Overcoming Catastrophic Forgetting in Zero-Shot Cross-Lingual Generation
Tu Vu\textsuperscript{1,2}, Aditya Barua\textsuperscript{1}, Brian Lester\textsuperscript{1}, Daniel Cer\textsuperscript{1}, Mohit Iyery\textsuperscript{2}, Noah Constant\textsuperscript{1}

**Zero-shot cross-lingual generation (XGen)**

- **Model Tuning vs. Prompt Tuning**
  - **Model Tuning**:
    - Pre-trained Model (11B params)
    - Task A Batch
    - Task A Model (11B params)
    - Task B Model (11B params)
    - Task C Model (11B params)
  - **Prompt Tuning**
    - Task Prompts (20K params each)
    - Mixed task Batch
    - Pre-trained Model (11B params)

  - **Model Tuning** fine-tunes the entire model on each task
  - **Prompt Tuning** only requires a small amount of additional parameters while keeping the entire model frozen

- **Prompt Tuning is better than Model Tuning on larger language shifts!**

- **Zero-shot inference on other languages than English is challenging for both methods. Interestingly, Prompt Tuning can provide large gains over Model Tuning.**

**Mitigating catastrophic forgetting**

- **Mixing in unlabeled multilingual training data (MIX-UNSUP)**
  - 1% unsupervised training task (i.e., span corruption) either from the target language
  - 99% WikiLingua-0

- **Factorized prompts (FP)**
  - Explicitly factoring soft prompts into “task” and “language” components that can be recombined in novel pairings during inference

- **Mixing in unsupervised multilingual data prevents catastrophic forgetting**

**References**