VarMAE: Pre-training of Variational Masked Autoencoder for Domain-adaptive Language Understanding

Dou Hu[1,2,3*], **Xiaolong Hou**[3], **Xiyang Du**[3], **Mengyuan Zhou**[3], **Lianxin Jiang**[3], **Yang Mo**[3], **Xiaofeng Shi**[3]
[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] Ping An Life Insurance Company of China, Ltd.
hudou@iie.ac.cn, {houxiaolong430, duxiyang037, zhoumengyuan425, jianglianxin769, moyang853, shixiaofeng309}@pingan.com.cn

Abstract

Pre-trained language models have achieved promising performance on general benchmarks, but underperform when migrated to a specific domain. Recent works perform pre-training from scratch or continual pre-training on domain corpora. However, in many specific domains, the limited corpus can hardly support obtaining precise representations. To address this issue, we propose a novel Transformer-based language model named VarMAE for domain-adaptive language understanding. Under the masked autoencoding objective, we design a context uncertainty learning module to encode the token's context into a smooth latent distribution. The module can produce diverse and well-formed contextual representations. Experiments on science- and finance-domain NLU tasks demonstrate that VarMAE can be efficiently adapted to new domains with limited resources.

1 Introduction

Pre-trained language models (PLMs) have achieved promising performance in natural language understanding (NLU) tasks on standard benchmark datasets (Wang et al., 2018; Xu et al., 2020). Most works (Devlin et al., 2019; Liu et al., 2019) leverage the Transformer-based pre-train/fine-tune paradigm to learn contextual embedding from large unsupervised corpora. Masked autoencoding, also named masked language model in BERT (Devlin et al., 2019), is a widely used pre-training objective that randomly masks tokens in a sequence to recover. The objective can lead to a deep bidirectional representation of all tokens in a BERT-like architecture. However, these models, which are pre-trained on standard corpora (e.g., Wikipedia), tend to underperform when migrated to a specific domain due to the *distribution shift* (Lee et al., 2020).

Recent works perform pre-training from scratch (Gu et al., 2022; Yao et al., 2022) or continual pre-training (Gururangan et al., 2020; Wu et al., 2022) on large domain-specific corpora. But in many specific domains (e.g., finance), effective and intact unsupervised data is difficult and costly to collect due to data accessibility, privacy, security, etc. The limited domain corpus may not support pre-training from scratch (Zhang et al., 2020), and also greatly limit the effect of continual pre-training due to the *distribution shift*. Besides, some scenarios (i.e., non-industry academics or professionals) have limited access to computing power for training on a massive corpus. Therefore, how to obtain effective contextualized representations from the limited domain corpus remains a crucial challenge.

Relying on the distributional similarity hypothesis (Mikolov et al., 2013a) in linguistics, that similar words have similar contexts, masked autoencoders (MAEs) leverage co-occurrence between the context of words to learn word representations. However, when pre-training on the limited corpus, most word representations can only be learned from fewer co-occurrence contexts, leading to sparse word embedding in the semantic space. Besides, in the reconstruction of masked tokens, it is difficult to perform an accurate point estimation (Li et al., 2020) based on the partially visible context for each word. That is, the possible context of each token should be diverse. The limited data only provides restricted context information, which causes MAEs to learn a relatively poor context representation in a specific domain.

To address the above issue, we propose a novel **Var**iational **M**asked **A**utoencoder (**VarMAE**), a regularized version of MAEs, for a better domain-adaptive language understanding. Based on the vanilla MAE, we design a context uncertainty learning (CUL) module for learning a precise context representation when pre-training on a limited corpus. Specifically, the CUL encodes the token's point-estimate context in the semantic space into a smooth latent distribution. And then, the module

*This work was done when the author was at Ping An.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6276–6286
December 7-11, 2022

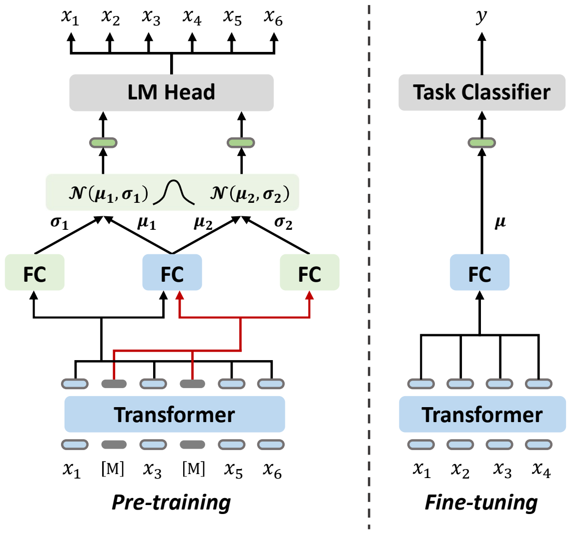

Figure 1: The architecture of VarMAE. Based on the vanilla MAE, a CUL module is used to learn diverse and well-formed context representations for all tokens.

reconstructs the context using feature regularization specified by prior distributions of latent variables. In this way, latent representations of similar contexts can be close to each other and vice versa (Li et al., 2019). Accordingly, we can obtain a smoother space and more structured latent patterns.

We conduct continual pre-training on unsupervised corpora in two domains (science and finance) and then fine-tune on the corresponding downstream NLU tasks. The results consistently show that VarMAE outperforms representative language models including vanilla pre-trained (Liu et al., 2019) and continual pre-training methods (Gururangan et al., 2020), when adapting to new domains with limited resources. Moreover, compared with masked autoencoding in MAEs, the objective of VarMAE can produce a more diverse and well-formed context representation.

2 VarMAE

In this section, we develop a novel Variational Masked Autoencoder (VarMAE) to improve vanilla MAE for domain-adaptive language understanding. The overall architecture is shown in Figure 1. Based on the vanilla MAE, we design a context uncertainty learning (CUL) module for learning a precise context representation when pre-training on a limited corpus.

2.1 Architecture of Vanilla MAE

Masking We randomly mask some percentage of the input tokens and then predict those masked tokens. Given one input tokens $X = \{x_1, ..., x_n\}$ and n is the sentence length, the model selects a random set of positions (integers between 1 and n) to mask out $M = \{m_1, ..., m_k\}$, where $k = \lceil 0.15n \rceil$ indicates 15% of tokens are masked out. The tokens in the selected positions are replaced with a `[MASK]` token. The masked sequence can be denoted as $X^{\text{masked}} = \text{REPLACE}(X, M, \texttt{[MASK]})$.

Transformer Encoder Vanilla MAE usually adopts a multi-layer bidirectional Transformer (Vaswani et al., 2017) as basic encoder like previous pre-training model (Liu et al., 2019). Transformer can capture the contextual information for each token in the sentence via self-attention mechanism, and generate a sequence of contextual embeddings. Given the masked sentence X^{masked}, the context representation is denoted as $\mathbf{C} = \{\mathbf{c}_1, ..., \mathbf{c}_N\}$.

Language Model Head We adopt the language model (LM) head to predict the original token based on the reconstructed representation. The number of output channels of LM head equals the number of input tokens. Based on the context representation $\mathbf{c}_i$, the distribution of the masked prediction is estimated by: $p_\theta(\mathbf{x}_i|\mathbf{c}_i) = softmax(\mathbf{W}\mathbf{c}_i + \mathbf{b})$, where $\mathbf{W}$ and $\mathbf{b}$ denote the weight matrices of one fully-connected layer. θ refers to the trainable parameters. The predicted token can be obtained by $x' = arg\max_i p_\theta(\mathbf{x}_i|\mathbf{c}_i)$, where x' denotes the predicted original token.

2.2 Context Uncertainty Learning

Due to the flexibility of natural language, one word may have different meanings under different domains. In many specific domains, the limited corpus can hardly support obtaining precise representations. To address this, we introduce a context uncertainty learning (CUL) module to learn regularized context representations for all tokens. These tokens include masked tokens with more noise and unmasked tokens with less noise. Inspired by variational autoencoders (VAEs) (Kingma and Welling, 2014; Higgins et al., 2017), we use latent variable modeling techniques to quantify the *aleatoric uncertainty*[1] (Der Kiureghian and Ditlevsen, 2009; Abdar et al., 2021) of these tokens.

Let us consider the input token x is generated with an unobserved continuous random variable $\mathbf{z}$. We assume that x_i is generated from a conditional

[1]The aleatoric uncertainty, or data uncertainty, is the uncertainty that captures noise inherent in the observations.

distribution $p_\theta(\mathbf{x}|\mathbf{z})$, where $\mathbf{z}$ is generated from an isotropic Gaussian prior distribution $p_\theta(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$. To learn the joint distribution of the observed variable x and its latent variable factors $\mathbf{z}$, the optimal objective is to maximize the marginal log-likelihood of x in expectation over the whole distribution of latent factors $\mathbf{z}$:

$$\max_\theta \mathbb{E}_{p_\theta(\mathbf{z})}[\log p_\theta(\mathbf{x}|\mathbf{z})]. \tag{1}$$

Since masked and unmasked tokens have relatively different noise levels, the functions to quantify the *aleatoric uncertainty* of these two types should be different. We take CUL for masked tokens as an example. Given each input masked token x_i^m and its corresponding context representation $\mathbf{c}_i^m$, the true posterior $p_\theta(\mathbf{z}^m|x_i^m)$ is approximated as $p_{\theta'}(\mathbf{z}^m|\mathbf{c}_i^m)$ due to the distributional similarity hypothesis (Mikolov et al., 2013a). Inspired by Kingma and Welling (2014), we assume $p_{\theta'}(\mathbf{z}^m|\mathbf{c}_i^m)$ takes on an approximate Gaussian form with a diagonal covariance, and let the variational approximate posterior be a multivariate Gaussian with a diagonal covariance structure. This variational approximate posterior is denoted as $q_\phi(\mathbf{z}^m|\mathbf{c}_i^m)$:

$$q_\phi(\mathbf{z}^m|\mathbf{c}_i^m) = \mathcal{N}(\mathbf{z}^m; \boldsymbol{\mu}_i^m, \boldsymbol{\sigma}_i^{m2}\mathbf{I}), \tag{2}$$

where $\mathbf{I}$ is diagonal covariance, ϕ is the variational parameters. Both parameters (mean as well as variance) are input-dependent and predicted by MLP (a fully-connected neural network with a single hidden layer) , i.e., $\boldsymbol{\mu}_i^m = f_{\phi_\mu}(\mathbf{c}_i^m)$, $\boldsymbol{\sigma}_i^m = f_{\phi_\sigma}(\mathbf{c}_i^m)$, where ϕ_μ and ϕ_σ refer to the model parameters respectively w.r.t output $\boldsymbol{\mu}_i^m$ and $\boldsymbol{\sigma}_i^m$. Next, we sample a variable $\mathbf{z}_i^m$ from the approximate posterior $q_\phi(\mathbf{z}^m|\mathbf{c}_i^m)$, and then feed it into the LM head to predict the original token.

Similarly, CUL for each unmasked token $x_i^{\bar{m}}$ adopts in a similar way and samples a latent variable $z_i^{\bar{m}}$ from the variational approximate posterior $q_\phi(\mathbf{z}^{\bar{m}}|\mathbf{c}_i^{\bar{m}}) = \mathcal{N}(\mathbf{z}^{\bar{m}}; \boldsymbol{\mu}_i^{\bar{m}}, \boldsymbol{\sigma}_i^{\bar{m}2}\mathbf{I})$, where $\mu_i^{\bar{m}}$ and $\boldsymbol{\sigma}_i^{\bar{m}}$ are predicted by MLP.

In the implementation, we adopt f_{ϕ_μ} with shared parameters to obtain $\boldsymbol{\mu}^m$ and $\boldsymbol{\mu}^{\bar{m}}$. Conversely, two f_{ϕ_σ} with independent parameters are used to obtain $\boldsymbol{\sigma}^m$ and $\boldsymbol{\sigma}^{\bar{m}}$, for x^m with more noise and $x^{\bar{m}}$ with less noise, respectively. After that, batch normalization (Ioffe and Szegedy, 2015) is applied to avoid the *posterior collapse*[2] (Zhu et al., 2020). By applying the CUL module, the context representation is not a deterministic point embedding any more, but a stochastic embedding sampled from $\mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\sigma}^2\mathbf{I})$ in the latent space. Based on the reconstructed representation, the LM head is adopted to predict the original token.

[2]The posterior collapse, or KL vanishing, is that the decoder in VAE learns to reconstruct data independent of the latent variable $\mathbf{z}$, and the KL vanishes to 0.

2.3 Training Objective

To learn a smooth space where latent representations of similar contexts are close to each other and vice versa, the objective function is:

$$\begin{aligned} &\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}}[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{c})}[\log p_\theta(\mathbf{x}|\mathbf{z})]], \\ &\qquad \text{s.t. } D_{KL}(q_\phi(\mathbf{z}|\mathbf{c})\|p_\theta(\mathbf{z})) < \delta, \end{aligned} \tag{3}$$

where $\delta > 0$ is a constraint, and $q_\phi(\mathbf{z}|\mathbf{c})$ is the variational approximate posterior of the true posterior $p_\theta(\mathbf{z}|x)$ (see Section 2.2). $D_{KL}(\cdot)$ denotes the KL-divergence term, which serves as the regularization that forces prior distribution p_θ to approach the approximated posterior q_ϕ. Then, for each input sequence, the loss function is developed as a weighted sum of loss functions for masked tokens $\mathcal{L}^m$ and unmasked tokens $\mathcal{L}^{\bar{m}}$. The weights are normalization factors of masked/unmasked tokens in the current sequence.

$$\begin{aligned} \mathcal{L}^\tau &= \mathbb{E}_{\mathbf{z}^\tau\sim q_\phi(\mathbf{z}^\tau|\mathbf{c}^\tau)}[\log p_\theta(\mathbf{x}^\tau|\mathbf{z}^\tau)] \\ &- \lambda^\tau D_{KL}(q_\phi(\mathbf{z}^\tau|\mathbf{c}^\tau)\|p_\theta(\mathbf{z}^\tau)), \tau \in \{m, \bar{m}\}, \end{aligned} \tag{4}$$

where λ^m and $\lambda^{\bar{m}}$ are trade-off hyper-parameters. Please see Appendix B for more details.

As the sampling of $\mathbf{z}_i$ is a stochastic process, we use *re-parameterization* trick (Kingma and Welling, 2014) to make it trainable: $\mathbf{z}_i = \boldsymbol{\mu}_i + \boldsymbol{\sigma}_i \odot \epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I})$, where $\odot$ refers to an element-wise product. Then, KL term $D_{KL}(\cdot)$ is computed as:

$$D_{KL}(q_\phi(\mathbf{z}|\mathbf{c})\|p_\theta(\mathbf{z})) = -\frac{1}{2}(1 + \log\boldsymbol{\sigma}^2 - \boldsymbol{\mu}^2 - \boldsymbol{\sigma}^2). \tag{5}$$

For all tokens, the CUL forces the model to be able to reconstruct the context using feature regularization specified by prior distributions of latent variables. Under the objective of VarMAE, latent vectors with similar contexts are encouraged to be smoothly organized together. After the pre-training, we leverage the Transformer encoder and f_{ϕ_μ} to fine-tune on downstream tasks.

3 Experiments

We conduct experiments on science- and finance-domain NLU tasks to evaluate our method.

Model	Science-domain					Finance-domain				
	ACL-ARC	*SciCite*	*JNLPBA*	*EBM-NLP*	Avg.	*OIR*	*MTC*	*IEE*	*PSM*	Avg.
	CLS		NER	SE		CLS		NER	TM	
RoBERTa	74.58	84.85	73.09	75.11	76.91	66.64	54.95	67.77	46.65	59.00
TAPT	68.10	86.23	72.54	74.09	75.24	65.16	53.18	68.80	49.71	59.21
DAPT	70.02	84.20	73.85	75.88	75.99	65.54	54.49	65.90	46.47	58.10
VarMAE	**76.50**	**86.32**	**74.43**	**76.01**	**78.32**	**68.77**	**56.58**	**70.15**	**53.68**	**62.30**

Table 1: Results on science- and finance-domain downstream tasks. All compared pre-trained models are fine-tuned on the task dataset. For each dataset, we run three random seeds and report the average result of the test sets. We report the micro-average F1 score for CLS and TM, entity-level F1 score for NER, and token-level F1 score for SE. Best results are highlighted in bold.

Corpus Size	Science-domain		Finance-domain	
	DAPT	VarMAE	DAPT	VarMAE
$\|\mathcal{D}\|/3$	76.77	77.82	59.56	62.04
$\|\mathcal{D}\|$	75.99	78.32	58.10	62.30

Table 2: Average results on all downstream tasks against different corpus sizes of pre-training. $|\mathcal{D}|$ is the corpus size for corresponding domain.

Masking Ratio	*Science-domain*	*Finance-domain*
5%	77.27	58.54
15%	78.32	62.30
30%	76.95	59.12

Table 3: Average results of VarMAE on all downstream tasks against different masking ratios of pre-training.

3.1 Domain Corpus and Downstream Tasks

Domain Corpus For science domain, we collect 0.6 million English abstracts (0.1B tokens) of computer science and broad biomedical fields, which are sampled from Semantic Scholar corpus (Ammar et al., 2018). For finance domain, we collect 2 million cleaned Chinese sentences (0.3B tokens) from finance-related online platforms (e.g., *Sina Finance*[3], *Weixin Official Account Platform*[4], and *Baidu Zhidao*[5]) and business scenarios[6]. The 1 million sentences in this corpus are from finance news, sales/claims cases, product introduction/clauses, and finance encyclopedia entries, while the remaining 1 million sentences are collected from the internal corpus and log data in business scenarios.

Downstream Tasks and Datasets We experiment with four categories of NLP downstream tasks: text classification (CLS), named entity recognition (NER), span extraction (SE), and text matching (TM). For science domain, we choose four public benchmark datasets: ACL-ARC (Jurgens et al., 2018) and SciCite (Cohan et al., 2019) for citation intent classification task, JNLPBA (Collier and Kim, 2004) for bio-entity recognition task, EBM-NLP (Nye et al., 2018) for PICO extraction task. For finance domain, we choose four real-world financial business datasets[6]: OIR for outbound intent recognition task, MTC for multi-label topic classification task, IEE for insurance-entity extraction task, and PSM for pairwise search match task. The details of datasets are included in Appendix C.1.

3.2 Experimental Setup

We compare VarMAE with the following baselines: **RoBERTa** (Liu et al., 2019) is an optimized BERT with a masked autoencoding objective, and is to directly fine-tune on given downstream tasks. **TAPT** (Gururangan et al., 2020) is a continual pre-training model on a task-specific corpus. **DAPT** (Gururangan et al., 2020) is a continual pre-training model on a domain-specific corpus.

Experiments are conducted under PyTorch[7] and using 2/1 NVIDIA Tesla V100 GPUs with 16GB memory for pre-training/fine-tuning. During pre-training, we use *roberta-base*[8] and *chinese-roberta-wwm-ext*[8] to initialize the model for science (English) and finance domains (Chinese), respectively. During the pre-training of VarMAE, we freeze the embedding layer and all layers of Transformer encoder to avoid catastrophic forgetting (French, 1993; Arumae et al., 2020) of previously general learned knowledge. And then we optimize other network parameters (e.g., the LM Head and CUL module) by using Adam optimizer (Kingma and Ba, 2015) with the learning rate of $5e^{-5}$. The number of epochs is set to 3. We use gradient accumulation step of 50 to achieve the large batch sizes (i.e., the batch size is 3200). The trade-off co-

[3]https://finance.sina.com.cn/
[4]https://mp.weixin.qq.com/
[5]https://zhidao.baidu.com/
[6]https://life.pingan.com/
[7]https://pytorch.org/
[8]https://huggingface.co/

No.	Example	Gold	Pred. (RoBERTa)	Pred. (DAPT)	Pred. (VarMAE)
1	*Can forearm superficial injury insure accidental injury?* (前臂浅表损伤是否投保意外保险？)	*Accident* (意外); *Disease underwriting* (疾病核保)	*Disease underwriting*	*Accident*	*Accident*; *Disease underwriting*
2	*Medical demands inspire quality care.* (医疗需求激发品质养老。)	*Pension* (养老); *Risk education* (风险教育)	*Pension*	*Pension*	*Pension*; *Risk education*
3	*How does high incidence cancer protection calculate the risk insurance?* (高发癌症保障计划如何计算风险保额？)	*Critical illness* (重疾); *Insurance rules* (投保规则)	*Insurance rules*	*Insurance rules*	*Critical illness*; *Insurance rules*
4	*What are the features of ABC Comprehensive Care Program?* (ABC全面呵护计划特色包括什么内容？)	*Product introduction* (产品介绍); *Critical illness* (重疾)	*Product introduction*	*Product introduction*	*Product introduction*

Table 4: Case studies in the multi-label topic classification (MTC) task of financial business scenarios. The table shows four examples of spoken dialogues in the test set, their gold labels and predictions by three methods (RoBERTa, DAPT and VarMAE). We translate original Chinese to English version for readers.

efficient λ is set to 10 for both domains selected from $\{1, 10, 100\}$. For fine-tuning on downstream tasks, most hyperparameters are the same as in pre-training, except for the following settings due to the limited computation. The batch size is set to 128 for OIR, and 32 for other tasks. The maximum sequence length is set to 64 for OIR, and 128 for other tasks. The number of epochs is set to 10. More details are listed in Appendix C.2.

3.3 Results and Analysis

Table 1 shows the results on science- and finance-domain downstream tasks. In terms of the average result, VarMAE yields 1.41% and 3.09% absolute performance improvements over the best-compared model on science and finance domains, respectively. It shows the superiority of domain-adaptive pre-training with context uncertainty learning. DAPT and TAPT obtain inferior results. It indicates that the small domain corpus limits the continual pre-training due to the *distribution shift*.

We report the average results on all tasks against different corpus sizes of pre-training in Table 2 (see Appendix D.1 for details). VarMAE consistently achieves better performance than DAPT even though a third of the corpus is used. When using full corpus, DPAT's performance decreases but VarMAE's performance increases, which proves our method has a promising ability to adapt to the target domain with a limited corpus.

Table 3 shows the average results of VarMAE on all tasks against different masking ratios of pre-training (see Appendix D.2 for details). Under the default masking strategies[9], the best masking rate is 15%, which is the same as BERT and RoBERTa.

[9] 80% for replacing the target token with [MASK] symbol, 10% for keeping the target token as is, and 10% for replacing the target token with another random token.

3.4 Case Study

As shown in Table 4, we randomly choose several samples from the test set in the multi-label topic classification (MTC) task.

For the first case, RoBERTa and DAPT each predict one label correctly. It shows that both general and domain language knowledge have a certain effect on the domain-specific task. However, none of them identify all the tags completely. This phenomenon reflects that the general or limited continual PLM is not sufficient for the domain-specific task. For the second and third cases, these two comparison methods cannot classify the topic label *Risk education* and *Critical illness*, respectively. It indicated that they perform an isolated point estimation and have a relatively poor context representation. Unlike other methods, our VarMAE can encode the token's context into a smooth latent distribution and produce diverse and well-formed contextual representations. As expected, VarMAE predicts the first three examples correctly with limited resources.

For the last case, all methods fail to predict *Critical illness*. We notice that *ABC Comprehensive Care Program* is a product name related to critical illness insurance. Classifying it properly may require some domain-specific structured knowledge.

4 Conclusion

We propose a novel Transformer-based language model named VarMAE for domain-adaptive language understanding with limited resources. A new CUL module is designed to produce a diverse and well-formed context representation. Experiments on science- and finance-domain tasks demonstrate that VarMAE can be efficiently adapted to new domains using a limited corpus. Hope that VarMAE can guide future foundational work in this area.

Limitations

All experiments are conducted on a small pre-training corpus due to the limitation of computational resources. The performance of VarMAE pre-training on a larger corpus needs to be further studied. Besides, VarMAE cannot be directly adapted to downstream natural language generation tasks since our model does not contain a decoder for the generation. This will be left as future work.

Acknowledgements

This research is supported by Ping An Life Insurance. We thank the reviewers for their insightful and constructive comments.

References

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew E. Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. 2018. Construction of the literature graph in semantic scholar. In *NAACL-HLT (3)*, pages 84–91. Association for Computational Linguistics.

Kristjan Arumae, Qing Sun, and Parminder Bhatia. 2020. An empirical investigation towards efficient multi-domain language model pre-training. In *EMNLP (1)*, pages 4854–4864. Association for Computational Linguistics.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *EMNLP/IJCNLP (1)*, pages 3613–3618. Association for Computational Linguistics.

Dimitri P Bertsekas. 1997. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan R. Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*. European Language Resources Association.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Re-train or train from scratch? comparing pre-training strategies of BERT in the medical domain. In *LREC*, pages 2626–2633. European Language Resources Association.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: "preparing the muppets for court'". In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 2898–2904. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *ICLR*. OpenReview.net.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *NAACL-HLT (1)*, pages 3586–3596. Association for Computational Linguistics.

Nigel Collier and Jin-Dong Kim. 2004. Introduction to the bio-entity recognition task at JNLPBA. In *NLPBA/BioNLP*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese BERT. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3504–3514.

Armen Der Kiureghian and Ove Ditlevsen. 2009. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Shizhe Diao, Jiaxin Bai, Yan Song, Tong Zhang, and Yonggang Wang. 2020. ZEN: pre-training chinese text encoder enhanced by n-gram representations. In

EMNLP (Findings), volume EMNLP 2020 of *Findings of ACL*, pages 4729–4740. Association for Computational Linguistics.

Robert M. French. 1993. Catastrophic interference in connectionist networks: Can it be predicted, can it be prevented? In *NIPS*, pages 1176–1177. Morgan Kaufmann.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Heal.*, 3(1):2:1–2:23.

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *ACL*, pages 8342–8360. Association for Computational Linguistics.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR (Poster)*. OpenReview.net.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *ACL (1)*, pages 328–339. Association for Computational Linguistics.

Dou Hu, Zhou Mengyuan, Xiyang Du, Mengfei Yuan, Jin Zhi, Lianxin Jiang, Mo Yang, and Xiaofeng Shi. 2022. Pali-nlp at semeval-2022 task 4: Discriminative fine-tuning of transformers for patronizing and condescending language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 335–343.

Dou Hu and Lingwei Wei. 2020. SLK-NER: exploiting second-order lexicon knowledge for chinese NER. In *SEKE*, pages 413–417. KSI Research Inc.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *CoRR*, abs/1904.05342.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 448–456. JMLR.org.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Trans. Assoc. Comput. Linguistics*, 8:64–77.

David Jurgens, Srijan Kumar, Raine Hoover, Daniel A. McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Trans. Assoc. Comput. Linguistics*, 6:391–406.

William Karush. 2014. Minima of functions of several variables with inequalities as side conditions. In *Traces and Emergence of Nonlinear Programming*, pages 217–245. Springer.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. In *ISMB (Supplement of Bioinformatics)*, pages 180–182.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *ICLR*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880. Association for Computational Linguistics.

Bohan Li, Junxian He, Graham Neubig, Taylor Berg-Kirkpatrick, and Yiming Yang. 2019. A surprisingly effective fix for deep latent variable modeling of text. In *EMNLP/IJCNLP (1)*, pages 3601–3612. Association for Computational Linguistics.

Chunyuan Li, Xiang Gao, Yuan Li, Baolin Peng, Xiujun Li, Yizhe Zhang, and Jianfeng Gao. 2020. Optimus: Organizing sentences via pre-trained modeling of a latent space. In *EMNLP (1)*, pages 4678–4699. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *ICLR (Workshop Poster)*.

Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119.

Benjamin E. Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain James Marshall, Ani Nenkova, and Byron C. Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *ACL (1)*, pages 197–207. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. ACL.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*, pages 2227–2237. Association for Computational Linguistics.

Yujia Qin, Yankai Lin, Ryuichi Takanobu, Zhiyuan Liu, Peng Li, Heng Ji, Minlie Huang, Maosong Sun, and Jie Zhou. 2021. ERICA: improving entity and relation understanding for pre-trained language models via contrastive learning. In *ACL/IJCNLP (1)*, pages 3350–3363. Association for Computational Linguistics.

Yujia Qin, Jiajie Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. ELLE: efficient lifelong pre-training for emerging data. In *ACL (Findings)*, pages 2789–2810. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223.

Wen Tai, H. T. Kung, Xin Dong, Marcus Z. Comiter, and Chang-Fu Kuo. 2020. exbert: Extending pre-trained models with domain-specific vocabulary under constrained training resources. In *EMNLP (Findings)*, volume EMNLP 2020 of *Findings of ACL*, pages 1433–1439. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*, pages 353–355. Association for Computational Linguistics.

Wei Wang, Bin Bi, Ming Yan, Chen Wu, Jiangnan Xia, Zuyi Bao, Liwei Peng, and Luo Si. 2020. Structbert: Incorporating language structures into pre-training for deep language understanding. In *ICLR*. OpenReview.net.

Lingwei Wei, Dou Hu, Wei Zhou, Xuehai Tang, Xiaodan Zhang, Xin Wang, Jizhong Han, and Songlin Hu. 2020. Hierarchical interaction networks with rethinking mechanism for document-level sentiment analysis. In *ECML/PKDD (3)*, volume 12459 of *Lecture Notes in Computer Science*, pages 633–649. Springer.

Tongtong Wu, Massimo Caccia, Zhuang Li, Yuan-Fang Li, Guilin Qi, and Gholamreza Haffari. 2022. Pretrained language model in continual learning: A comparative study. In *ICLR*. OpenReview.net.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A chinese language understanding evaluation benchmark. In *COLING*, pages 4762–4772. International Committee on Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*, pages 5754–5764.

Xingcheng Yao, Yanan Zheng, Xiaocong Yang, and Zhilin Yang. 2022. Nlp from scratch without large-scale pretraining: A simple and efficient framework. In *International Conference on Machine Learning*, pages 25438–25451. PMLR.

Yunzhi Yao, Shaohan Huang, Wenhui Wang, Li Dong, and Furu Wei. 2021. Adapt-and-distill: Developing small, fast and effective pretrained language models for domains. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 460–470. Association for Computational Linguistics.

Rong Zhang, Revanth Gangi Reddy, Md. Arafat Sultan, Vittorio Castelli, Anthony Ferritto, Radu Florian, Efsun Sarioglu Kayi, Salim Roukos, Avirup Sil, and Todd Ward. 2020. Multi-stage pre-training for low-resource domain adaptation. In *EMNLP (1)*, pages 5461–5468. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. In *ACL (1)*, pages 1441–1451. Association for Computational Linguistics.

Qile Zhu, Wei Bi, Xiaojiang Liu, Xiyao Ma, Xiaolin Li, and Dapeng Wu. 2020. A batch normalized inference network keeps the KL vanishing away. In *ACL*, pages 2636–2649. Association for Computational Linguistics.

Appendix Overview

In this supplementary material, we provide: (i) the related work, (ii) objective derivation of the proposed VarMAE, (iii) detailed description of experimental setups, (iv) detailed results, and (v) our contribution highlights.

A Related Work

A.1 General PLMs

Traditional works (Mikolov et al., 2013b; Pennington et al., 2014) represent the word as a single vector representation, which cannot disambiguate the word senses based on the surrounding context. Recently, unsupervised pre-training on large-scale corpora significantly improves performance, either for Natural Language Understanding (NLU) (Peters et al., 2018; Devlin et al., 2019; Cui et al., 2021) or for Natural Language Generation (NLG) (Raffel et al., 2020; Brown et al., 2020; Lewis et al., 2020). Following this trend, considerable progress (Liu et al., 2019; Yang et al., 2019; Clark et al., 2020; Joshi et al., 2020; Wang et al., 2020; Diao et al., 2020) has been made to boost the performance via improving the model architectures or exploring novel pre-training tasks. Some works (Sun et al., 2019; Zhang et al., 2019; Qin et al., 2021) enhance the model by integrating structured knowledge from external knowledge graphs.

Due to the flexibility of natural language, one word may have different meanings under different domains. These methods underperform when migrated to specialized domains. Moreover, simple fine-tuning (Howard and Ruder, 2018; Hu and Wei, 2020; Wei et al., 2020; Hu et al., 2022) of PLMs is also not sufficient for domain-specific applications.

A.2 Domain-adaptive PLMs

Recent works perform pre-training from scratch (Gu et al., 2022; Yao et al., 2022) or continual pre-training (Alsentzer et al., 2019; Huang et al., 2019; Lee et al., 2020; Gururangan et al., 2020; Wu et al., 2022; Qin et al., 2022) on domain-specific corpora.

Remarkably, Beltagy et al. (2019); Chalkidis et al. (2020) explore different strategies to adapt to new domains, including pre-training from scratch and further pre-training. Boukkouri et al. (2022) find that both of them perform at a similar level when pre-training on a specialized corpus, but the former requires more resources. Yao et al. (2022) jointly optimize the task and language modeling objective from scratch. Zhang et al. (2020); Tai et al. (2020); Yao et al. (2021) extend the vocabulary of the LM with domain-specific terms for further gains. Gururangan et al. (2020) show that domain- and task-adaptive pre-training methods can offer gains in specific domains. Qin et al. (2022) present an efficient lifelong pre-training method for emerging domain data.

In most specific domains, collecting large-scale corpora is usually inaccessible. The limited data makes pre-training from scratch infeasible and restricts the performance of continual pre-training. Towards this issue, we investigate domain-adaptive language understanding with a limited target corpus, and propose a novel language modeling method named VarMAE. The method performs a context uncertainty learning module to produce diverse and well-formed contextual representations, and can be efficiently adapted to new domains with limited resources.

B Derivation of Objective Function

Here, we take the objective for masked tokens as the example to give derivations of the loss function. The objective for unmasked tokens is similar. For simplifying description, we omit the superscripts that use to distinguish masked tokens from unmasked tokens. To learn a smooth space of masked tokens where latent representations of similar contexts are close to each other and vice versa, the objective function is:

$$\begin{aligned}\max_{\phi,\theta} \mathbb{E}_{x\sim\mathbf{D}}[\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{c})}[\log p_\theta(\mathbf{x}|\mathbf{z})]],\\ \text{s.t. } D_{KL}(q_\phi(\mathbf{z}|\mathbf{c})\|p_\theta(\mathbf{z}))<\delta,\end{aligned} \tag{6}$$

where $\delta > 0$ is a constraint, and $q_\phi(\mathbf{z}|\mathbf{c})$ is the variational approximate posterior of the true posterior $p_\theta(\mathbf{z}|x)$ (see Section 2.2). $D_{KL}(\cdot)$ denotes the KL-divergence term, which serves as the regularization that forces the prior distribution p_θ to approach the approximated posterior q_ϕ.

In order to encourage this disentangling property in the inferred (Higgins et al., 2017), we introduce a constraint δ over $q_\phi(\mathbf{z}|\mathbf{c})$ by matching it to a prior $p_\theta(\mathbf{z})$. The objective can be computed as a Lagrangian under the KKT condition (Bertsekas, 1997; Karush, 2014). The above optimization problem with only one inequality constraint is equivalent to maximizing the following equation,

$$\begin{aligned}\mathcal{F}(\theta,\phi,\lambda;\mathbf{c},\mathbf{z}) = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{c})}[\log p_\theta(\mathbf{x}|\mathbf{z})]\\ -\lambda(D_{KL}(q_\phi(\mathbf{z}|\mathbf{c})\|p_\theta(\mathbf{z}))-\delta),\end{aligned} \tag{7}$$

	Dataset Name	Task Name	Train	Dev	Test	# Entities	Avg/Min/Max	Class	Source
Science	ACL-ARC	Citation Intent Classification	1,688	114	139	-	42/4/224	6	NLP field
	SciCite	Citation Intent Classification	7,320	916	1,861	-	34/7/228	3	Multiple scientific fields
	JNLPBA	Bio-entity Recognition	16,807	1,739	3,856	59,963	27/2/204	5	Biomedical field
	EBM-NLP	PICO Extraction	27,879	7,049	2,064	77,360	37/1/278	3	Clinical medicine field
Finance	OIR	Outbound Intent Recognition	36,885	9,195	3,251	-	16/2/69	34	F1, F2
	MTC	Multi-label Topic Classification	66,670	2,994	4,606	-	15/2/203	39	F1, F2, F3, F4
	IEE	Insurance-entity Extraction	19,136	4,784	19,206	13,128	21/1/388	2	F1, F2
	PSM	Pairwise Search Match	11,812	1,476	1,477	-	7/2/100; 14/1/134	4	F1, F2

Table 5: Dataset statistics of science- and finance-domain downstream tasks. Avg, Min, and Max indicate the average, minimum, and maximum length of sentences, respectively. "Class" refers to the number of classes. F1, F2, F3 and F4 mean the insurance, sickness, job and legal fields, respectively.

Hyperparameter	**Assignment**
Number of Epoch	3
Trade-off Weight λ	10
Number of Layers	12
Hidden size	768
FFN inner hidden size	3072
Attention heads	12
Attention head size	64
Dropout	0.1
Attention Dropout	0.1
Peak Learning Rate	$5e^{-5}$
Maximum Length	128
Batch Size	64
Gradient Accumulation Steps	50
Optimization Steps	{504, 1830}
Weight Decay	0.0
Adam ϵ	$1e^{-6}$
Adam β_1	0.9
Adam β_2	0.98

Table 6: Hyperparameters for pre-training on a domain-specific corpus for each domain. The optimization steps are 504 and 1830 for science- and finance-domain, respectively.

where the KKT multiplier λ is the regularization coefficient that constrains the capacity of the latent information channel $\mathbf{z}$ and puts implicit independence pressure on the learnt posterior due to the isotropic nature of the Gaussian prior $p_\theta(\mathbf{z})$. Since $\delta, \lambda > 0$, the function is further defined as,

$$\begin{aligned} \mathcal{F}(\theta, \phi, \lambda; \mathbf{c}, \mathbf{z}) \geq \mathcal{L}(\theta, \phi; \mathbf{c}, \mathbf{z}, \lambda) \\ = \mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{c})}[\log p_\theta(\mathbf{x}|\mathbf{z})] \\ - \lambda D_{KL}(q_\phi(\mathbf{z}|\mathbf{c}) \| p_\theta(\mathbf{z})), \end{aligned} \quad (8)$$

where the multiplier λ can be considered as a hyperparameter. λ not only encourages more efficient latent encoding but also creates a trade-off between context reconstruction quality and the extent of disentanglement. We train the model by minimizing the loss $\mathcal{L}$ to push up its evidence lower bound.

Hyperparameter	**Assignment**
Number of Epoch	10
Maximum Length	{64, 128}
Batch Size	{32, 128}
Learning Rate	$5e^{-5}$
Dropout	0.1
Weight Decay	0.0
Warmup ratio	0.06

Table 7: Hyperparameters for fine-tuning on science- and finance-domain downstream tasks. The maximum sequence length is set to 64 for OIR, and is set to 128 for other tasks. The batch size is set to 128 for OIR, and is set to 32 for other tasks.

C Detailed Experimental Setup

C.1 Datasets of Downstream Tasks

The statistics of datasets and their corresponding tasks are reported in Table 5.

Science Domain We choose four public benchmark datasets from the science domain.

ACL-ARC (Jurgens et al., 2018) is a dataset of citation intents based on a sample of papers from the ACL Anthology Reference Corpus (Bird et al., 2008) in the NLP field.

SciCite (Cohan et al., 2019) is a dataset of citation intents. It provides coarse-grained categories and covers a variety of scientific domains.

JNLPBA (Collier and Kim, 2004) is a named entity dataset in the biomedical field and is derived from five superclasses in the GENIA corpus (Kim et al., 2003).

EBM-NLP (Nye et al., 2018) annotates PICO (Participants, Interventions, Comparisons and Outcomes) spans in clinical trial abstracts. The corresponding PICO Extraction task aims to identify the spans in clinical trial abstracts that describe the respective PICO elements.

Finance Domain We choose four real-world business datasets[6] from the financial domain.

Corpus Size	Model	*Science-domain*					*Finance-domain*				
		ACL-ARC	*SciCite*	*JNLPBA*	*EBM-NLP*	Avg.	*OIR*	*MTC*	*IEE*	*PSM*	Avg.
		CLS		NER	SE		CLS		NER	TM	
$\|\mathcal{D}\|/3$	DAPT	72.42	85.92	73.38	75.35	76.77	72.65	47.09	66.13	52.38	59.56
$\|\mathcal{D}\|/3$	VarMAE	76.98	84.67	74.73	74.91	77.82	70.50	53.93	67.72	56.02	62.04
$\|\mathcal{D}\|$	DAPT	70.02	84.20	73.85	75.88	75.99	65.54	54.49	65.90	46.47	58.10
$\|\mathcal{D}\|$	VarMAE	76.50	86.32	74.43	76.01	**78.32**	68.77	56.58	70.15	53.68	**62.30**

Table 8: Results of DAPT and VarMAE on all downstream tasks against different corpus sizes of pre-training. $|\mathcal{D}|$ is the corpus size. For each dataset, we run three random seeds and report the average result of the test sets. We report the micro-average F1 score for CLS and TM, entity-level F1 score for NER, and token-level F1 score for SE.

Masking Ratio	Model	*Science-domain*					*Finance-domain*				
		ACL-ARC	*SciCite*	*JNLPBA*	*EBM-NLP*	Avg.	*OIR*	*MTC*	*IEE*	*PSM*	Avg.
		CLS		NER	SE		CLS		NER	TM	
5%	VarMAE	76.02	85.12	73.86	74.09	77.27	67.80	46.33	66.72	53.32	58.54
15%	VarMAE	76.50	86.32	74.43	76.01	**78.32**	68.77	56.58	70.15	53.68	**62.30**
30%	VarMAE	73.62	85.69	73.75	74.73	76.95	70.57	45.68	65.00	55.23	59.12

Table 9: Results of VarMAE on all downstream tasks against different masking ratios of pre-training. For each dataset, we run three random seeds and report the average result of the test sets. We report the micro-average F1 score for CLS and TM, entity-level F1 score for NER, and token-level F1 score for SE.

OIR is a dataset of the outbound intent recognition task. It aims to identify the intent of customer response in the outbound call scenario.

MTC is a dataset of the multi-label topic classification task. It aims to identify the topics of the spoken dialogue.

PSM is a dataset of the pairwise search matching task. It aims to identify the semantic similarity of a sentence pair in the search scenario.

IEE is a dataset of the Insurance-entity extraction task. Its goal is to locate named entities mentioned in the input sentence.

For OIR and MTC, we use an ASR (automatic speech recognition) tool to convert acoustic signals into textual sequences in the pre-processing phase.

C.2 Implementation Details

C.2.1 Pre-training Hyperparameters

Table 6 describes the hyperparameters for pre-training on a domain-specific corpus.

C.2.2 Fine-tuning Hyperparameters

Table 7 reports the fine-tuning hyperparameters for downstream tasks.

D Detailed Results

In this part, we provide detailed results on science- and finance-domain downstream tasks.

D.1 Results Against Different Corpus Sizes

The detailed results of DAPT and VarMAE on all downstream tasks against different corpus sizes of pre-training are reported in Table 8.

D.2 Results Against Different Masking Ratios

The detailed results of VarMAE on all downstream tasks against different masking ratios of pre-training are reported in Table 9.

E Contribution and Future Work

The main contributions of this work are as follows: **1)** We present a domain-adaptive language modeling method named VarMAE based on the combination of variational autoencoders and masked autoencoders. **2)** We design a context uncertainty learning module to model the point-estimate context of each token into a smooth latent distribution. The module can produce diverse and well-formed contextual representations. **3)** Extensive experiments on science- and finance-domain NLU tasks demonstrate that VarMAE can be efficiently adapted to new domains with limited resources.

For future works, we will build domain-specific structured knowledge to further assist language understanding, and apply our method for domain-adaptive language generation.