Adapting a Language Model While Preserving its General Knowledge

Zixuan Ke[1], Yijia Shao[2], Haowei Lin[2], Hu Xu[3], Lei Shu[4] and Bing Liu[1]
University of Illinois at Chicago[1], Peking University[2], Meta AI[3] and Google Resaerch[4]

DA-training --- General knowledge preservation and LM Adaptation

- DA-training (a.k.a., domain-adaptive pre-training or pre-finetuning or post-training) helps an LM achieves better results
- **However,** existing DA-training simply use MLM loss
  - It does not explicitly identify what should be **preserved** and what should be **updated**
- Two needs:
  - General language knowledge should be preserved as much as possible, because target domain data is not large enough to learn that
  - LM should be specialized/adapted to the target domain due to polysemy (focus of the existing methods, may destroy useful general knowledge)
- **Our goal:** a more **informed adaptation**
  - Preserve the general knowledge
  - Integrate the adapted domain knowledge and the preserved general knowledge

Background: Pruning

- Large LM is over-parameterized. Many parameters are prune-able without affecting the performance
- A popular way is to prune a parameter by its absolute value or importance (indicated by gradient)
- From gradient to Importance

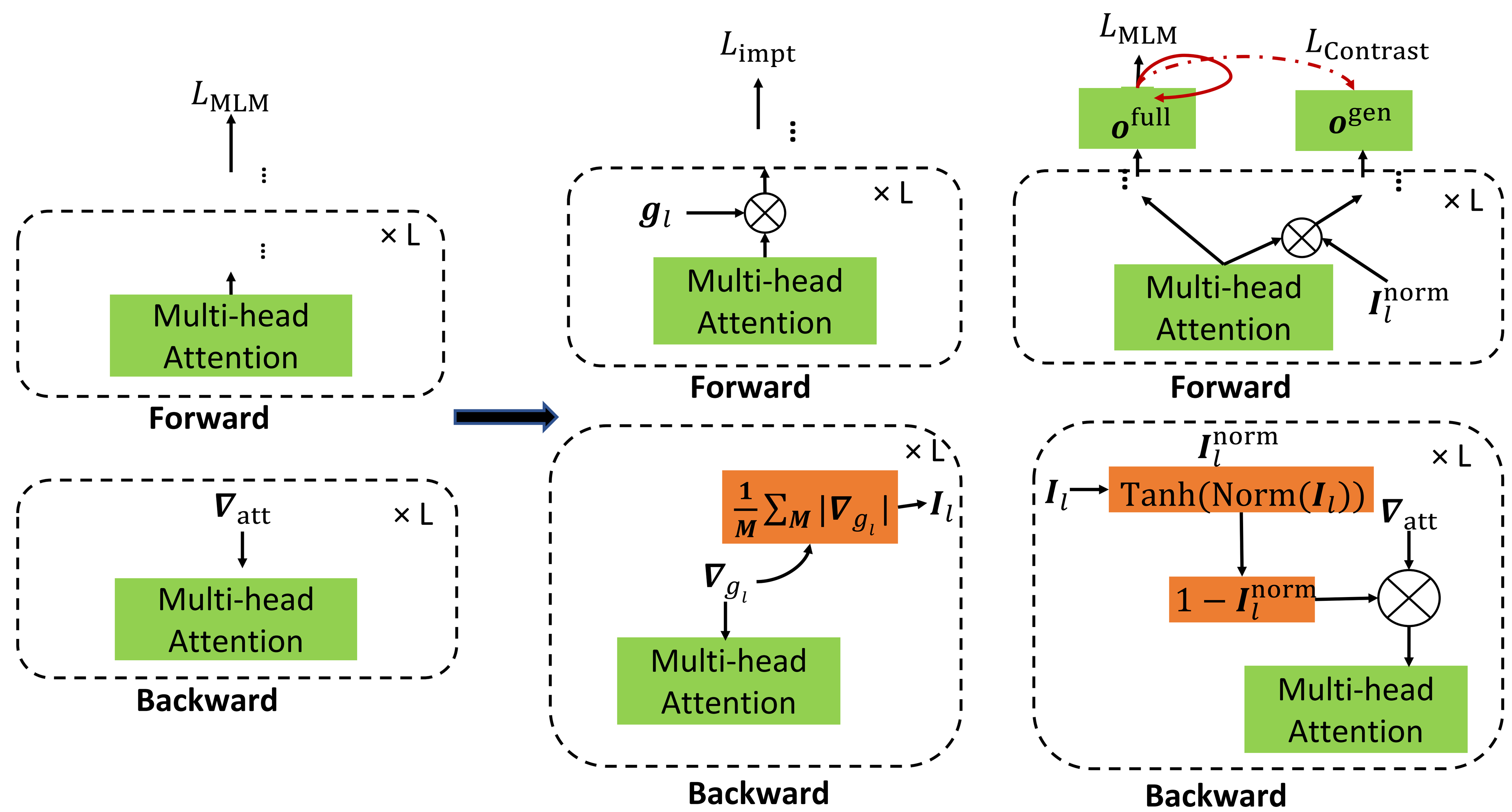
$$\nabla_{g_l} = \frac{\partial \mathcal{L}_{\text{impt}}(x_m, y_m)}{\partial g_l} \qquad I_l = \frac{1}{M}\sum_M |\nabla_{g_l}|$$

- $\mathcal{L}_{\text{impt}}$ is typically the cross-entropy loss since pruning is typically for a supervised task

- **Challenge**: if we use the domain data at hand and MLM as $\mathcal{L}_{\text{impt}}$, ∇_{g_l} only indicates the importance score of domain-specific knowledge.
- The **key** is to decide the $\mathcal{L}_{\text{impt}}$

Code, data, post-trained models:
https://github.com/ UIC-Liu-Lab /DGA

Proposed DGA Model: soft-masking and contrastive learning



Existing Post-training **DGA**

- **Preserving the general knowledge by soft-masking**
  - **Idea:** Detect importance of units for general knowledge (inspired from pruning)

  - Our **goal** is to estimate the importance of units for *general knowledge,* which requires the data used in pre-training the LM. However, this is not accessible to DA-training users.
  - **Proposed solution: Proxy KL-divergence loss**

$$\mathcal{L}_{\text{impt}} = \text{KL}(f_1(x_m), f_2(x_m))$$

  - We use *robustness* as the proxy: if an importance has high score, it indicates that it is important to the LM's robustness because its change can cause the LM to change a great deal
  - To compute the robustness, we input the domain data twice and compute the KL-divergence. f_1 and f_2 are the LM with different dropout masks (already implemented in the standard Transformer)
  - **Proposed solution: Soft-masking**

$$\nabla'_l = (1 - \text{Tanh}(\text{Norm}(I_l))) \otimes \nabla_l$$

  - Soft-mask the gradient to protect the important units for general knowledge

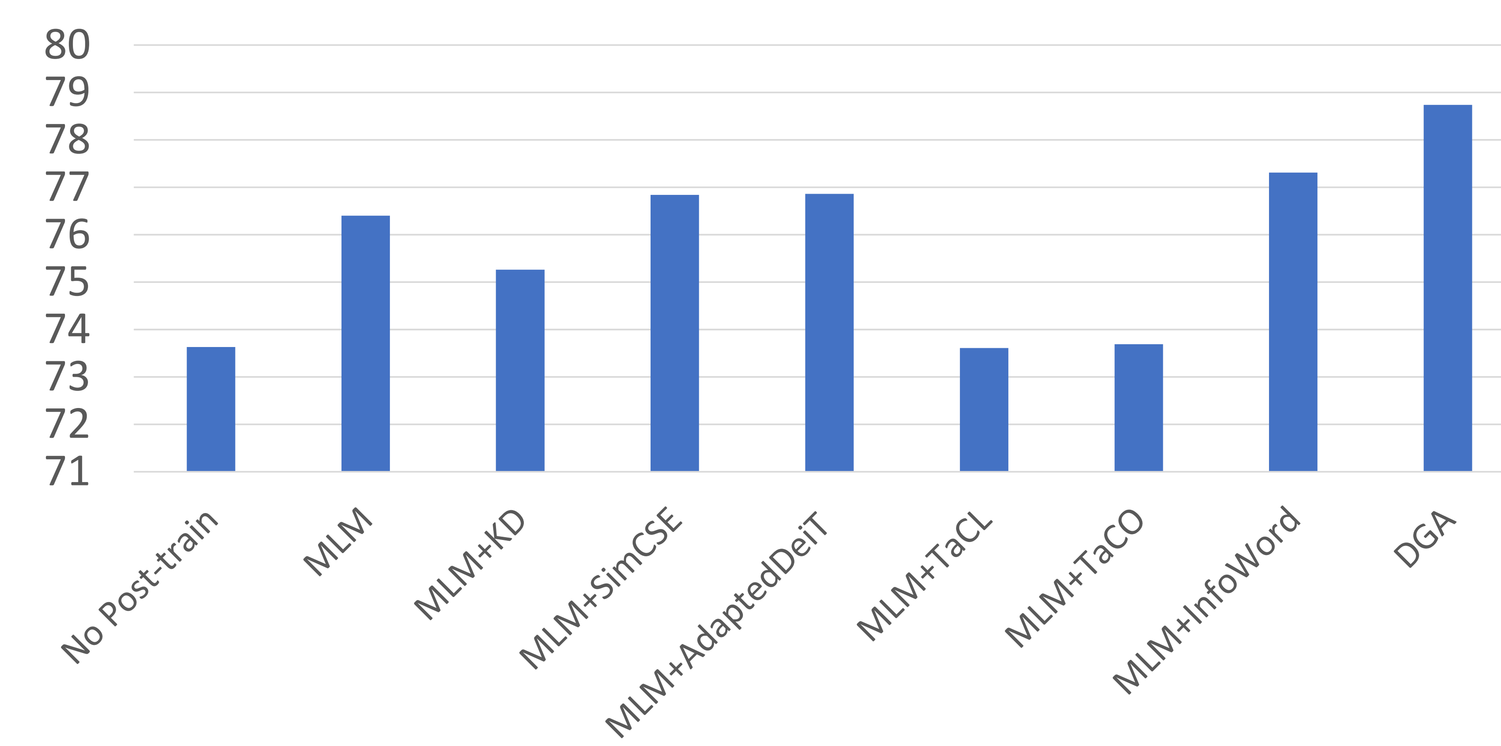
- **Integrate knowledge via contrastive learning**
  - **Idea:** encourage the learning of domain-specific knowledge in o^{full} that is not already in the general knowledge and yet related to and integrated with the general knowledge o^{gen}.
  - **Obtaining general knowledge:** plugin the importance score I_l
  - **Obtaining full knowledge:** using all units in the layer
  - **Contrastive learning:**

$$\text{L}_{\text{contrast}} = -\log \frac{e^{\text{sim}\left(o_m^{\text{full}}, o_m^{\text{full}+}\right)/\tau}}{\sum_{j=1}^{N}\left(e^{\text{sim}\left(o_m^{\text{full}}, o_j^{\text{full}}\right)/\tau} + e^{\text{sim}\left(o_m^{\text{full}}, o_j^{\text{gen}}\right)/\tau}\right)}$$

Dataset

Unlabeled Domain Datasets			End-Task Classification Datasets				
Source	Dataset	Size	Dataset	Task	#Training	#Testing	#Classes
Reviews	Yelp Restaurant	758MB	Restaurant	Aspect Sentiment Classification (ASC)	3,452	1,120	3
	Amazon Phone	724MB	Phone	Aspect Sentiment Classification (ASC)	239	553	2
	Amazon Camera	319MB	Camera	Aspect Sentiment Classification (ASC)	230	626	2
Academic Papers	ACL Papers	867MB	ACL	Citation Intent Classification	1,520	421	6
	AI Papers	507MB	AI	Relation Classification	2,260	2,388	7
	PubMed Papers	989MB	PubMed	Chemical-protein Interaction Prediction	2,667	7,398	13

Experimental Results



✓ Outperforms **10** SOTA baselines, including MLM, KD, SimCSE, TaCL, InfoWord etc.

Summary

✓ An effective DA-training method that can effectively integrate the domain knowledge to the general knowledge in the LM