**EXT5: Towards Extreme Multi-Task Scaling for Transfer Learning**

The key to recent advancements in natural language processing has been transfer learning. The idea of multi-task co-training has shown some promise in earlier pioneering studies like T5 and MT-DNN. The main challenges of these methods are that the tasks must be selected carefully to show the positive affinity with respect to downstream transfer. In most the cases, the negative transfer can also be expected. While the self-supervised language modeling aim is generally used in conventional pre-training, some skills, like as common-sense knowledge, are only learned slowly even when using large volumes of unlabeled data. The construction of considerably more sample-efficient pre-training settings, therefore, becomes more crucial as ever larger models are trained, and it could be solved via multi-task learning. In this paper, Extreme Multi-Task Scaling is proposed. This is based on the rationale that, even though negative transfer is frequently observed during fine-tuning, a large and varied collection of pre-training tasks is typically preferable to an expensive search for the optimal combination of pre-training tasks. The EXT5: a T5 model, which pre-trained on a combination of supervised EXMIX and self-supervised C4 span denoising is proposed.

The EXMiX is used to explore the Extreme Task Scaling paradigm and all the tasks are formatted as text-text examples which allows multi-task training. The datasets are proportioned in such a way that they are proportionate to the individual dataset size. Subsets are created from the dataset to evaluate the transfer among the task families in the multi-task learning set up. The main goal is to pretrain a model on this to improve the downstream performance. A model is fined tuned on each pair of task families. It was observed that a set of task-family pairs show positive transfer and negative transfer, the negative transfer was predominantly seen in across task families when compared to intra-family training. It can also be observed that negative transfer during multi-task fine-tuning doesn’t inhibit the pre-training necessarily.

A trained sequence-to-sequence model called EXT5 is introduced. It is a transformer encoder-decoder model based on the well-known T5 framework. The T5.1.1 architecture is used. The standard encoder-decoder transformer is used in the T5.1.1. The encoder and decoder each consist of 12 layers which has an output dimensionality of 768. The FFN layers have hidden size with 3072 as the output dimensions. The GLU-variant-based FFN layers have three weight matrices instead of two, and the hidden layer is reduced to the output dimensions of 2048, to maintain the same parameter and operation counts as the base model. The EXT5 models uses GEGLU instead of ReLU and Adafactor is used for optimization along with inverse square root as the learning rate. The representations learned by EXT5 are more general adaptable to a new objective. Improved generalization can be seen.

This paper mainly explores how supervised multi-task learning can be used on a large scale to improve existing self-supervised pre-training strategies for NLP models. The experiments proved that even though negative transfer occurs frequently when fine-tuning on several tasks, adding more tasks to a multi-task pre-training setup will result in stronger downstream performance with improved sample efficiency. However, it can be seen that the datasets being explicitly abstracted into task families is mainly dependent on nuances pertaining to the expressiveness, nature and domain of the task family’s representative datasets. The solutions like gradient manipulation can improve extreme multi-task scaling further.

**META-DDIE: Predicting Drug–Drug Interaction Events with Few-Shot Learning**

One of the main issues in pharmaceutical research is drug-drug interactions (DDIs), and several computational techniques have been developed to determine whether two medications will interact or not. Recently, events brought on by DDIs have received greater attention, which is better for examining the mechanism underlying the use of many drugs or unfavorable reactions. Some uncommon events, however, could only have a few examples, making it difficult to accurately predict them. The methods currently in use are often intended to determine whether two drugs will interact or not, and they have significantly improved knowledge of DDIs. However, DDIs can have many biological effects, making them more beneficial for understanding the mechanism behind medication interactions or negative reactions. To represent the drug molecules, NLP techniques are used to compress the molecular representation into a denser vector. MAML is a few-shot learning meta-approach that takes the training phase into account. With a limited number of gradient descent steps, it aims to learn a representation that is simple to fit to new data. A computational method called META-DDIE, which comprises of a representation module and a comparison module, to predict DDI occurrences in order to overcome the existing challenges has been proposed in this paper. Through the representation module, META-DDIE learns the interpretable representations of DDIs using drug structures as input. It helps to identify the crucial elements that may lead to DDI events and reveals the relationships between various actions. The main aim is to build a framework, that improves the drug–drug interaction event prediction, especially on the rare events and identify the causes of DDI events through interpretable drug–drug interaction representation.

The drug chemical structures are the input for META-DDIE, which then builds a comparing module to determine whether two representations are similar and learns the representations of drug-drug interactions from those representations. This classification of rare events with a limited number of labeled examples is accomplished using META-DDIE. In each DDI event prediction, a DDI matrix consists of drugs and types of DDI-associated events. Each DDI is associated with one event. META-DDIE has two components: the drug–drug interaction representation module and the drug–drug interaction comparing module. The interpretable representations of drug–drug interactions are learnt from drug chemical structures. In order to determine whether two drug-drug interaction representations are comparable, the drug-drug interaction comparing module constructs a relation network. It mainly aims to judge whether two drug–drug interactions have the same event or not. It then uses this network under the few-shot learning framework to predict DDI events with relatively few labeled samples. The data-driven Chemical Sequential Pattern Mining (SPM) algorithm is used to learn frequent substructures of drugs from their SMILES. To obtain representation of the drug–drug interaction between the drugs, OR operation is performed on the vector of two drugs. This representation is again decoded by the latent embedding using another neural network. The sigmoid function is used to limit the range of output from 0 to 1. For the few-shot learning, the resulting projection coefficient serves as the final representation of the drug-drug interaction. Given two drug–drug interactions in the query set, the drug–drug interaction representation module produces feature vectors. This is fed into the comparing module⁠, which eventually produces an output in range of 0 to 1 that represents the similarity. This is implemented with a neural network. A 1D convolution layer with 32 kernels, each of size 5, a padding size of 1, batch normalization layer, ReLU activation and max pooling layer with a size of 5 is adopted. This is then flattened into a 1D vector. 3 fully connected layers with 25664 are used along with ReLU, dropout with a dropout rate of 0.5 and batch normalization except the output layer. In the meta-training, DDIs from support set and query set are utilized to train META-DDIE. Scores of C dimensions for each drug–drug interaction is obtained in the query set. The highest score is selected and is used to classify the drug–drug interaction. The loss function used is mean square error (MSE).

The META-DDIE produces better accuracy, and it outperforms the other methods. This indicates the model’s ability to predicting fewer events. META-DDIE produces satisfying accuracy on rare events and outperforms the existing methods. This result can be used as a pre-screening indication for the downstream experiments. This is extremely important since, in order to determine whether unusual drug interaction events exist, typical machine learning algorithms fall short and necessitate extensive, time-consuming, labor-intensive experiments. However, to improve feature extraction from the molecules while keeping interpretability, graph models may be utilized. In cases where there is an instance of only one example, these can be either in support set or query set⁠. This prevents them from being classified. Zero-shot learning offers a viable approach to resolving problem.