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, asiu}@bht-berlin.de

*Both authors contributed equally.

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 detecting, 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[1] and plan to upload trained models to `huggingface.co`

2 Methods

An overview of our method is depicted in Figure 2 **A)**. 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.

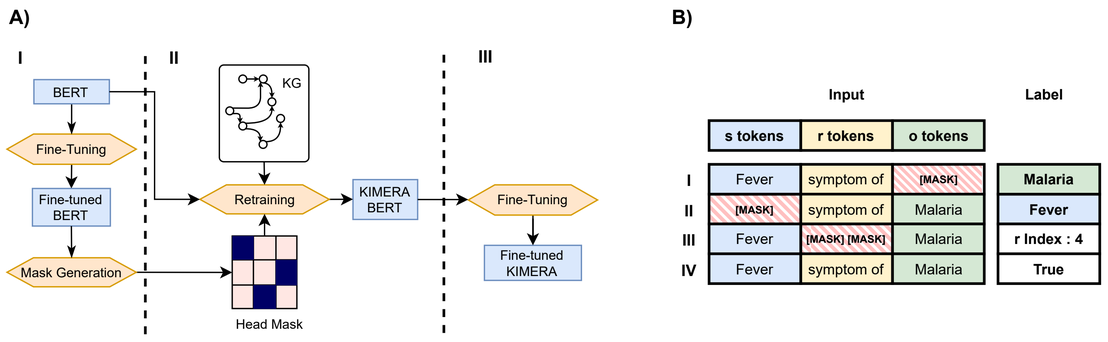

Figure 1: **A)** KIMERA consists of three phases: **I** A transformer model is fine-tuned and a head-mask is computed by identifying redundancies. **II** The computed mask is then used in conjunction with a multi-task training based on knowledge graph completion. Finally, the model is fine-tuned on the target task. **III** The retrained model is fine-tuned on the domain-specific task to culminate the domain transfer. **B)** Examples of KG retraining tasks. **I** and **II** *Entity Prediction* with a Masked Language Modelling objective. **III** *Relation Prediction* with a multi-class classification objective, and **IV** *Triplet Classification* with a binary classification objective.

$$W_{i+1}^{lh} = W_i^{lh} - \eta(1 - m^{lh})\nabla\mathcal{L} \tag{1}$$

3 Datasets and downstream tasks

3.1 Knowledge Graphs

UMLSDBLP:journals/nar/Bodenreider04 The Unified Medical Language System is 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 Passage 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) [**?**] 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 [**?**] and evaluate our models using the Cross Encoder Architecture [**?**], which calculates matching scores over the joint sequence of all query and passage pairs. We use the same training and evaluation described in [**?**] and train on Wikipedia articles, and evaluate on WikiSectionQAwikisectionqa, Mimic-III clinical notesmimiciii, MedQuadmedquad, and HealthQAhealthqa datasets. In this setting, we create only one joint attention-head mask for all four tasks. This mask is generated on a dataset that is combined from held out parts of the test sets of each of the datasets.

3.3 Clinical Outcome Prediction(COP)

We adopt the admission notes dataset by [**?**] 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.

[1] https://anonymous.4open.science/r/kg-transformers/README.md

4 Experiments and results

Model	MedQuad		HealthQA		Mimic-III		Wiki		MP	LOS	DIA	PRO
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	AUROC	AUROC	AUROC	AUROC
BERT-base	52.63	60.80	40.30	81.82	59.74	72.07	35.44	77.66	81.13	**70.40**	82.08	85.84
BERT-base(pruned)	50.71	60.45	39.92	78.12	61.96	72.64	35.23	75.12	81.07	70.14	80.21	83.48
KIMERA scratch	32.88	74.17	31.23	**83.45**	23.63	41.77	20.63	59.85	75.75	65.74	51.1	64.91
KIMERA no-mask	64.68	92.33	49.01	80.31	65.68	79.78	50.38	80.44	81.63	69.55	82.47	85.91
KIMERA hard-mask	**71.94**	**94.52**	**50.53**	82.71	67.13	80.52	**51.73**	80.72	**81.88**	69.02	**82.59**	**85.95**
KIMERA soft-mask	70.33	93.81	49.50	81.69	67.94	**81.82**	51.25	**81.31**	81.20	68.11	82.35	85.49
KIMERA b+f	70.41	93.91	49.22	80.99	**68.07**	80.43	50.81	81.24	65.72	55.36	81.45	84.21
BioBERT	78.86	97.06	62.07	91.59	64.89	78.81	61.31	90.69	82.55	**71.59**	82.81	86.36
KIMERA BioBERT	**79.74**	**97.93**	**64.14**	**92.26**	**65.22**	**79.02**	**62.48**	**94.32**	**82.87**	71.42	**83.56**	**88.44**

Figure 2: Results across the four CAPR datasets using the Cross Encoder architecture(left) and four COP tasks(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 performance, with the exception of the LOS task.

Model	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	WNLI	Mean
BERT-base	59.05	**93.34**	**89.37**	**88.79**	89.84	**85.12**	**91.78**	**69.31**	49.30	79.54
BioBERT	43.70	91.28	88.51	88.15	89.59	83.97	90.84	67.50	32.39	75.10
KIMERA no-mask	60.17	92.20	87.71	88.12	89.53	84.49	90.35	67.50	60.17	80.02
KIMERA hard-mask	**62.06**	93.00	88.93	88.53	**90.63**	84.65	91.15	69.12	**62.05**	**81.13**

Figure 3: Results of the GLUE benchmark, choosing the best of 10 seeds. KIMERA consistently outperforms BioBERT, and shows improvements over BERT-base in 3 tasks, having the highest mean score of tested models.

5 Discussion

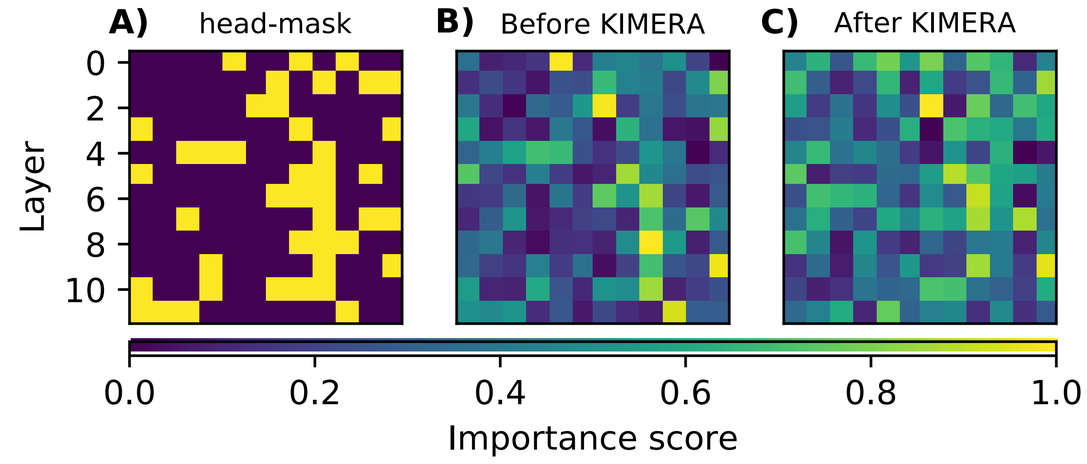

Figure 4: Attention head importance with and without KIMERA for the CAPR task. **A)** Head mask used for retraining. **B)** and **C)** present the head importances I_h before and after using KIMERA, respectively. Our method results in relatively higher and more homogeneous importance of the heads.

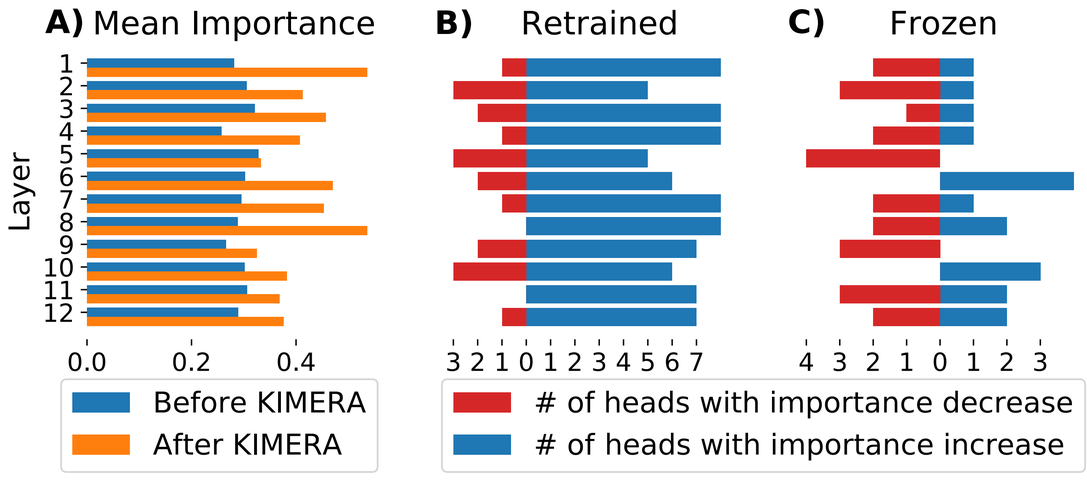

Figure 5: Importance changes per layer for the CAPR task. **A)** Average importance I_h per layer before and after KIMERA. **B)** Number of *retrained* heads that saw an increase/decrease in their importance after KIMERA. **C)** Number of *frozen* heads that saw an increase/decrease in importance with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

Heads	I_h Before KIMERA	I_h After KIMERA
Frozen	0.60	0.53
Retrained	0.17	0.37

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