KIMERA: Injecting Domain Knowledge into Vacant Transformer Heads

Benjamin Winter*, Alexei Figueroa*, Alexander Löser, Felix A. Gers, Amy Siu

Berliner Hochschule für Technik

[Benjamin.Winter, afigueroa, aloeser, gers, aisy]@bht-berlin.de

Training transformer language models requires vast amounts of text and computational resources hindering their usage in niche domains where task-domain-specific training data is scarce. We choose the clinical domain because of this, whereas only structured data is readily available. We leverage recent findings in model compression and propose KIMERA (Knowledge Injection via Mask Enforced Retraining of Attention) for detection, retraining and instilling attention heads with complementary structured domain knowledge. Our novel multi-task training scheme effectively identifies and targets individual attention heads that are least useful for a given downstream task and improves their representation with information from structured data. KIMERA achieves significant performance boosts on seven datasets in the medical domain in Information Retrieval and Clinical Outcome Prediction settings. We apply KIMERA to BERT-base to evaluate the extent of the domain transfer and also improve on the already strong results of BioBERT in the clinical domain.

1 Contributions

• Applying model compression-based analysis for targeted retraining of attention heads
• A novel Multi-Task retraining scheme based on Knowledge Graph Completion to integrate structured knowledge
• Experiments on 5 different strategies to employ our method
• An evaluation on domain adaptation to the medical domain in 8 downstream tasks over both BERT-base and BioBERT
• We publish PyTorch code and plan to upload trained models to huggingface.co

2 Methods

An overview of our method is depicted in Figure 1A). We start with a pre-trained transformer model, a domain-specific knowledge graph, and a downstream task within that domain that we desire to improve on. KIMERA is composed of three major steps:

1. Compute the attention head importance of a fine-tuned model on the downstream task we intend to improve on.
2. Retrain the less essential heads (using the attention mask generated in step 1) of a pre-trained model using a multi-task knowledge graph generation scheme.
3. Fine-tune and evaluate the retrained model on the downstream task.

3 Datasets and downstream tasks

3.1 Knowledge Graphs

UMLS[DBLP:journals/nar/Bodenreider04] is the Unified Medical Language System an aggregation of different medical knowledge sources. This work specifically focuses on UMLS’ Metathesaurus, which contains diseases, symptoms, medications, etc., and the relations between them. From the 80 million relationship triplets in UMLS, we filter for relevant relation types, triplets that are complete, and choose to keep only well-populated sub-relations with more than 10k sample triplets. This results in our training corpus of 600k triples.

3.2 Clinical Answer Retrieval (CAPR)

Retrieving documents and passages from clinical documents is an important task in the medical domain. We evaluate our models on the clinical answer passage retrieval task (CAPR) [7] in a zero-shot setting and across four different datasets. The zero-shot setting puts an even higher burden on each individual model since each model is evaluated as-is, and not fine-tuned to the evaluated datasets. We follow [7] and evaluate our models using the Cross Encoder Architecture [7], which calculates matching scores over the joint sequence of all query and passage pairs. We use the same training and evaluation described in [7] and train on Wikipedia articles, and evaluate on WikiSectionMimicIII, Mimic-III clinical notessummarize, MedQuadmedquad, and HealthQAhealthqa datasets. In this setting, we create only one retrained model for the Clinical Outcome Prediction tasks. They are based on relevant relation types, triplets that are complete, and choose to keep only well-populated sub-relations with more than 10k sample triplets. This results in our training corpus of ~600k triplets.

3.3 Clinical Outcome Prediction (COP)

We adopt the admission notes dataset by [7] for the Clinical Outcome Prediction tasks. They are based on special filtering of Mimic-III’s discharge summaries that simulate patient information at the time of admission. This is achieved by only keeping the following sections: Chief complaint, (History of Present illness, Medical history, Admission Medications, Allergies, Physical exam, Family history, Social history. In particular, this filtering hides all information about the course and outcome of treatment of the patient during their stay.

4 Experiments and results

<table>
<thead>
<tr>
<th>Model</th>
<th>MSquad</th>
<th>HealthQA</th>
<th>Mimic-III</th>
<th>Wiki</th>
<th>MP</th>
<th>LOS</th>
<th>DIA</th>
<th>PRO</th>
</tr>
</thead>
<tbody>
<tr>
<td>R0 v. R1</td>
<td>59.05</td>
<td>78.82</td>
<td>50.71</td>
<td>32.88</td>
<td>70.61</td>
<td>64.89</td>
<td>54.89</td>
<td>79.69</td>
</tr>
<tr>
<td>BERT-base</td>
<td>81.13</td>
<td>88.79</td>
<td>59.05</td>
<td>70.61</td>
<td>64.89</td>
<td>54.89</td>
<td>79.69</td>
<td>80.69</td>
</tr>
<tr>
<td>KIMERA scratch</td>
<td>60.17</td>
<td>87.82</td>
<td>55.36</td>
<td>75.10</td>
<td>64.89</td>
<td>54.89</td>
<td>79.69</td>
<td>80.69</td>
</tr>
<tr>
<td>KIMERA no-mask</td>
<td>64.68</td>
<td>77.82</td>
<td>68.07</td>
<td>81.82</td>
<td>64.89</td>
<td>54.89</td>
<td>79.69</td>
<td>80.69</td>
</tr>
<tr>
<td>KIMERA mask</td>
<td>73.70</td>
<td>81.82</td>
<td>71.87</td>
<td>82.71</td>
<td>64.89</td>
<td>54.89</td>
<td>79.69</td>
<td>80.69</td>
</tr>
<tr>
<td>BioBERT</td>
<td>78.86</td>
<td>97.35</td>
<td>66.89</td>
<td>81.81</td>
<td>64.89</td>
<td>54.89</td>
<td>79.69</td>
<td>80.69</td>
</tr>
</tbody>
</table>

Figure 2: Results across the four CAPR datasets using the Cross Encoder architecture[7] and four COP task(right). Top part shows scores for models based on BERT-base, bottom part scores for models on BioBERT. KIMERA improves on both BERT-base and BioBERT with the exception of the LOS task.

5 Discussion

A) Head-mask used for retraining. B) and C) present the head importances $h_i$ before and after using KIMERA, respectively. Our method results in relatively higher and more homogeneous importance of the heads.

Figure 5: Importance changes per layer for the CAPR task. A) Average importance $l_h$ per layer before and after KIMERA. B) Number of retrained heads that saw an increase/decrease in their importance after KIMERA. C) Number of frozen heads that saw an increase/decrease in importance with our method. The retrained heads present an overall increase in importance, whereas the frozen heads show mixed results.

Figure 6: Mean importance scores $l_h$ before and after KIMERA for frozen and retrained heads in the CAPR task. $l_h$ more than doubles for the retrained heads while it moderately decreases for the frozen heads.