Improving Your Baseline ET-EN Model

Description: In your homeworks, you have trained baseline MT models which translate from Estonian into English, analyzed their performance, and tried to improve them. Now you can try to fix more problems and make your models even better. To evaluate your model's performance, rely both on automatic metrics, such as BLEU, and on manual evaluation, aiming to fix specific issues.

Supervisor: Lisa Korotkova

Prompting GPT4Est

Description: TartuNLP has trained a few GPT-like models for Estonian: gpt-4-est-large and gpt-4-est-base. Unfortunately, these models have not yet been evaluated for any of the natural language processing tasks such as question-answering, text summarization and classification, entailment prediction, and casual/commonsense reasoning.
The most common approach to teaching such language models to solve a task is fine-tuning. However, it has been shown that few-shot, one-shot, and zero-shot learning can achieve competitive results on these tasks while preserving the model's generalization ability.
Your task would be to first get familiar with the concepts of different learning methods (for example, by reading the relevant parts in the papers "Language Models are Unsupervised Multitask Learners" and "Language Models are Few-Shot Learners"). Then, pick a suitable Estonian dataset (I can help with that ;)) and report the results with different learning methods. In order to make the scope of this project up to 3 ECTS, you can either experiment with more than one dataset or different prompting strategies (please discuss it with your TA and/or supervisor).

Supervisor: Hele-Andra Kuulmets

Self-Supervised Cell Discovery in Histopathology Images Using Vision Transformer

Description: Vision Transformers (ViT) are able to extract the information about image structure through self-supervised pretraining (DINO). This ability may prove extremely useful in medical imaging where data annotation is cumbersome and expensive. Recently, DINO was used on histopathology images and successfully distinguished cells from the background structures in different attention heads [1]. We hypothesize that transformers can learn visually similar cell types the same way. To confirm this hypothesis, we propose to train DINO on the testis histopathology image dataset from East-Tallinn Central Hospital and inspect how attention visualization matches human annotation.

Plan:

- Train DINO on testis dataset from ImageNet or histopathology weights checkpoint
- Visualize attention maps, match attention heads to four cell types (if possible)

- (Optional) fine-tune ViT encoder from DINO for cell detection task

Supervisor: Mikhail Papkov

3D Image Restoration Using Swin Transformer

Description: Shifted Window (Swin) Transformer shows superior performance in image restoration task (SwinIR). Currently, SwinIR works with 2D data only with dimensionality hardcoded in many places. We would like to see how SwinIR performs on 3D microscopy denoising and deconvolution tasks. Fortunately, 3D Swin blocks were already implemented in Video Swin, so we can use them.

Plan:

- Fork SwinIR repository
- Move 3D Swin blocks from Video Swin to SwinIR
- Modify SwinIR so it would handle both 2D and 3D data
- Denoise 3D microscopy data with Swin, compare the results to the standard U-Net denoising
- (optional) Incorporate SwinIR in self-supervised denoising framework such as Noise2Same, evaluate the results
- (optional) Incorporate SwinIR in self-supervised deconvolution framework such as SSI, evaluate the results

Supervisor: Mikhail Papkov

Transformer Encoders for Image Segmentation

Description: Using segmentation_models.pytorch and PyTorch Image Models, experiment with different encoders for microscopy image segmentation.

Supervisor: Mikhail Papkov

Tackling Unfair Music Label Contracts

Description: Contracts with a record label are difficult to review for musicians. Our system helps them find non-standard clauses or omissions from the contract that are detrimental to the musician. The data for training the system consists of contracts with phrases to be recognized annotated with tags.

Contact: Anna Aljanaki, Kristjan Heinmets

MT Quality Estimation by Reconstruction

Description: When we have parallel data, we can *evaluate* the performance of a machine translation system by comparing its outputs to the reference translations and computing automatic metrics. But when we use the system to translate new data and don't have a

reference translation, we need to rely on quality *estimation,* that is, predicting the quality of a model's output given just the input.
In this project, you can try doing quality estimation by reconstruction. The idea is to add a second, very small decoder to an MT model, and make it copy the input text. The hypothesis is that the better the model is at translating a certain sentence, the easier it will be to reproduce the source text based on its internal representations.
However, copying is a very easy task for a machine translation model, so you will probably need to start with making your systems *fail* at copying input text.

Supervisor: Lisa Korotkova

MT Domain Adaptation with Prompting

Description: We know large pre-trained language models can show interesting behaviors when prompted in specific ways (e.g. translation, question answering, code generation). Could we use prompting to modify the behavior of our translation models as well? If we have an MT model trained on multiple corpora, for example, OpenSubtitles ("Hey, dude!") and Europarl ("Dear ladies and gentlemen…"), can we use prompting to make it produce outputs in the style of a certain training corpus?
In this project, you can train an MT model and experiment with different ways of forcing it to modify some characteristics of its output. Optionally, you can experiment with a pre-trained model as well.
See also this recent paper: https://arxiv.org/abs/2202.11822

Supervisor: Lisa Korotkova

Multilingual MT with Language-Clustered Vocabularies

Description: When we create vocabularies for multilingual models, we usually train a subword segmentation model on the entire multilingual corpus at once. This may lead to poor decomposition for low-resource languages or unnecessary subword sharing between languages. This paper proposes to modify this vocabulary generation approach: the languages are clustered based on similarities of their individual subword vocabularies, SentencePIece is applied to each cluster separately, and the resulting vocabularies are combined to form the unified multilingual vocabulary. The authors experiment with massively multilingual (104 languages) Transformer language models, and their performance is assessed on downstream tasks. In this project, you can apply the proposed vocabulary generation method to multilingual machine translation (with a smaller number of languages) and see if it can improve MT quality.

Supervisor: Lisa Korotkova

M2M-100

Description: Experiment with the M2M-100 multilingual model (paper: https://arxiv.org/abs/2010.11125), fine-tune it to English↔Estonian data.

Supervisor: Lisa Korotkova

An online competition

Description: You may participate in an ongoing data science competition of your choosing. Here are a couple of options: U.S. Patent Phrase to Phrase Matching, Amazon KDD Cup'22. As our course is on Transformers, you will need to try some approaches that use Transformers. That does not mean your final system *has* to involve Transformers. If other methods work better, and you show what you tried, what worked, and what didn't, that is totally valid.

Your own idea

Description: If you have your own idea for a project, you can definitely work on that. Maybe you've found an exciting research paper and would like to replicate it, or you have some work in progress where you could apply Transformers, or you would like to experiment with a pre-trained model. Let us know, and, if you need a supervisor, we will try to find someone for you.