Regularising Fisher Information Improves Cross-lingual Generalisation

Asa Cooper Stickland
University of Edinburgh
a.cooper.stickland@ed.ac.uk

Iain Murray
University of Edinburgh
i.murray@ed.ac.uk

1 Introduction

Multilingual pre-trained representations (Devlin et al., 2019; Huang et al., 2019; Conneau et al., 2020) are ubiquitous in state-of-the-art methods for cross-lingual transfer (Wu and Dredze, 2019; Pires et al., 2019). These methods learn from raw textual data in up to hundreds of languages. A typical pipeline transfers to another language by fine-tuning a downstream task in a high-resource language, often English.

Many recent works use 'consistency regularisation' to improve the generalisation of fine-tuned pre-trained models, both multilingual and English-only (Jiang et al., 2020; Aghajanyan et al., 2020; Wang et al., 2021; Zheng et al., 2021; Park et al., 2021; Liang et al., 2021). These works encourage model outputs to be similar between a perturbed and normal version of the input, usually via penalising the Kullback–Leibler (KL) divergence between the probability distribution of the perturbed and normal model. 'Generic' perturbations can be adversarial inputs (Jiang et al., 2020) or inputs with Gaussian or uniform noise (Aghajanyan et al., 2020). For cross-lingual generalisation in particular, probabilistic subword segmentations (Kudo, 2018) of the input or translations of the input generated by machine translation can be used (Wang et al., 2021; Zheng et al., 2021). Other work has found improvement by enforcing consistency for perturbations *within* models in addition to at the input (Hua et al., 2021; Liang et al., 2021).

While these works show improved generalisation compared to 'vanilla' fine-tuning, they present multiple explanations of the effectiveness of consistency regularisation. They also rarely compare to traditional regularisation methods like dropout or L2 regularisation. Finally such methods either require a complex adversarial training step (Jiang et al., 2020; Park et al., 2021), or tuning many hyper-parameters like type of noise, level of noise, and weight given to the consistency loss term.

We believe that consistency losses may be implicitly regularizing the loss landscape. In particular, we build on the work of Jastrzebski et al. (2021), who hypothesize that implicitly or explicitly regularizing trace of the Fisher Information Matrix (FIM), $\mathrm{Tr}(F)$, amplifies the implicit bias of SGD to avoid memorization. Briefly, the FIM is defined as $\mathbf{F}(\boldsymbol{\theta}) = E_{x\sim\mathcal{X}, y\sim p_{\boldsymbol{\theta}}(y|x)}[g(x,y)g(x,y)^T]$, where $g(x,y)$ is the gradient w.r.t to $\boldsymbol{\theta}$ on the loss for label y and input x. Jastrzebski et al. (2021) propose directly penalising a proxy of $\mathrm{Tr}(F)$, the **Fisher penalty** defined as $\|\frac{1}{B}\sum_{i=1}^{B} g(x_i, y_i)\|^2$.

In the multilingual setting we may wish to first fine-tune on a high-resource language like English, then further fine-tune on a smaller amount data in a lower-resource language, a 'two-stage' fine-tuning procedure. The FIM is a measure of the local curvature, and a small $\mathrm{Tr}(F)$ at the end of training implies a flatter minimum. Intuitively, such flat minima imply we can 'travel further' in parameter space before reaching a region of high loss, allowing for better performance in the two-stage fine-tuning setting.

Our (preliminary) key contributions are

- We show that the trace of the FIM is correlated with generalisation, confirming that the results of Jastrzebski et al. (2021) apply to cross-lingual transfer.
- Adding a direct Fisher penalty can achieve similar results to subword consistency regularization.
- We show for models fine-tuned on an English downstream task, improvements from fine-tuning on data from *another* language are correlated with low curvature (i.e. small trace of the FIM) in the English fine-tuned model.

Proceedings of the 1st Workshop on Multilingual Representation Learning, pages 238–241
November 11, 2021. ©2021 Association for Computational Linguistics

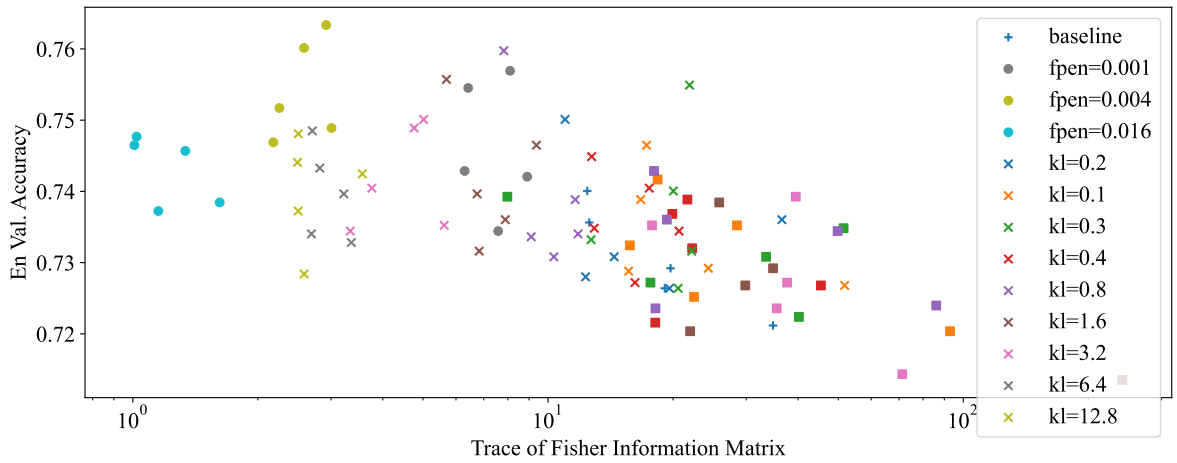

Figure 1: English validation set accuracy on a subset of the XNLI dataset vs. trace of the FIM at the end of training, for models with different weight given to consistency regularisation (crosses) and fisher penalty (circles) losses. For clarity we leave various models with different L2 penalties (squares) off the legend. kl=k means a consistency loss was given weight k during training. fpen=k means the Fisher penalty was given weight k.

2 Initial Results

We first present a theoretical argument that in some simple situations, consistency losses penalize the trace of the FIM. We perturb model parameters $\boldsymbol{\theta}$ with small, zero-mean, i.i.d. noise $\boldsymbol{\epsilon}$. For small $\boldsymbol{\epsilon}$, we can Taylor expand the KL divergence between models with these parameters: $KL\left[p_{\boldsymbol{\theta}}(y\,|\,x)\;||\;p_{\boldsymbol{\theta}+\boldsymbol{\epsilon}}(y\,|\,x)\right] \approx \frac{1}{2}\boldsymbol{\epsilon}^T F \boldsymbol{\epsilon}$, (see e.g., Dabak and Johnson, 2003). Taking expectations w.r.t. $\boldsymbol{\epsilon}$ and writing $\boldsymbol{\epsilon}^T F \boldsymbol{\epsilon}$ as a sum, we have,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\epsilon}\sim p(\boldsymbol{\epsilon})}[\boldsymbol{\epsilon}^T F \boldsymbol{\epsilon}] &= \textstyle\sum_{i,j} \mathbb{E}[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_j F_{i,j}] \\ &= \textstyle\sum_{i\neq j} \mathbb{E}[\boldsymbol{\epsilon}_i]\mathbb{E}[\boldsymbol{\epsilon}_j] F_{i,j} + \sum_i \mathbb{E}[\boldsymbol{\epsilon}_i^2] F_{i,i} \\ &= 0 + C \textstyle\sum_i F_{i,i} = C\,\mathrm{Tr}(F), \end{aligned} \tag{1}$$

where C is the variance of the i.i.d. noise.

Consistency losses that use larger or more structured perturbations could potentially have a useful effect not captured by the Fisher Information Matrix alone. We empirically investigate the relationship between a subword segmentation consistency loss and penalizing the FIM. We use the same loss and hyper-parameters as Wang et al. (2021). All experiments use multilingual BERT.

Figure 1 presents results from an experiment on a subset (20k examples, with the small size chosen due to compute constraints) of the XNLI dataset (Conneau et al., 2018). It shows 1) a correlation between generalisation and small $\mathrm{Tr}(F)$, and 2) decreasing $\mathrm{Tr}(F)$ with increasing weight given to the consistency loss term. Additionally we see that directly penalising the FIM (models marked 'fpen=') has a similar effect to these consistency losses.

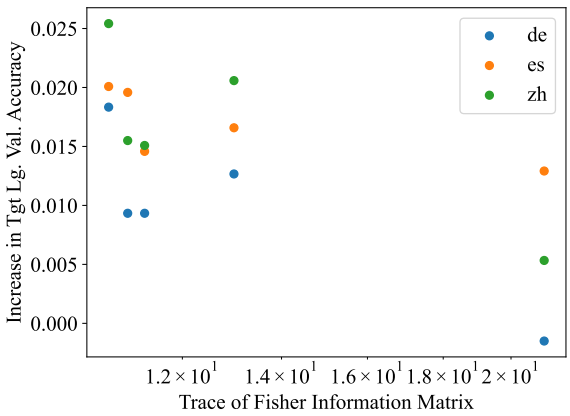

Figure 2: Increase in target language (either de, es or zh) validation set accuracy when fine-tuning on data in the target language vs. trace of the FIM after training on English, for the PAWS-X dataset (Yang et al., 2019).

Figure 2 shows the effect of small $\mathrm{Tr}(F)$, i.e. flat minima, on fine-tuning a model trained on English data on another language. To obtain non-English training data, we split the 2000 dev set examples in two, leaving 1000 training examples in each language and 1000 new dev examples. We see that improvements on the non-English language are correlated with flat minima.

We aim to confirm these initial results on more datasets, and use our insights to develop better multilingual fine-tuning techniques.

References

Armen Aghajanyan, Akshat Shrivastava, Anchit Gupta, Naman Goyal, Luke Zettlemoyer, and Sonal Gupta. 2020. Better fine-tuning by reducing representational collapse. *arXiv preprint arXiv:2008.03156*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Anand Dabak and Don Johnson. 2003. Relations between Kullback-Leibler distance and Fisher information.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hang Hua, Xingjian Li, Dejing Dou, Chengzhong Xu, and Jiebo Luo. 2021. Noise stability regularization for improving BERT fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3229–3241, Online. Association for Computational Linguistics.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Stanislaw Jastrzebski, Devansh Arpit, Oliver Åstrand, Giancarlo Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J. Geras. 2021. Catastrophic Fisher explosion: Early phase Fisher matrix impacts generalization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4772–4784. PMLR.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. 2020. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2177–2190, Online. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. 2021. R-drop: Regularized dropout for neural networks. *CoRR*, abs/2106.14448.

Jungsoo Park, Gyuwan Kim, and Jaewoo Kang. 2021. Consistency training with virtual adversarial discrete perturbation. *CoRR*, abs/2104.07284.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2021. Multi-view subword regularization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 473–482, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Bo Zheng, Li Dong, Shaohan Huang, Wenhui Wang, Zewen Chi, Saksham Singhal, Wanxiang Che, Ting Liu, Xia Song, and Furu Wei. 2021. Consistency regularization for cross-lingual fine-tuning. *CoRR*, abs/2106.08226.