000 GPU, with a batch size of 32, a learning rate of 3e-5, and maximum sequence length of 384 running for 20 epochs. For optimization purposes, mixed-precision training is employed, and gradient clipping with a max gradient norm of 1.0 is applied to improve training stability. The distillation process took 5.3 hours to complete. We utilize the HuggingFace library (Wolf et al., 2020) for our training implementation and the source code is available at https://github.com/anonymous/anonymous.

A.3 Experimental results on model efficiency

In Table 2 we show a comparison of model sizes and inference latency between the baseline models and our SpanishTinyRoBERTa model. The results were obtained by running on a single NVIDIA RTX A6000 GPU and averaging over 10 different runs.

Model	Layers	Hidden Size	Feed-forward Size	Size	Latency (ms)
Spanish RoBERTa-large (teacher)	24	1024	4096	355M	1683
Multilingual BERT (mBERT)	12	768	3072	179M	1187
Spanish BERT-base	12	768	3072	109M	942
Spanish RoBERTa-base	12	768	3072	124M	1015
SpanishTinyRoBERTa (ours)	6	512	3072	51M	392

Table 2: Comparison of model sizes and inference latency (in milliseconds for a single query) between the baselines and our distilled model. The number of layers does not include the embedding and prediction layers.