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), $\text{Tr}(F)$, amplifies the implicit bias of SGD to avoid memorization. Briefly, the FIM is defined as $F(\theta) = E_{x \sim X, y \sim y(x)}[g(x, y)g(x, y)^T]$, where $g(x, y)$ is the gradient w.r.t to $\theta$ on the loss for label $y$ and input $x$. Jastrzebski et al. (2021) propose directly penalising a proxy of $\text{Tr}(F)$, 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 $\text{Tr}(F)$ 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.
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 $\theta$ with small, zero-mean, i.i.d. noise $\epsilon$. For small $\epsilon$, we can Taylor expand the KL divergence between models with these parameters:

$$\text{KL}[p_{\theta}(y \mid x) \mid \mid p_{\theta+\epsilon}(y \mid x)] \approx \frac{1}{2} \epsilon^T F \epsilon,$$

(see e.g., Dabak and Johnson, 2003). Taking expectations w.r.t. $\epsilon$ and writing $\epsilon^T F \epsilon$ as a sum, we have,

$$\mathbb{E}_{\epsilon \sim p(\epsilon)}[\epsilon^T F \epsilon] = \sum_{i,j} \mathbb{E}[\epsilon_i \epsilon_j F_{i,j}] = \sum_{i \neq j} \mathbb{E}[\epsilon_i] \mathbb{E}[\epsilon_j] F_{i,j} + \sum_i \mathbb{E}[\epsilon_i^2] F_{i,i} = 0 + C \sum_i F_{i,i} = \text{Tr}(F),$$

(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 $\text{Tr}(F)$, and 2) decreasing $\text{Tr}(F)$ 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 shows the effect of small $\text{Tr}(F)$, 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


