Robust Optimization for Multilingual Translation with Imbalanced Data

Xian Li, Hongyu Gong
xianl@fb.com

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

Problem: Multilingual training as multi-task learning (MTL) faces the challenge of task data imbalance.

🤫 Training data sub-sampling is a common trick in training multilingual models.
🤔 How to choose an appropriate sampling temperature (T)? Depends on dataset.
😩 Double bind for low-resource languages, easily underfit (T=1) or overfit (T=5,100)

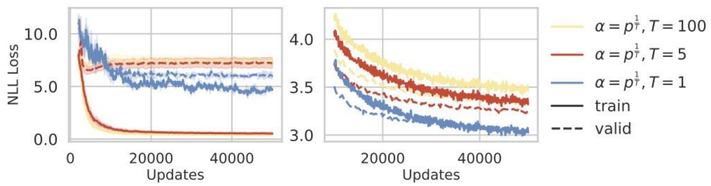

Figure 1. Train and validation loss (token-level negative log-likelihoood, NLL) for low resource (**Left**) and high resource (**Right**) from the same multilingual model. We can see that addressing data imbalance with a temperature hyperparameter T is not robust to changing data distribution.

Understanding the Why from an Optimization Perspective

🔦 Abrasive gradients between high-resource and low resource tasks.
👎 Interference (negative transfer)

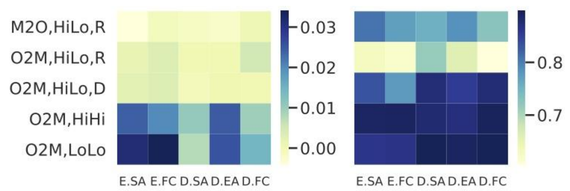

Figure 2. Gradients similarity of different Transformer parameters (x-axis) in common multilingual translation tasks (y-axis), measured as gradients direction similarity (**Left**), and gradients norm similarity (**Right**).

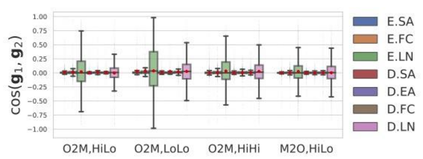

Figure 3. Gradients alignment (y-axis) for Transformer parameters across common multilingual translation tasks (x-axis), such as M2O (many-to-one), O2M (one-to-many).

Curvature Aware Task Scaling (CATS) 🐱

Key Idea

- Regularizing curvature of the shared loss surface can mitigate negative transfer (interference)
- Guide gradients to point to “flatter” regions
- Projecting gradients is expensive. Scale and combine.

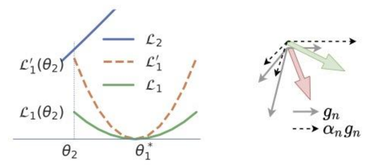

Figure 4. **Left**: Change of one language's loss L_1 after shared parameters being updated to $\boldsymbol{\theta}_2$ driven by ∇L_2 from another language is affected by the curvature of the loss landscape around previous critical point $\boldsymbol{\theta}_1^*$. **Right**: Illustration of the proposed algorithm, **C**urvature **A**ware **T**ask **S**caling (**CATS**), to learn task weighting re-scaling $\boldsymbol{\alpha}$ such that the combined gradients will guide the optimization trajectory to low curvature region (pointed by the green arrow).

Results on Optimization

Before: 👎 Competition between HiRes and LoRes; LoRes loses.
- High resource dominates the loss surface (dashed lines)
- Upsampling low resource implicitly regularizes it

After 🐱: 👍 Positive crosslingual transfer
- Reduced “sharpness”
- Increased gradients similarity

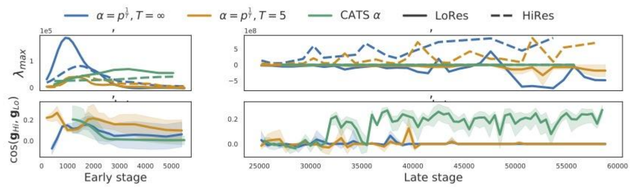

Figure 5. Local curvature measured by top eigenvalue (**Top**) and gradient direction similarity (**Bottom**) of multilingual training with high-resource (HiRes) and low resource (LoRes) languages measured on TED corpus. In the beginning of training (\textbf{Left}), HiRes and LoRes competes to increase the sharpness the loss landscape, with HiRes dominating the optimization trajectory during the rest of the training (\textbf{Right}) and their gradients are almost orthogonal. Our proposed method (CATS α) effectively reduced local curvature and improved gradients alignment.

Results on Generalization

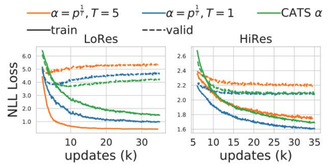

Figure 6. Train and validation loss (token-level negative log-likelihoood, NLL) for low resource (**Left**) and high resource (**Right**) on the TED dataset. CATS improves both high resource and low resource while sampling with temperature hyperparameter either overfit or underfit low resource.

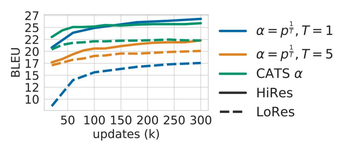

Figure 7. CATS is very effective in highly imbalanced datasets (WMT), where common sampling approaches T=1,5 sacrifice either low resources (LoRes) or high resources (HiRes). CATS significantly improves generalization on LoRes while demonstrating better sample efficiency.

Scalability

Table 1. Performance on OPUS-100. CATS can easily apply to training at the scale of ~100 languages and improves low resources.

	LoRes	MidRes	HiRes	Avg.
# examples per language # languages	< 100K 18	 29	> 1M 45	
T=1	16.4	22.8	**22.1**	21.2
T=5	26.8	24.6	18.9	22.3
CATS	**28.1**	**26.0**	19.9	**23.4**
Prev. SoTA (CLSR [1], +43% params.)	27.2	26.3	21.7	23.3

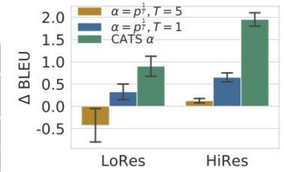

Figure 8. CATS is robust in large batch size training (4x batch size from 33K to 131K tokens).

Robust to overparameterization

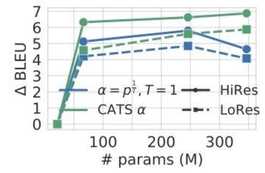

Figure 9. CATS improves generalization to overparameterized models, while standard approach suffers from overfitting.

Summary

- ✓ Remove the tradeoff between high resource and low resource performance. Improve low resource without hurting high resource.
- ✓ Robust to different data distributions.
- ✓ No more ad-hoc temperatures (T) tuning.

Broader Impact

⚖️ Recent progress in NLP enabled by scaling up model size and data is widening the gap of technology **equity** between high resource languages (such as English) and low resource languages.

🌎 Multilingual model is a promising approach to **close this gap**.

⚠️ However, current *language agnostic* multilingual models **does not** effectively improve low resources for various reasons to be understood.

📍 Our investigation in optimization is one step towards building ***truly inclusive multilingual models***.

References

[1] Biao Zhang, Ankur Bapna, Rico Sennrich, and Orhan Firat. Share or not? learning to schedule language-specific capacity for multilingual translation. In International Conference on Learning Representations, 2021.

Take a photo to learn more: