

Adversarial Training Methods for Cross-Domain Question Answering

Stanford CS224N Default Project (Robust QA track)

Gino Tesei
Stanford University
gtesei@stanford.edu

Abstract

Despite recent work demonstrated considerable advantages on many NLP tasks by pre-training on a large corpus of text and then fine-tuning on a specific task, this approach can hardly scale as it requires task-specific fine-tuning data for any new task and, unfortunately, this is also the case when it comes to adapt one model trained for a specific task and dataset to a new dataset. To address this problem, domain adaptation techniques aim to adapt models trained for a task on in-domain datasets to a target domain by using efficiently samples from the latter. On the contrary, domain generalization techniques aim to incentivate the model to learn domain-invariant features directly from in-domain data to generalize the model for any out-of-domain dataset, pushing to learn task-relevant features and preventing overfitting on in-domain data. We like to compare this approach the way humans learn a task, as they can generally perform the same task on different domains from only a few examples. However, domain generalization is often performed by augmenting in-domain data by applying semantic-preserving transformations to challenge the model during training, leveraging some kind of rules or domain knowledge. Contrarily, in this project our goal is to explore domain generalization techniques applied to question answering based on adversarial training without leveraging any set of rules or domain knowledge, but using adversarial terms to make more robust the regular loss with or without adopting task-agnostic critic networks. Such extremely general methodology does not suffer from the limitations of synonym replacement approaches and can be applied to other NLP tasks. Our best variant combines two different and complementary approaches of adversarial training on a DistilBERT baseline, achieving $>3\%$ F1-score improvement over the regular fine-tuning process, outperforming several other adversarial and energy-based approaches.

1 Introduction

Even though many deep learning models surpass human-level performance on tasks like question answering when evaluated on in-domain test-sets, they might perform relatively poorly on out-of-domain datasets. To solve this problem, domain adaptation techniques aim to use efficiently samples of the target domain, but still its effectiveness depends, although not linearly, on the number of out-of-domain samples available [HDB19].

On the contrary, domain generalization techniques aim to generalize models by stimulating them to learn domain-invariant features directly on in-domain data through robust training methods such as adversarial training, which is the process of training a model to correctly perform the task on both unmodified examples and adversarial examples, improving both robustness to adversarial examples and generalization performance for original examples. While in computer vision adversarial perturbations can be applied directly on input images [GSS15], NLP typically adopts high-dimensional input one-hot vectors, which do not admit infinitesimal perturbations. This limitation motivated rule-based approaches [RSG18] or synonym replacement strategies [ASE⁺18a] to generate adversarial

examples, which typically suffer from the intrinsic limitation of making the model more robust only against pre-defined transformations.

In this work, we focus on domain generalization techniques applied to question answering without leveraging any prior linguistic or domain set of rules, defining adversarial perturbations on continuous word and positional embeddings instead of discrete word inputs, and experimenting adversarial terms with and without adopting task-agnostic critic networks. Based on these assumptions, we extend a DistilBERT [SDCW20] baseline with the adversarial training approach described in [LKP19], where the model is incentivated to learn domain-invariant features from the embedding representations of tokens like [CLS] or [SEP] against a discriminator network trained to correctly classify the sample domain. We provide empirical evidence that such model suffers from vanishing gradient problems and, to fix such issue, we experiment Wasserstein GAN with gradient penalty (WGAN-GP) [ACB17], which is designed to address this problem in a different domain. We found this approach moderately successful achieving ~ 0.86 F1-score improvement over the regular fine-tuning process, identifying the main limitation to overcome in future improvements.

Moreover, we explore the adversarial training method to regularize supervised learning algorithms exposed in [MDG17] for text classification on LSTMs. This approach generates small adversarial perturbations of the embedding representations designed to significantly increase the loss and pretending the model handle them correctly. In addition, we experiment the related virtual adversarial training method [MiMK⁺16, MDG17], which extends such idea to the semi-supervised regime and unlabeled examples. This is done by regularizing the model so that given an example, the model will produce the same output distribution as it produces on an adversarial perturbation of that example. Our variant based on the supervised approach achieves ~ 1.21 F1-score improvement over the regular fine-tuning process, proving its effectiveness. Combined together with the previous adversarial training method through linear ensembling we achieve $>3\%$ F1-score improvement over the regular fine-tuning process, proving both their effectiveness and complementarity.

Lastly, we explore energy-based models (EBMs) [LCH⁺06, Hin02] as a means for robust modeling. They require the definition of an energy function defined over the whole input sequence. Training EBMs aims at shaping the energy function such that regions of high density of training data points have lower energy than elsewhere. Specifically, shaping this energy function at training time might involve to identify the most offending incorrect answer (i.e. the answer that has the lowest energy among all the incorrect answers) and minimizing the difference between the energies of the correct answer and the most offending incorrect answer, penalized linearly when larger than a margin. We experimented such loss identifying a major limitation and modifying it with a contrastive term able to achieve a significant improvement.

2 Related Work

Generating adversarial samples at training time traditionally involved the application of semantic-preserving perturbations using a set of rules [RSG18], computing the nearest neighbors of selected words according to the distance in the GloVe embedding space [ASE⁺18b], or various synonym replacement strategies [ASE⁺18a], and such approaches have been applied on tasks such as cross-domain universal dependency parsing [SMNM17], acronym identification [ZLZ⁺21] and question answering [RSG18, LKP19]. Synonym replacement strategies and semantically equivalent adversarial rules (e.g. *What color* $\rightarrow$ *Which color*) are effective for specific transformations and typically non redundant, but they are based on a predefined set of rules to replace specific words in sentences. On the other hand, in question answering we can challenge the model in many ways, including adding irrelevant details to the paragraph, and it is difficult to capture them all in predefined schemes. This limitation holds even stronger for approaches based on word replacement from word embeddings or WordNet [Mi95], where the semantic equivalence is not always guaranteed. Unlike those methods, we don't leverage any linguistic or domain set of rules.

Adversarial training was originally proposed [GSS15] for Generative Adversarial Networks (GANs) applied to image generation. GANs have been experimented also in NLP, for example applied to text generation [YZWY17]. The same author co-developed the adversarial training technique described in [MDG17] for text classification on a LSTM-based architecture [GS05], which has been applied to acronym identification [ZLZ⁺21], neural semantic parsers [HLQP21], neural machine translation [CJME20]. The related semi-supervised extension of such approach [MDG17] is based on the idea

of local distributional smoothness expressed in [MiMK⁺16]. Our extension to question answering of such approaches on transformers architectures removes some ingredients of the original recipe, e.g. we removed the embedding normalization after observing interfering with positional embeddings.

EBMs [LCH⁺06, Hin02] in NLP have been applied to several tasks, including text generation [DBO⁺20], text ranking [YF16], natural language understanding [HMXHA21].

3 Approach

Baseline. The baseline system finetunes DistilBERT [SDCW20] on all the training data. We report the performance of both DistilBERT finetuned only on train in-domain datasets (baseline) vs. finetuned on all the train datasets (baseline_ft) to estimate the finetuning gain.

VAR-1. The first variation to our DistilBERT baseline is the one described in [LKP19]. Let consider K in-domain datasets $D = \{D_i\}_{i=1}^K$, each one consisting of triplets of passage c , question q , and answer y , i.e. $D_i = \left\{c_i^{(j)}, q_i^{(j)}, y_i^{(j)}\right\}_{j=1}^{N_i}$ where $y_i^{(j)} = \left(s_i^{(j)}, e_i^{(j)}\right) \in \mathbb{N}^2$. Each D_i can be considered a sample of size N_i from the underlying domain population p_i . Let consider a sampling procedure that first samples from a categorical variable l having uniform distribution over K classes, i.e. $l \sim U_k$, and then generates a data point by sampling from p_l . We denote the related distribution as p , i.e. $(l, c, q, y) \sim p$. Hence, the QA model is trained to minimize

$$\mathcal{L}_{QA} = \mathbb{E}_{(l, c, q, y) \sim p} [-\log(P_{start}(y_s | c, q; \theta)) - \log(P_{end}(y_e | c, q; \theta))] \quad (1)$$

, where P_{start}, P_{end} are the softmax distributions for the start/end locations. To regularize the model with adversarial training, a discriminator D_{adv} is trained to classify the joint embedding of question and passage from the QA model, denoted as $E(c, q; \theta) \in \mathbb{R}^d$, into the given K domains, minimizing the loss

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}_{(l, c, q, y) \sim p} [-\log(D_{adv}(E(c, q; \theta); \phi)_l)] \quad (2)$$

, where we denote with $D_{adv}(E(c, q; \theta); \phi)_l$ the l -th component of the (softmax) distribution $D_{adv}(E(c, q; \theta); \phi)$. Adversarially, the QA model tries to maximize the entropy of $D_{adv}(E(c, q; \theta); \phi)_l$, or equivalently tries to minimize the Kullback-Leibler (KL) divergence between the uniform distribution over K classes and the discriminator’s prediction (and not just the l -th component), i.e.

$$\mathcal{L}_{ADV} = KL[U_k(\cdot) \parallel D_{adv}(\cdot)].$$

The full objective of the QA model is

$$\mathcal{L}(QA, D_{adv}; \lambda) = \mathcal{L}_{QA} + \lambda \mathcal{L}_{ADV} \quad (3)$$

where λ controls the relative importance of the two objective terms. In our first sub-variation (var_1_base¹), we adopt a discriminator based on a multilayer perceptron updating both the QA model and the discriminator at each iteration sampling with equal probability from the in-domain train sets as in [LKP19]. During experimentation, we noticed KL vanishing problems on $\mathcal{L}_{ADV}$ and we adopted KL annealing [FLL⁺19] to fix it (var_1_anneal¹). We noticed that along with [CLS] token also the representation of [SEP] tokens can be extracted to feed the discriminator (var_1_anneal_sep_embed²). Also, we pushed to the extreme such approach, extracting not only the [SEP] and [CLS] tokens but also k random token representations from DistilBERT (var_1_random_token²).

VAR-2. The second variation (var_2_base²) adopts

$$\mathcal{L}_{ADV} = -\mathbb{E}_{(l, c, q, y) \sim p} [\log(1 - D_{adv}(E(c, q; \theta); \phi)_l)]$$

in equation 3, which is the experimental setting of the original GAN paper [GPAM⁺14]. As last layer of the discriminator a sigmoid is adopted, i.e. $D_{adv}(E(c, q; \theta); \phi) = \sigma(h_\phi(E(c, q; \theta)))$, where h_ϕ

¹Adapted code from the repository of [LKP19] to our baseline and datasets: <https://github.com/seanie12/mrqa>

²Original part of our approach.

are the domain logits. At the begin of the training $D_{adv}(E(c, q; \theta); \phi)_l \approx 0$, and it can be shown that the derivative of $\mathcal{L}_{ADV}$ with respect to θ is approximately 0, i.e.

$$\begin{aligned}\frac{\partial \mathcal{L}_{ADV}}{\partial \theta} &= -\mathbb{E}_{(l, c, q, y) \sim p} \left[\frac{\sigma'(h_\phi(E(c, q; \theta)))_l}{1 - \sigma(h_\phi(E(c, q; \theta)))_l} \frac{\partial h_\phi(E(c, q; \theta))}{\partial \theta} \right] \\ &= -\mathbb{E}_{(l, c, q, y) \sim p} \left[D_{adv}(E(c, q; \theta); \phi)_l \cdot \frac{\partial h_\phi(E(c, q; \theta))}{\partial \theta} \right]\end{aligned}$$

Hence, this variation was designed to be a confirmation of this problem, which was indeed observed.

VAR-3. As we observed vanishing gradient issues, we experimented Wasserstein GAN [ACB17] with gradient penalty (var_3_base²), i.e.

$$\begin{aligned}\mathcal{L}_D &= -\mathbb{E}_{(l, c, q, y) \sim p} [D_{adv}(E(c, q; \theta); \phi)_l] + \lambda_P \mathbb{E}_{(l, c, q, y) \sim p} \left[(\|\nabla_\phi D_{adv}(E(c, q; \theta); \phi)\|_2 - 1)^2 \right] \\ \mathcal{L}_{ADV} &= \mathbb{E}_{(l, c, q, y) \sim p} [D_{adv}(E(c, q; \theta); \phi)_l]\end{aligned}$$

where λ_P controls the straight of the penalty (we adopted $\lambda_P = 10$, which usually works). Notice that in the gradient penalty term we did not use the the random interpolation distribution, as here we don't have a generator.

VAR-4. We applied the technique for generating adversarial samples in the embedding layer described in [MDG17]. We extend this approach from text-classification on LSTMs to question answering on DistilBERT, adding the following term to $\mathcal{L}_{QA}$

$$\mathcal{L}_{ADV} = -\log(P_{start}(y_s | E(c, q; \theta) + r_{adv}; \theta)) - \log(P_{end}(y_e | E(c, q; \theta) + r_{adv}; \theta))$$

where $E(c, q; \theta) \in \mathbb{R}^d$ is the output of the embedding layer related to the passage c and the question q , while

$$r_{adv} = \arg \min_{r, \|r\|_2 \leq \epsilon} \log(P_{start}(y_s | E(c, q; \theta) + r; \theta)) + \log(P_{end}(y_e | E(c, q; \theta) + r; \theta))$$

approximated as

$$\begin{aligned}r_{adv} &= -\epsilon g / \|g\|_2 \\ \text{where } g &= \nabla_x \log(P_{start}(y_s | E(c, q; \theta); \theta)) + \log(P_{end}(y_e | E(c, q; \theta); \theta))\end{aligned}$$

For each iteration, var_4_adv_loss² computes and back-propagates two different losses, i.e. one of regular question answering and another for $\mathcal{L}_{ADV}$, while var_4_base² add $\mathcal{L}_{ADV}$ to $\mathcal{L}_{QA}$. In [MDG17] it is observed the model could trivially learn to make the perturbations insignificant by learning embeddings with very large norm and, to prevent this pathological solution, var_4_base_norm_embed² is designed to replace the embeddings with normalized embeddings.

VAR-5. A further variation is to apply the technique for generating virtual adversarial samples in the embedding layer described in [MDG17]. We extend such approach from text-classification on LSTMs to question answering on DistilBERT, adding the following term to $\mathcal{L}_{QA}$

$$\begin{aligned}\mathcal{L}_{ADV}(E(c, q; \theta), r; \theta) &= KL[P_{start}(\cdot | E(c, q; \theta); \theta) \| P_{start}(\cdot | E(c, q; \theta) + r_{v-adv}; \theta)] \\ &\quad + KL[P_{end}(\cdot | E(c, q; \theta); \theta) \| P_{end}(\cdot | E(c, q; \theta) + r_{v-adv}; \theta)] \\ \text{where } r_{v-adv} &= \arg \max_{r, \|r\|_2 \leq \epsilon} KL[P_{start}(\cdot | E(c, q; \theta); \theta) \| P_{start}(\cdot | E(c, q; \theta) + r; \theta)] \\ &\quad + KL[P_{end}(\cdot | E(c, q; \theta); \theta) \| P_{end}(\cdot | E(c, q; \theta) + r; \theta)]\end{aligned}$$

Because $\mathcal{L}_{ADV}(E(c, q; \theta), r; \theta)$ takes minimum value at $r = 0$, assuming is differentiable with respect to θ and x almost everywhere, its first derivative $\nabla_r \mathcal{L}_{ADV}(E(c, q; \theta), r; \theta)|_{r=0}$ is zero. Therefore we need the second-order Taylor approximation, requiring the computation of the related Hessian, which requires $O(d^3)$ becoming unfeasibly large for high dimensional input space. Hence, we calculate an approximated virtual adversarial perturbation by performing a single iteration of the power method exposed in [MiMK⁺16] on

$$\begin{aligned}r_{v-adv} &= \epsilon g / \|g\|_2 \\ \text{where } g &= \nabla_{s+d} KL[P_{start}(\cdot | s; \theta) \| P_{start}(\cdot | s + d; \theta)] \\ &\quad + KL[P_{end}(\cdot | s; \theta) \| P_{end}(\cdot | s + d; \theta)]\end{aligned}$$

We experimented several values of ϵ (var_5_base²), including $\epsilon = 5$ which produced the best performance in [MDG17] for text-classification on LSTMs, several values for ξ , which is a parameter of the power method to approximate the dominant eigenvector of the Hessian [MiMK⁺16]. We also experimented a second iteration of the power method (var_5_pow_2²).

VAR-6. We experimented energy based modeling following [LCH⁺06, Hin02]. The energy function for this problem is defined as

$$\Gamma(c, q, y; \theta) = - \frac{\log(P_{start}(y_s | c, q; \theta)) + \log(P_{end}(y_e | c, q; \theta))}{2}$$

Our baseline can be seen as an energy based model with loss functions equals to the energy function, which in the context of EBMs is usually considered a sub-optimal choice. Given a training sample (c, q, y) , let's define the *most offending incorrect answer* $\bar{y} = (\bar{s}, \bar{e}) \in \mathbb{N}^2$ as the start / end position pair that has the lowest energy among all the answers that are different from the correct answer. Hence, our hinge loss (var_6_hinge²) is defined as

$$\mathcal{L}_{hinge}(c, q, y; \theta) = \max \{0, m + \Gamma(c, q, y; \theta) - \Gamma(c, q, \bar{y}; \theta)\} \quad (4)$$

where m is the positive margin. Adopting hinge loss the model is incentivated to minimize the difference between the energies of the correct answer and the most offending incorrect answer penalized linearly when larger than $-m$. The hinge loss only depends on energy differences, hence individual energies are not constrained to take any particular value, which during experimentation was observed to be a great problem, as explained in section 5. Hence, we modified the hinge loss with a contrastive term defined as the weighted average of all the answers that are different from the correct answer (var_6_neg²), i.e.

$$\mathcal{L}_{neg}(c, q, y; \theta) = \Gamma(c, q, y; \theta) - \frac{1}{\gamma} \cdot \frac{1}{\sum_{\tilde{y}: \tilde{y} \neq y} 1} \sum_{\tilde{y}: \tilde{y} \neq y} \Gamma(c, q, \tilde{y}; \theta) \quad (5)$$

ENS. Linear ensembles of best performant variants.

4 Experiments

Data. In-domain datasets are SQuAD [RZLL16] (50,000 train-samples, 10,507 dev-samples), NewsQA [TWY⁺17] (50,000 train-samples, 4,212 dev-samples), Natural Questions [KPR⁺19] (50,000 train-samples, 12,836 dev-samples). Out-of-domain datasets are DuoRC [SAKS18] (127 train-samples, 126 dev-samples, 1248 test-samples), RACE [LXL⁺17] (127 train-samples, 126 dev-samples, 419 test-samples), RelationExtraction [LSCZ17] (127 train-samples, 126 dev-samples, 2693 test-samples).

Evaluation method. Evaluation metrics are F1score and Exact Match (EM), as defined in [RZLL16].

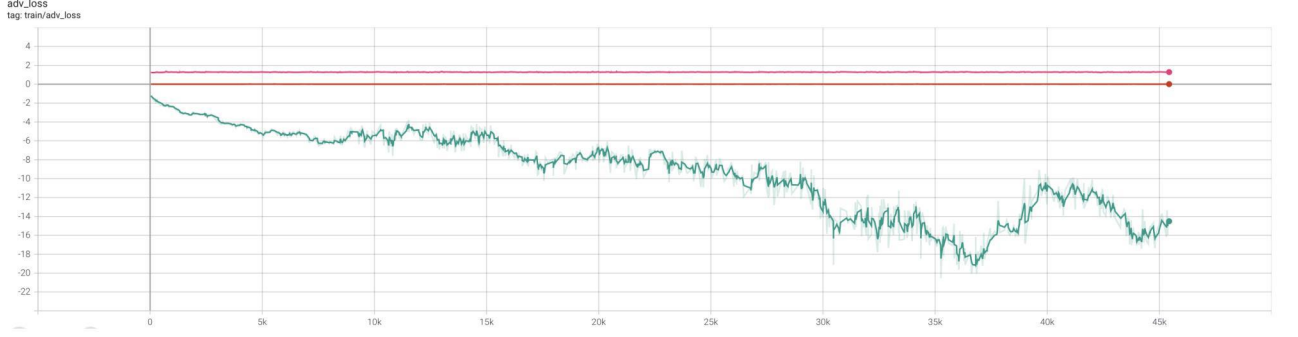
Experimental details. Unless otherwise specified in table 1, all experiments adopted learning rate $3e-5$, batch size 16, AdamW solver [LH17], models were trained on 3 epochs on in-domain train-sets and 20 epochs on out-of-domain train-sets. The specific hyperparameters adopted are indicated in table 1.

4.1 Results

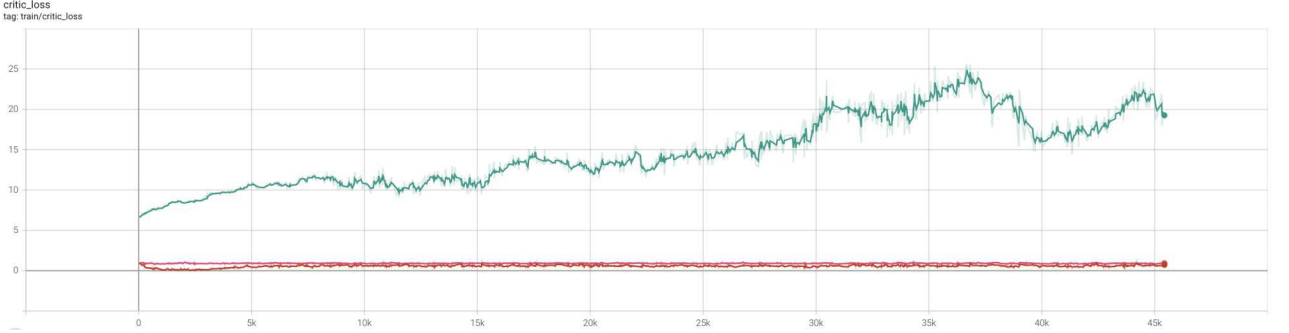
Table 1 reports quantitative results obtained from the different variants. The VAR-1 sub-variant var_1_anneal_sep_embed ($\lambda = 0.03$), even if not fine-tuned on the out-of-domain trainset, outperforms the baseline by $\sim 0.86\%$ F1-score. In [LKP19], for validation datasets (BioProcess [TSBW⁺13], ComplexWebQuestion [TB18], MCTest [RBR13], QAMR [MSH⁺17], QAST [JHT09] and TREC [VOO01]), the average F1-score of the models was about $\sim 1.5\%$ higher than the baseline, hence in our case, although we were able to outperform the baseline, we observe a in-domain bias reduction a bit lower. Also, our optimal variation adopts $\lambda = 0.03$ while the best one in [LKP19] adopts $\lambda = 0.01$,

Variation	Sub-variation	In-domain		Out-of-domain		
		Train F1/EM	Dev F1/EM	Train F1/EM	Dev F1/EM	Test F1/EM
baseline	baseline	81.97 / 66.32	70.56 / 54.70	49.49 / 32.81	49.17 / 35.34	-
	baseline_ft	78.56 / 61.64	66.58 / 50.45	82.39 / 68.50	50.36 / 34.82	-
VAR-1	var_1_base ($\lambda = 0.03$)	79.65 / 62.70	71.03 / 54.82	47.73 / 29.66	49.01 / 32.20	-
	var_1_base_ft ($\lambda = 0.03$)	-	-	82.26 / 67.98	50.20 / 34.82	-
	var_1_anneal ($\lambda = 0.01$)	81.54 / 65.31	70.81 / 54.37	46.93 / 29.92	48.28 / 31.68	-
	var_1_anneal_ft ($\lambda = 0.01$)	-	-	81.40 / 67.72	49.07 / 33.25	-
	var_1_anneal_sep_embed ($\lambda = 0.01$)	78.44 / 61.09	70.34 / 54.33	46.88 / 30.97	48.82 / 33.51	-
	var_1_anneal_sep_embed_ft ($\lambda = 0.01$)	-	-	82.11 / 68.50	50.45 / 35.34	-
	var_1_anneal_sep_embed ($\lambda = 0.03$) [♡]	78.61 / 61.29	70.33 / 54.53	47.69 / 29.66	51.22 / 34.55	-
	var_1_anneal_sep_embed_ft ($\lambda = 0.03$) [◇]	-	-	79.68 / 65.62	51.11 / 34.82	-
	var_1_random_token ($\lambda = 0.03$)	-	-	48.38 / 30.18	45.41 / 29.32	-
VAR-2	var_2_base ($\lambda = 0.03$)	78.44 / 61.09	70.34 / 54.33	46.88 / 30.97	48.43 / 33.25	-
	var_2_base_ft ($\lambda = 0.03$)	-	-	82.11 / 68.50	49.69 / 35.34	-
VAR-3	var_3_base ($\lambda = 0.03$)	78.05 / 60.96	70.91 / 54.87	50.22 / 34.12	47.39 / 33.77	-
	var_3_base_ft ($\lambda = 0.03$)	-	-	80.71 / 67.19	48.07 / 32.98	-
VAR-4	var_4_adv_loss ($\epsilon = 5$)	82.20 / 67.17	70.90 / 54.66	48.29 / 32.28	48.14 / 33.51	-
	var_4_adv_loss_ft ($\epsilon = 5$)	-	-	80.92 / 68.24	48.67 / 32.98	-
	var_4_base ($\epsilon = 5$)	79.88 / 63.70	72.09 / 56.15	51.47 / 35.43	49.69 / 35.34	-
	var_4_base_ft ($\epsilon = 5$) [♣]	-	-	72.90 / 56.17	51.57 / 36.39	59.41 / 41.42
	var_4_base ($\epsilon = 8$)	80.61 / 64.70	72.03 / 56.18	48.19 / 32.28	47.45 / 31.94	-
	var_4_base_ft ($\epsilon = 8$)	-	-	48.19 / 32.28	47.45 / 31.94	-
	var_4_base ($\epsilon = 2$)	83.50 / 68.25	71.78 / 55.71	49.57 / 31.76	48.54 / 35.08	-
	var_4_base_ft ($\epsilon = 2$) [♣]	-	-	80.80 / 66.93	51.26 / 35.08	-
	var_4_base_norm_embed ($\epsilon = 5$)	79.25 / 62.93	71.55 / 55.72	50.84 / 33.86	48.10 / 33.77	-
	var_4_base_norm_embed_ft ($\epsilon = 5$)	-	-	50.18 / 33.33	48.59 / 33.77	-
	var_4_base_norm_embed ($\epsilon = 8$)	78.38 / 61.97	71.60 / 55.83	49.89 / 34.12	49.72 / 34.03	-
	var_4_base_norm_embed_ft ($\epsilon = 8$)	-	-	49.25 / 33.86	48.77 / 33.25	-
	var_5_base ($\epsilon = 5, \xi = 1e - 1$)	77.84 / 60.22	70.22 / 54.06	48.84 / 31.50	49.14 / 33.77	-
VAR-5	var_5_base_ft ($\epsilon = 5, \xi = 1e - 1$)	-	-	49.07 / 32.02	48.94 / 33.25	-
	var_5_base ($\epsilon = 8, \xi = 1e - 1$)	78.98 / 61.70	70.61 / 54.46	45.58 / 28.87	44.96 / 29.06	-
	var_5_base_ft ($\epsilon = 8, \xi = 1e - 1$)	-	-	46.76 / 29.92	44.86 / 28.80	-
	var_5_base ($\epsilon = 2, \xi = 1e - 1$)	75.56 / 57.92	71.60 / 55.47	49.59 / 32.02	48.34 / 32.46	-
	var_5_base_ft ($\epsilon = 2, \xi = 1e - 1$)	-	-	49.37 / 32.02	48.50 / 32.98	-
	var_5_base ($\epsilon = 5, \xi = 1e - 6$)	77.84 / 60.22	70.22 / 54.06	48.84 / 31.50	49.14 / 33.77	-
	var_5_base_ft ($\epsilon = 5, \xi = 1e - 6$)	-	-	49.07 / 32.02	48.94 / 33.25	-
	var_5_pow_2 ($\epsilon = 5, \xi = 1e - 1$)	77.84 / 60.22	70.22 / 54.06	48.84 / 31.50	49.14 / 33.77	-
	var_5_pow_2_ft ($\epsilon = 5, \xi = 1e - 1$)	-	-	49.07 / 32.02	48.94 / 33.25	-
	var_6_hinge ($m = 1$)	-	-	11.60 / 04.72	12.24 / 04.19	-
VAR-6	var_6_neg ($\gamma = 10$)	78.35 / 60.56	69.34 / 53.53	45.57 / 27.82	45.84 / 30.89	-
ENS	$0.4^{\clubsuit} + 0.4^{\clubsuit} + 0.15^{\heartsuit} + 0.05^{\diamond}$	-	-	-	52.65 / 36.13	60.384 / 42.133
	$0.05^{\clubsuit} + 0.35^{\clubsuit} + 0.15^{\heartsuit} + 0.45^{\diamond}$	-	-	-	53.46 / 37.70	61.288 / 43.486

Table 1: Quantitative results obtained from the different variants. Sub-variants with the suffix “_ft” indicate models trained on in-domain trainsets (3 epochs) and on out-of-domain trainsets (20 epochs), while sub-variants without such suffix indicate models trained only on in-domain trainsets. Devset F1-score / EM in bold outperformed the baseline finetuned on oo-domain trainset (baseline_ft). We have five variants and two linear ensembles able to outperform in F1-score baseline_ft. Var_1_anneal_sep_embed ($\lambda = 0.03$), although not finetuned on oo-domain trainset, was able to outperform baseline_ft by a margin of approximately 0.86%, while var_4_base_ft ($\epsilon = 5$) was able to outperform baseline_ft by 1.21%. Our best linear ensemble was able to outperform baseline_ft by 3.10%. Test F1-score / EM are the ones obtained on the test leaderboard.



(a) $\mathcal{L}_{ADV}$ for *var_1_anneal_sep_embed* (pink), *var_2_base* (red), *var_3_base* (green), exponential smoothing averaged ($\alpha = 0.6$).



(b) $\mathcal{L}_{\mathcal{D}}$ for *var_1_anneal_sep_embed* (pink), *var_2_base* (red), *var_3_base* (green), exponential smoothing averaged ($\alpha = 0.6$).

Figure 1: $\mathcal{L}_{ADV}$ (a) and $\mathcal{L}_{\mathcal{D}}$ (b) for *var_1_anneal_sep_embed* (pink), *var_2_base* (red), *var_3_base* (green), exponential smoothing averaged ($\alpha = 0.6$).

and we adopted a different procedure to extract [SEP] tokens than the one in [LKP19], that we think was not implemented correctly. As explained in section 5, we found confirmation $\mathcal{L}_{\mathcal{D}}$ is affected by vanishing gradients both in VAR-1 and VAR-2, although this problem affects VAR-2 more seriously. VAR-3 was designed to solve the problem adopting Wasserstein GAN, but doing so it creates a bigger problem, as explained in section 5. The best sub-variant of VAR-4, *var_4_base_ft* ($\epsilon = 5$), was able to outperform *baseline_ft* by 1.21%, achieving a superior improvement than the one found in [MDG17] ($\sim 1.11\%$ error rate). On the other hand, no VAR-5 sub-variant was able to outperform *baseline_ft*, which is pretty surprising considering that virtual adversarial training achieved the best improvement over the baseline in [MDG17]. Energy based sub-variants of VAR-6 performed pretty poorly, although the one adopting hinge loss was affected by higher bias, as explained in section 5. Finally, our best linear ensemble was able to outperform *baseline_ft* by 3.10%.

5 Analysis

The gap between in-domain train F1-score vs. dev F1-score is a measure of the *variance* of the learning algorithm, and we can see that in the baseline it is around $\sim 11.41\%$, while the gap between in-domain dev F1-score vs. out-of-domain dev F1-score is a measure of the *in-domain bias* of the learning algorithm, and we can see that in the baseline it is around $\sim 21.39\%$. On the other hand, when a model is finetuned on out-of-domain train-sets, the gap between out-of-domain dev F1-score of the finetuned model vs. out-of-domain dev F1-score of the non-finetuned one is a measure of the *out-of-domain finetuning gain*, and we can see that in baseline vs. *baseline_ft* it is around $\sim 1.19\%$. Domain adaptation techniques focus on maximizing the out-of-domain finetuning gain, while domain generalization techniques, which are the ones we adopted, focus on minimizing the in-domain bias.

Fig. 1 shows $\mathcal{L}_{ADV}$ and $\mathcal{L}_{\mathcal{D}}$ for key sub-variants of VAR-1, VAR-2 and VAR-3. Specifically, we found confirmation $\mathcal{L}_{\mathcal{D}}$ is affected by vanishing gradients both in VAR-1 (KL-divergence loss in [LKP19]) and VAR2 (the discriminator loss in the original GAN paper [GPAM⁺14]). VAR-3 was designed to solve the problem adopting Wasserstein GAN [ACB17] with gradient penalty and indeed it fixed the vanishing gradient problem, but it is clear that the discriminator is outmatched

by the QA model. This suggests the choice in [LKP19] of adopting a discriminator based on a multilayer perceptron is not optimal considering the QA model is a transformer with a so larger capacity. This observation suggested to extract not only the [SEP] and [CLS] tokens but also k random token representations from DistilBERT (*var_1_random_token*) to enrich the feature space of the discriminator. Better results could be probably achieved designing and training a larger capacity discriminator, possibly based on the same architecture of the QA model (choice, that usually works in computer vision).

VAR-4 in its default setting (i.e. *var_4_base_ft* with $\epsilon = 5$) leverages amazingly the idea of local distributional smoothness (LDS) expressed in [MiMK⁺16]. We can see this is not only the best sub-variant but also the one with the highest in-domain dev F1-score, which is a wonderful confirmation of the regularization effects of LDS, i.e. promoting smoothness of the model distribution. Interestingly, as showed in [MiMK⁺16], LDS regularization is parametrization invariant, i.e. if $\theta^* = \arg \max_{\theta} T(\theta)$ then $f(\theta^*) = \arg \max_{\theta} T(f(\theta))$, property not satisfied by the more popular L_q regularization [FT01]. This could explain why we found so good performance with our first-guess parameters. Indeed, we experimented $\epsilon > 5$ more as double check without really expecting improvements, as the adversarial perturbation is clearly too far from the initial input embeddings, while $\epsilon < 5$ still assures good performance. This potential correlation between model regularization and in-domain bias reduction should be investigated in future study.

On the other hand, the virtual adversarial regularization method experimented with VAR-5 affects the model with high bias, as proven by the low in-domain train F1-score, leading to pretty similar performance regardless of the hyper-parameters adopted. This is a bit surprising considering the results reported in [MDG17] for text classification on LSTMs, and probably further study should investigate ways to mitigate these side effects while preserving the regularization properties.

Experimenting the EBMs of VAR-6, we observed the hinge loss of equation 4 only depends on energy differences, hence individual energies are not constrained to take any particular value, which was a major problem. While the model can learn the correct answer, it can hardly learn the most offending incorrect answer, as clearly shown by the related learning curves. The loss with the contrastive term of equation 5 improves significantly the performance, but not enough to beat the baseline.

Lastly, the fact that our best linear ensemble, combining VAR-1 and VAR-4 sub-variants, was able to outperform *baseline_ft* by 3.10%, significantly better than our best single model, proves an high ensemble gain which, in turn, suggests these variants are able to regularize the model producing accurate predictions relatively weakly correlated.

6 Conclusion

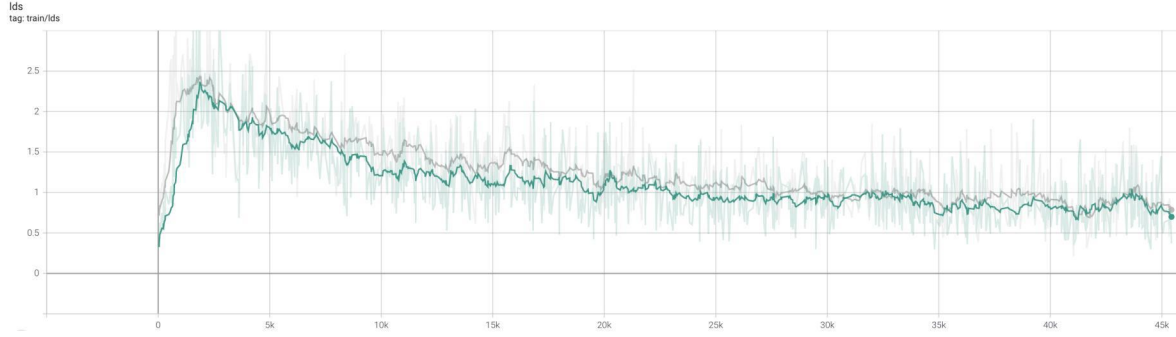
We experimented adversarial training, virtual adversarial training and energy-based methods to reduce in-domain bias in cross-domain question answering. We found successful two different and complementary approaches of adversarial training. One is based on incentivating the model to learn domain-invariant features from specific tokens and against a discriminator network, while the other is based on challenging the model through adversarial perturbations on word and positional embeddings, promoting local distributional smoothness. Combined together through linear ensembling, they achieve >3% F1-score improvement over the regular fine-tuning process, proving both their effectiveness and complementarity. Virtual adversarial training turned out to have an overwhelming regularization effect on the training process, while experimented energy-based methods biased excessively the model. Further research would be advisable to investigate two interesting directions. One is designing a more balanced variant of the first adversarial training approach, having a discriminator network with comparable capacity of the QA model and based on transformers. In this way, adopting Wasserstein GAN to prevent vanishing gradients, the discriminator would not be easily outmatched, pushing even more the QA model to learn domain-invariant features. The other is investigating whether the side effects of virtual adversarial training can be mitigated preserving the regularization properties.

References

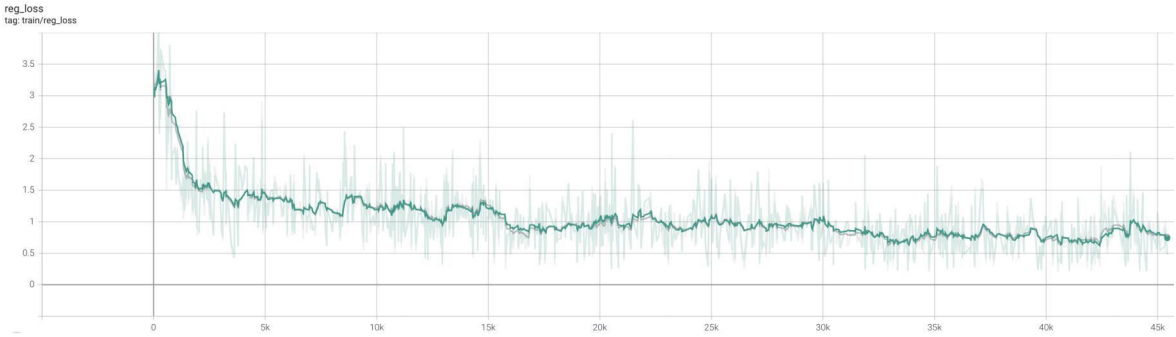
[ACB17] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017. 1, 3, 5

- [ASE⁺18a] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2890–2896, Brussels, Belgium, October–November 2018. Association for Computational Linguistics. 1, 2
- [ASE⁺18b] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang. Generating natural language adversarial examples. *CoRR*, abs/1804.07998, 2018. 2
- [CJME20] Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. Advaug: Robust adversarial augmentation for neural machine translation, 2020. 2
- [DBO⁺20] Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc’Aurelio Ranzato. Residual energy-based models for text generation, 2020. 2
- [FLL⁺19] Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing, 2019. 3
- [FT01] Hastie Trevor Friedman, Jerome and Robert Tibshirani. *The elements of statistical learning*. Springer series in statistics Springer. Springer, 2001. 5
- [GPAM⁺14] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. 3, 5
- [GS05] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005. 2
- [GSS15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. 1, 2
- [HDB19] Timothy J. Hazen, Shehzaad Dhuliawala, and Daniel Boies. Towards domain adaptation from limited data for question answering using deep neural networks. *CoRR*, abs/1911.02655, 2019. 1
- [Hin02] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. 1, 2, 3
- [HLQP21] Shuo Huang, Zhuang Li, Lizhen Qu, and Lei Pan. On robustness of neural semantic parsers, 2021. 2
- [HMXHA21] Tianxing He, Bryan McCann, Caiming Xiong, and Ehsan Hosseini-Asl. Joint energy-based model training for better calibrated natural language understanding models, 2021. 2
- [JHT09] Wittawat Jitkrittum, Choochart Haruechaiyasak, and Thanaruk Theeramunkong. QAST: Question answering system for ThaiWikipedia. In *Proceedings of the 2009 Workshop on Knowledge and Reasoning for Answering Questions (KRAQ 2009)*, pages 11–14, Suntec, Singapore, August 2009. Association for Computational Linguistics. 4.1
- [KPR⁺19] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*, 2019. 4
- [LCH⁺06] Yann LeCun, Sumit Chopra, Raia Hadsell, Fu Jie Huang, and et al. A tutorial on energy-based learning. In *PREDICTING STRUCTURED DATA*. MIT Press, 2006. 1, 2, 3

- [LH17] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017. 4
- [LKP19] Seanie Lee, Donggyu Kim, and Jangwon Park. Domain-agnostic question-answering with adversarial training. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 196–202, Hong Kong, China, November 2019. Association for Computational Linguistics. 1, 2, 3, 3, 1, 4.1, 5
- [LSCZ17] Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. Zero-shot relation extraction via reading comprehension. *CoRR*, abs/1706.04115, 2017. 4
- [LXL⁺17] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. RACE: large-scale reading comprehension dataset from examinations. *CoRR*, abs/1704.04683, 2017. 4
- [MDG17] Takeru Miyato, Andrew M. Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification, 2017. 1, 2, 3, 3, 4.1, 5
- [Mil95] George A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995. 2
- [MiMK⁺16] Takeru Miyato, Shin ichi Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional smoothing with virtual adversarial training, 2016. 1, 2, 3, 5
- [MSH⁺17] Julian Michael, Gabriel Stanovsky, Luheng He, Ido Dagan, and Luke Zettlemoyer. Crowdsourcing question-answer meaning representations, 2017. 4.1
- [RBR13] Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. 4.1
- [RSG18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia, July 2018. Association for Computational Linguistics. 1, 2
- [RZLL16] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. *CoRR*, abs/1606.05250, 2016. 4, 4
- [SAKS18] Amrita Saha, Rahul Aralikkatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. Duorc: Towards complex language understanding with paraphrased reading comprehension. *CoRR*, abs/1804.07927, 2018. 4
- [SDCW20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. 1, 3
- [SMNM17] Motoki Sato, Hitoshi Manabe, Hiroshi Noji, and Yuji Matsumoto. Adversarial training for cross-domain Universal Dependency parsing. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 71–79, Vancouver, Canada, August 2017. Association for Computational Linguistics. 2
- [TB18] Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions, 2018. 4.1
- [TSBW⁺13] Aju Thalappillil Scaria, Jonathan Berant, Mengqiu Wang, Peter Clark, Justin Lewis, Brittany Harding, and Christopher D. Manning. Learning biological processes with global constraints. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1710–1720, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. 4.1



(a) Adversarial loss for key VAR-5 sub-variants, exponential smoothing averaged ($\alpha = 0.6$).



(b) Regular QA loss for key VAR-5 sub-variants, exponential smoothing averaged ($\alpha = 0.6$).

Figure 2: Adversarial and regular QA loss for key VAR-5 sub-variants, exponential smoothing averaged ($\alpha = 0.6$). We can see after initial $\sim 8,000$ iteration, both the adversarial and regular QA loss are pretty much unable to improve significantly.

[TWY⁺17] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada, August 2017. Association for Computational Linguistics. 4

[VOO01] ELLEN M. VOORHEES. The trec question answering track. *Natural Language Engineering*, 7(4):361–378, 2001. 4.1

[YF16] Mengwen Liu Yi Fang. A unified energy-based framework for learning to rank. In *ICTIR. IACM*, 2016. 2

[YZWY17] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. SeqGAN: Sequence Generative Adversarial Nets with Policy Gradient. In *The Thirty-First AAAI Conference on Artificial Intelligence (AAAI 2017)*, 2017. 2

[ZLZ⁺21] Danqing Zhu, Wangli Lin, Yang Zhang, Qiwei Zhong, Guanxiong Zeng, Weilin Wu, and Jiayu Tang. At-bert: Adversarial training bert for acronym identification winning solution for sdu@aaai-21, 2021. 2

7 Appendix

7.1 Learning Curves

Fig. 2 shows adversarial and regular QA loss for key VAR-5 sub-variants, while fig. 3 shows EBM’s hinge loss of equation 4 (red) and the improved one with the contrastive term of equation 5 (blue).

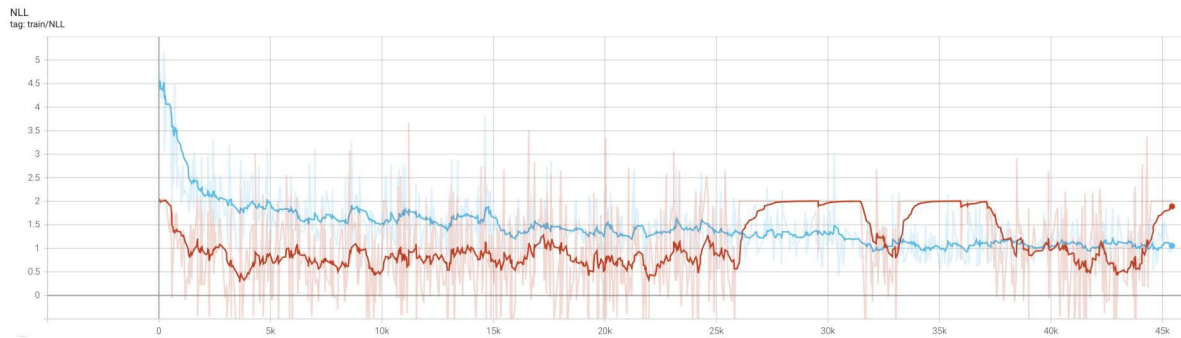


Figure 3: EBM’s hinge loss of equation 4 (red) and the improved one with the contrastive term of equation 5 (blue), exponential smoothing averaged ($\alpha = 0.6$). While the model can learn the correct answer, it can hardly learn the most offending incorrect answer.

7.2 Saliency Maps

Fig. 4 shows examples of output saliency maps generated by the baseline and var_4_base_ft ($\epsilon = 5$), while fig. 5 shows examples of output saliency maps generated by (a) the baseline and (b) var_1_anneal_sep_embed ($\lambda = 0.03$). They are obtained considering magnitudes of gradients along the sequence in the embedding space. Tokenization is performed at sub-word level.

[CLS] where is the longest mountain range in the world ? [SEP] our planet was born some billions of years ago . ever since this planet has developed , it has surprised us human beings . no matter how much we human beings have developed technological ##ly , we cannot ever create this kind of natural planet . the human race can only discover some amazing facts about it . let us find out what some of them are . on september 13 , 1922 el aziz ##ia , a place in libya , recorded a temperature of 136 degrees fa ##hre ##nh ##eit . this was the highest temperature ever measured on the planet earth . the highest waterfall in the world is angel falls in venezuela , which drops from an amazing height of 3 , 212 feet . the waterfall drops over the edge of au ##yan ##te ##pu ##i in the can ##ai ##ma national park . the gravitational pull of the earth can be escaped if someone moves at a speed of 6 . 95 miles per second . at such a speed , one could travel from new york to philadelphia in just 20 seconds . if the entire water in the world ' s oceans eva ##por ##ated , the salt we got from these oceans could cover the entire planet with a 500 - foot deep layer of salt . the deepest lake on earth is lake bai ##kal in russia . lake bai ##kal is the largest freshwater lake in the world , containing about 20 % of the world ' s un ##fr ##oz ##en surface fresh water . the andes in south america spreads over 4 , 525 miles . this is the longest mountain range in the world . did you know that the biggest valley in the world is the grand valley in mesa county , colorado and grand county , utah ? it is about 30 miles (48 km) long and 5 miles (8 km) wide , along the colorado river . the amazing facts about the earth are too many to be told completely above . [SEP]

(a) Output saliency map generated by the baseline.

[CLS] where is the longest mountain range in the world ? [SEP] our planet was born some billions of years ago . ever since this planet has developed , it has surprised us human beings . no matter how much we human beings have developed technological ##ly , we cannot ever create this kind of natural planet . the human race can only discover some amazing facts about it . let us find out what some of them are . on september 13 , 1922 el aziz ##ia , a place in libya , recorded a temperature of 136 degrees fa ##hre ##nh ##eit . this was the highest temperature ever measured on the planet earth . the highest waterfall in the world is angel falls in venezuela , which drops from an amazing height of 3 , 212 feet . the waterfall drops over the edge of au ##yan ##te ##pu ##i in the can ##ai ##ma national park . the gravitational pull of the earth can be escaped if someone moves at a speed of 6 . 95 miles per second . at such a speed , one could travel from new york to philadelphia in just 20 seconds . if the entire water in the world ' s oceans eva ##por ##ated , the salt we got from these oceans could cover the entire planet with a 500 - foot deep layer of salt . the deepest lake on earth is lake bai ##kal in russia . lake bai ##kal is the largest freshwater lake in the world , containing about 20 % of the world ' s un ##fr ##oz ##en surface fresh water . the andes in south america spreads over 4 , 525 miles . this is the longest mountain range in the world . did you know that the biggest valley in the world is the grand valley in mesa county , colorado and grand county , utah ? it is about 30 miles (48 km) long and 5 miles (8 km) wide , along the colorado river . the amazing facts about the earth are too many to be told completely above . [SEP]

(b) Output saliency map generated by var_4_base_ft ($\epsilon = 5$).

Figure 4: Examples of output saliency maps generated by (a) the baseline and (b) var_4_base_ft ($\epsilon = 5$). Q: *where is the longest mountain range in the world?* They are pretty similar, but we can see that while the most “influential” token for the prediction of the baseline is “America” (which is different from the predicted answer), the most “influential” token for the prediction of our VAR-4 sub-variant is “Andes”, which is the specific name of mountain ranges.

[CLS] who is lo ' s daughter ? [SEP] the film starts with a brief rec ##oll ##ection of the " killing angel " , a dangerous assassin that has been operating in most of southeast asia . the killing angel has no known acc ##omp ##lices or colleagues , is not affiliated with any government , religious , or criminal organization , and to this date their true identity has still remained secret . the killing angel is believed to be responsible for 6 murders in the past year alone , with only one thing connecting each of the assassination ##s - all the victims were criminals and deserved to die . then the movie quickly moves on to the killing angel ' s most recent killing , the well - publicized assassination of a notorious ex - ya ##ku ##za crime boss . working swiftly , the killing angel is able to overcome the criminal ' s defense systems , quickly inca ##pac ##itating the boss ' s bodyguards and iso ##lating his target . in a fit of desperation , the crime boss promises to triple whatever fee the killing angel is being paid to kill him , then goes on to warn the vi ##gil ##ante that there is a massive revenge fund set to avenge his killing . ignoring his pleas , the killing angel silently kills the crime boss . the scene then switches over to the story of fu (jet li) , a small - town gang member who lives an extremely simple and modest lifestyle , albeit being a contract killer . it is revealed that fu is very kind and has trouble killing innocent people , a trait that causes him to be disliked by the other low - rent triad members . when fu ' s boss reveals that there is now a

(a) Output saliency map generated by the baseline.

[CLS] who is lo ' s daughter ? [SEP] the film starts with a brief rec ##oll ##ection of the " killing angel " , a dangerous assassin that has been operating in most of southeast asia . the killing angel has no known acc ##omp ##lices or colleagues , is not affiliated with any government , religious , or criminal organization , and to this date their true identity has still remained secret . the killing angel is believed to be responsible for 6 murders in the past year alone , with only one thing connecting each of the assassination ##s - all the victims were criminals and deserved to die . then the movie quickly moves on to the killing angel ' s most recent killing , the well - publicized assassination of a notorious ex - ya ##ku ##za crime boss . working swiftly , the killing angel is able to overcome the criminal ' s defense systems , quickly inca ##pac ##itating the boss ' s bodyguards and iso ##lating his target . in a fit of desperation , the crime boss promises to triple whatever fee the killing angel is being paid to kill him , then goes on to warn the vi ##gil ##ante that there is a massive revenge fund set to avenge his killing . ignoring his pleas , the killing angel silently kills the crime boss . the scene then switches over to the story of fu (jet li) , a small - town gang member who lives an extremely simple and modest lifestyle , albeit being a contract killer . it is revealed that fu is very kind and has trouble killing innocent people , a trait that causes him to be disliked by the other low - rent triad members . when fu ' s boss reveals that there is now a

(b) Output saliency map generated by var_1_anneal_sep_embed ($\lambda = 0.03$).

Figure 5: Examples of output saliency maps generated by (a) the baseline and (b) var_1_anneal_sep_embed ($\lambda = 0.03$). Q: *who is Lo's daughter?* While there is no answer to such question in the paragraph, the baseline is highly affected by words like "Southeast Asia" and "triad members". On the contrary, our VAR-1 sub-variant looks to be weakly influenced by the details of the paragraph.