Analyzing the Forgetting Problem in Pretrain-Finetuning of Open-domain Dialogue Response Models

Tianxing He
MIT
tianxing@mit.edu

Jun Liu
Facebook
junliu@fb.com

Kyunghyun Cho
New York University
kyunghyun.cho@nyu.edu

Myle Ott
Facebook
myleott@fb.com

Bing Liu
Facebook
bingl@fb.com

James Glass
MIT
glass@mit.edu

Fuchun Peng
Facebook
fuchunpeng@fb.com

Abstract

In this work, we study how the finetuning stage in the pretrain-finetune framework changes the behavior of a pretrained neural language generator. We focus on the transformer encoder-decoder model for the open-domain dialogue response generation task. Our major finding is that after standard finetuning, the model *forgets* some of the important language generation skills acquired during large-scale pretraining. We demonstrate the forgetting phenomenon through a set of detailed behavior analysis from the perspectives of knowledge transfer, context sensitivity, and function space projection. As a preliminary attempt to alleviate the forgetting problem, we propose an intuitive finetuning strategy named "mix-review". We find that mix-review effectively regularizes the finetuning process, and the forgetting problem is alleviated to some extent. Finally, we discuss interesting behavior of the resulting dialogue model and its implications.

1 Introduction

Large-scale unsupervised pretraining (Peters et al., 2018; Devlin et al., 2018; Song et al., 2019; Yang et al., 2019; Liu et al., 2019) has recently been shown to greatly boost the performance of natural language processing (NLP) models. On a high level, the pretrain-finetune framework can be viewed as a simple two-stage procedure: (1) Use large-scale unsupervised text data to pretrain the model; (2) Use target task data to finetune the model.

Recently, multiple works (Radford et al., 2019; Jiang et al., 2020; Roberts et al., 2020; Talmor et al., 2019) have reported that pretrained language models (LM) have implicitly stored large amounts of "world knowledge" in its parameters, and are able to answer common-sense questions. While these studies are encouraging, during the finetuning stage the model is usually trained on a dataset that is very different from the pretraining data, which leads to the potential danger that the model could **forget** precious skills gained during pretraining. This is an important question for open-domain dialogue response generation, which is the focus of our work, because the knowledge acquired during pretraining can greatly help make the dialogue interaction more engaging or informative.

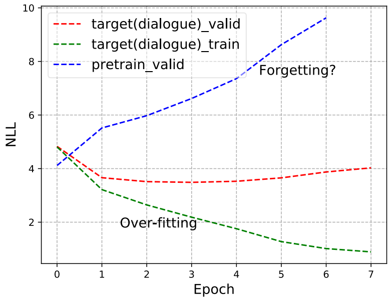


Figure 1: During finetuning, the model's performance on the pretraining data drastically degrades.

In Figure 1, we show that during finetuning, the model's performance on the pretraining data drastically degrades. While this drop is concerning, it does not necessarily mean that the skills from pretrained model are not well transferred to the end dialogue task, because the model should be evaluated in a dialogue setting.

To better answer the question about how finetuning changes the pretrained model's behavior, in this work we conduct a set of behavior analysis from the perspectives of knowledge transfer, context sensitivity, and function space projection. Our major finding is that in the finetuning stage, data separation causes the model to *forget* some of the important language generation skills acquired during pretraining. We also show that the forgetting problem can be alleviated by mixing pretraining and target-task data during finetuning.

2 Model Formulation

In this work we study the pretrain-finetune framework from the viewpoint of neural language generation (NLG). In particular, we focus on the open-domain dialogue response task, for the following reasons: (1) There is high similarity between the target dialogue response task (conditional NLG) and the pretraining language modeling (LM) objective, so we expect that language generation skills learnt during pretraining can be well transferred to the down-stream target task. (2) The sequence-to-sequence (seq2seq) nature of the model allows us to characterize the model's generation behavior in various ways (e.g., context sensitivity).

End-to-end dialogue response generation (Li et al., 2016) can be formulated as a sequence-to-sequence (seq2seq) task: Given a dialogue context (previous utterances), the model is asked to generate a high-quality response. In this work we adopt the encoder-decoder model architecture (Sutskever et al., 2014; Cho et al., 2014), which is widely used in NLG applications like dialogue response generation (Li et al., 2016), machine translation (Luong et al., 2015), etc. In particular, we use the transformer model (Vaswani et al., 2017), which has currently become the most popular encoder-decoder model architecture (Young et al., 2017). We use the same configuration as (Vaswani et al., 2017), which has 6 encoder/decoder layers, 16 attention heads, with an embedding dimension of 1024 and a feed-forward dimension of 4096.

During standard finetuning, the Adam optimizer (Kingma and Ba, 2014) is used to minimize the negative log-likelihood (NLL) of the reference target sentence $\boldsymbol{y}$ given the input context $\boldsymbol{x}$ in the data distribution (denoted as P_{data}):

$$\begin{aligned}\mathcal{L}_{\text{finetune}}(P_{data};\theta) &= E_{(\boldsymbol{x},\boldsymbol{y})\sim P_{data}}(-\log P_\theta(\boldsymbol{y}|\boldsymbol{x})) \\ &= E_{(\boldsymbol{x},\boldsymbol{y})\sim P_{data}}(-\sum_{t=1}^{m} \log P_\theta(y_t|\boldsymbol{y}_{<t},\boldsymbol{x})),\end{aligned} \tag{1}$$

where $\boldsymbol{y}_{<t}$ refers to $\{y_0, y_1, ..., y_{t-1}\}$, in which y_0 is set to a begin-of-sentence token `<BOS>`, and y_m is a end-of-sentence token `<EOS>`. In the dialogue response setting, the input $\boldsymbol{x}$ is a concatenation of previous utterances. We truncate the length of $\boldsymbol{x}$ to be at most 128 words, which typically includes around 6 previous utterances.

Given a trained seq2seq model, to generate a response for some contextual input, one needs to choose a decoding or sampling method. Recent research (Holtzman et al., 2019; Radford et al., 2019; Fan et al., 2018) has shown that a strategy called top-k sampling, in which the next word is sampled from the top k most probable choices, is a better choice than the traditional beam-search decoding, due to better diversity. Our preliminary experiments (Appendix A) have also verified this claim in the open-domain dialogue response setting. As a result, in this work, unless otherwise mentioned, we use top-k sampling as the default decoding method. In particular, we set k to 30 (we find it to work well in preliminary experiments).

Dialogue
Input: what did you do yesterday ? `<eou>` i watched the avengers movie .
Output: wow ! i am crazy about iron man !
Next-sentence Pretraining
Input: the avengers are super hot currently . `<eou>` the next movie will be on in April .
Output: fans are talking about iron man on the internet .
MASS Pretraining
Input: fans are talking about `<MASK>` `<MASK>` `<MASK>` will do on the internet .
Output: what iron man

Table 1: Illustrations of input-output pairs for typical dialogue response training, next-sentence pretraining, or MASS pretraining.

3 The Pretrain-Finetune Framework

In this section we first review the pretrain-finetune framework for encoder-decoder models. We discuss the language generation skills the model can acquire during pretraining, and more importantly, how we check whether the skills are "forgotten" during finetuning. Finally, as a preliminary attempt to alleviate the forgetting problem, we propose the mix-review finetuning strategy.

3.1 Pretraining

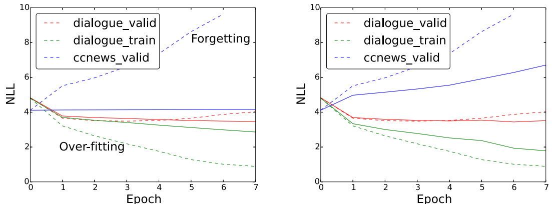
In this work, we consider pretraining the seq2seq model using large-scale unsupervised text data, and afterwards finetuning it using target dialogue data. We compare two representative strategies: next-sentence (NS) pretraining and masked sequence-to-sequence (MASS) pretraining (Song et al., 2019). Next-sentence pretraining is a natural extension of GPT-style LM training (Radford et al., 2019; Kiros et al., 2015) for encoder-decoder models. For every sentence in a given training document, we set the previous sentences as the contextual input, and ask the model to generate the next sentence. We omit the formulation of NS because it is very similar to Equation (1).

Masked sequence-to-sequence pretraining (MASS) can be regarded as an extension of the "BERT" (Devlin et al., 2018) pretraining for encoder-decoder models. For each sentence, a random segment of the sentence is masked, and the model is trained to generate the masked words on the decoder side. We refer readers to (Song et al., 2019) for more details.

In Table 1, we illustrate the similarity between NS pretraining and typical dialogue response training. Compared to NS pretraining, MASS has the disadvantage that it focuses on one single sentence at a time. However, the context of multiple previous sentences are very important for dialogue response generation.

3.2 Analyzing the Forgetting Problem

Although recently a number of pretraining strategies (Peters et al., 2018; Devlin et al., 2018; Song et al., 2019; Yang et al., 2019; Liu et al., 2019) have been proposed for various NLP tasks, the finetuning stage remains simple and straightforward: simply finetune all parameters with a relatively small learning rate.



(a) Mix-review (b) WD(θ_{pre})

Figure 2: Model's performance on different evaluation sets during the finetuning stage, for the Dailydialog dataset (to be described in Section 4.1). The dotted lines represent the original finetuning process, while the solid lines represent when mix-review or WD(θ_{pre}) is applied.

In Figure 2a, we show (with the dotted lines) the model's negative log-likelihood (NLL) on different evaluation sets during the finetuning stage. We identify two potential issues during finetuning: (1) **Over-fitting:** The gap between training-set NLL and validation-set NLL increases quickly. (2) **Forgetting:** The performance on the pretraining CCNEWS data (to be described in Section 4.1) drops drastically. Note that the forgetting phenomenon here is not necessarily "catastrophic" as in the sequential learning case (Atkinson et al., 2018; Robins, 1995), because the goal is to achieve the best performance on the target dialogue dataset, and the model does not need to maintain fidelity to the pretraining data. However, **it leads us to question whether the model has lost some important skills learned during pretraining**.

In this work we analyze two important generation capabilities that the model can acquire in the pretraining stage, and will be useful for the target dialogue setting. One is the *acquisition of knowledge*: the large-scale pretraining text data contains a large amount of knowledge, and can be used to make dialogue responses more informative and engaging (e.g., the model can learn about the "Avengers" movie, and use it as a topic). To quantify how knowledgeable the finetuned model is, we prepare a set of knowledge terms such as iphone, pokemon, etc., and the corresponding reference description. We then query the model about these knowledge terms, and compare its output against the reference. We also conduct multi-turn human evaluation in the setting of knowledgeable conversations. More details will be given in Section 5.1.

The other ability is the *utilization of contextual input*: as shown by (Sankar et al., 2019), the current open-domain dialogue models (without pretraining) are insensitive to contextual input, which gives rise to the generic response problem (Li et al., 2016). In our preliminary experiments with NS pretraining, we find that similarly to the GPT model (Radford et al., 2019) the pretrained model has the ability to generate closely related responses given the previous sentences as input. Ideally during finetuning, the model can transfer this skill to the target dialogue task. To quantify the model's sensitivity to context, following (Sankar et al., 2019), we add noise to the input, and measure the relative drop in perplexity. More details will be given in Section 5.2.

3.3 The Mix-review Finetuning Strategy

As a preliminary attempt to alleviate the forgetting problem, we propose a finetuning strategy named "mix-review (MR)": For each finetuning epoch, we mix the target dialogue data with a random subset of the pretraining data. This process introduces two hyper-parameters: *mix-ratio*, which controls how much pretraining data is mixed, and *mix-decay*, which decays mix-ratio by each epoch. For example, assume the target dialogue training set has

100k utterances, mix-ratio=4 and mix-decay=0.9, then in the first epoch of mix-review finetuning, 400k pretraining utterances will be mixed in, and for the second epoch the amount will be reduced to 360k utterances, etc.

We formulate the mix-review objective as below:

$$\begin{aligned}\mathcal{L}_{\text{mix-review}} = & \mathcal{L}_{\text{finetune}}(P_{\text{target-data}}; \theta) + \\ & \text{mix-ratio} \cdot \mathcal{L}_{\text{pretrain}}(P_{\text{pretrain-data}}; \theta).\end{aligned} \quad (2)$$

Note that the augmented mixing term can be viewed as a regularization term.

We tune the hyper-parameters (mix-ratio and mix-decay) in the grid of $\{1, 2, 4, 8, 16\} \times \{1, 0.9, 0.8, 0.7, 0.6, 0.5\}$ (using the same learning rate and other hyper-parameters with standard finetuning), and report with the best model based on the perplexity (PPL) performance on the validation set of the target task. We find that the performance gain of mix-review is not sensitive to hyper-parameter tuning: a small mix-ratio of 4 typically works well, which means the computational cost of mix-review is comparable to standard finetuning.

In Figure 2a, we show the loss curve for mix-review finetuning with a mix-ratio of 4 and a mix-decay of 0.7. We observe that the performance on the pretraining CCNEWS data is preserved, which strongly supports the motivation of mix-review. Furthermore, we observe a regularization effect from mix-review (narrowing the gap between training and testing performance).

We compare mix-review with the L_2 regularization (weight decay) toward the pretrained parameters θ_{pre} (Kirkpatrick et al., 2016). We denote it as WD(θ_{pre}) and formulate it as follows:

$$\mathcal{L}_{\text{finetune}}(P_{\text{target-data}}; \theta) + \lambda \cdot \left\|\theta - \theta_{\text{pre}}\right\|_2^2 . \quad (3)$$

In our experiments, we tune λ in the set $\{10^{-1}$,10^{-2},10^{-3},10^{-4},$10^{-5}\}$ and report with the best model based on PPL on the validation set.

In Figure 2b we show the loss curve for WD(θ_{pre}) with $\lambda = 0.1$. We observe that WD(θ_{pre}) also has a regularization effect, but it is not as strong as mix-review.

Additionally, we tried the following two basic regularization techniques: (1) Increase the rate of dropout; (2) Freeze the bottom layers of the model during finetuning. However, these two techniques show little or no improvement. The reason could be that the transformer is already a well-tuned model (e.g., it features dropout and layer normalization).

4 Datasets and Implementation Details

4.1 Datasets

For pretraining, we use the large-scale CCNEWS data (Bakhtin et al., 2019) which is a de-duplicated subset of the English portion of the CommonCrawl news dataset[1]. The dataset contains news articles published worldwide between September 2016 and February 2019. It has in total around 1 billion sentences or 27 billion words. To be able to complete experiments in a reasonable amount of time, we use the first 10 percent of the CCNEWS data for pretraining, which contains 100 million sentences and 2.7 billion words.

For finetuning, three open-domain conversational dialogue datasets are used: Dailydialog (1.3 million words) (Li et al., 2017), Switchboard (1.2 million words), and Cornell Movie (Danescu-Niculescu-Mizil and Lee, 2011) (4.5 million words). To save space, we defer the details of the data-sets to Appendix B.

To construct the vocabulary, we learn codes of Byte Pair Encoding (BPE) (Sennrich et al., 2016) from the CCNEWS-100m data with 50k merges. This results in a vocabulary of size 62k. We then apply the same BPE codes to all target dialogue datasets.

4.2 Implementation

Our code is based on the Fairseq toolkit (Ott et al., 2019). The Adam optimizer (Kingma and Ba, 2014) is used for all experiments. For pretraining of both MASS and NS, we use a mini-batch size of 2048, with the learning rate (LR) set to 0.0001. Following (Vaswani et al., 2017), the "inverse square root" LR scheduler with a warm-up stage is used. Pretraining is conducted on 32 GPUs and half-precision (float16) speed-up is used. For both MASS and NS, we stop the pretraining after the CCNEWS data is swept 20 times. For all our experiments, a dropout rate of 0.1 is applied to the transformer model. We follow Song et al. (2019) for the recommended hyper-parameter setting of MASS (e.g., how to select the mask span).

Finetuning is done on 2 GPUs without float16 speed-up. The learning rate is halved when the PPL on the validation set does not improve. In almost all finetuning experiments over-fitting is observed, and we do an early-stop when performance on the validation set starts to deteriorate. We tune the

---

[1] `http://commoncrawl.org/2016/10/news-dataset-available`

Training	**Test-PPL**		
	DD	**SB**	**CM**
Baseline(from scratch)	24.83	51.14	49.48
MASS+finetune	12.78	28.41	30.25
NS+finetune	11.54	26.37	28.06
NS+WD(θ_{pre})	11.19	26.25	27.80
NS+mix-review	**11.07**	**25.92**	**27.54**

Table 2: Perplexity on test set for different training process on the three dialogue datasets. The datasets are Dailydialogue (DD), Switchboard (SB), and Cornell Movie (CM).

learning rate from $\{10^{-3}, 10^{-4}, 10^{-5}\}$, and report the best model based on validation set perplexity.

5 Experiment Results

In this section, we conduct a set of detailed behavior analysis, characterising how different training strategies change the model's behavior. In particular, we aim to answer the crucial question about whether the model forgets precious language generation skills during standard finetuning, and whether mix-review helps the model remember the skills.

We first present perplexity results for different finetuning methods in Table 2. We observe the big improvement in perplexity (larger than 40%) for the pretrained models comparing to the baseline models trained from scratch. Comparing to MASS, the NS pretraining has more than 7% relative improvement. This confirms our earlier discussion that the model pretrained by NS better utilizes contextual input (which is further verified in Section 5.2). Based on this observation, we focus our analysis below on the NS pretraining.

Comparing to standard finetuning, mix-review further gives solid improvement. The gain is due to its strong regularization effect (which we study in the next three sections). However, the performance gap between mix-review and WD(θ_{pre}) is not significant. We believe the reason is that the benefit (e.g., knowledge transfer) from alleviate the forgetting problem is not be well demonstrated in single-turn response evaluation, because the context is limited to the narrow scope of the specific datasets. We address this concern with multi-turn human evaluation in the next section.

5.1 Behavior Analysis: Knowledge Transfer

As argued in Section 3.1, ideally the model can acquire common-sense (or world) knowledge from the large-scale pretraining data, which will be useful for the downstream open-domain dialogue task. In this section, we design a process to quantify how much knowledge the model has, and use it to monitor how the pretrain-finetune framework changes the model's behavior.

Since the pretraining CCNEWS data is in the public news domain, we expect the model to have knowledge about "big news". So, we utilize the Google trend data of the year 2016[2], which contains 365 trending terms (e.g., iPhone 7, Deadpool), and its corresponding description.

To query whether the model has knowledge of a certain term, we design three news-style and three dialogue-style "trigger templates" to trigger the model to generate responses related to the knowledge term. We collect 10 samples for each trigger, then we compute BLEU score of generated samples against the reference descriptions. We show some examples of trigger inputs in Table 3.

The BLEU scores are shown in Table 4. Note that for the pretrained model we feed news triggers, while for the other dialogue models dialogue triggers are used. We observe that although the finetuned model is more knowledgeable than the baseline model, its score is much lower than the pretrained model. This demonstrates the forgetting problem for the standard finetuning.

On the other hand, we find that mix-review and WD(θ_{pre}) can effectively retain the knowledge acquired during pretraining, giving a much higher BLEU score than the standard finetuned model. Mix-review shows higher BLEU scores than WD(θ_{pre}), demonstrating its superiority in facilitating knowledge retention. We showcase samples from different models in Table 3. To save space, we manually select and show the most related sample out of the 30 samples for each knowledge term. The observations agree with the quantitative results: The standard finetuning loses the detailed information about the knowledge term, and mix-review helps the model retain it. More importantly, the model is able to express the knowledge in a dialogue context.

To further investigate our model's ability to conduct knowledgeable dialogues with users, we use the ParlAI[3] platform to conduct multi-turn dialogue evaluation. For each session, the user will be assigned a random knowledge term from Google Trend, and have a 8-turn dialogue with the model under that topic. Ratings from around 600 dia-

---

[2] `https://www.google.com/intl/en-US/trends/2016records/`

[3] https://parl.ai/

News-style Triggers	**Dialogue-style Triggers**
now, some opinions about X .	what you do think about X ?
let me tell you about X .	please tell me about X .
here's some news about X .	do you have news about X ?
Reference Description: Pokemon first took the world by storm in the mid-90s, doing so once again this year with the release of Pokemon Go.	
NS Pretrained: the game , titled pokemon go : pocket camp , can be played in person ...	
Standard Finetuned: it 's a new game that can be played with kids .	
WD(θ_{pre}): pokemon go , it 's a type of game that only exists in the us .	
Mix-review: pokemon go is a popular mobile game , where you 're expected to catch pokemon .	
Reference Description: Deadpool: The wisecracking antihero, played by Ryan Reynolds in a movie of the same name, became the highest-grossing R-rated film of all time.	
NS Pretrained: ryan reynolds teased his upcoming movie as the character of deadpool .	
Standard Finetuned: it 's a popular movie .	
WD(θ_{pre}): yes , i really like him . he is a very funny character .	
Mix-review: ryan reynolds .	

Table 3: Templates for news or dialogue-style triggers. "X" is to be replaced by specific knowledge terms. They are followed by reference description and model samples for "pokemon" and "deadpool". Note that the pretrained model's sample is from news-style triggers, and the other samples are from dialogue-style triggers.

Training	**Dailydialog**	**Switchboard**	**Cornell**
Pretrained	**BLEU-2 0.347 / BLEU-3 0.153**		
Baseline	0.124/0.007	0.032/0.003	0.081/0.003
NS+finetune	0.162/0.047	0.187/0.052	0.207/0.071
NS+WD	0.226/0.080	0.203/0.070	0.285/0.114
NS+MR	**0.261/0.108**	**0.223/0.079**	**0.396/0.190**

Table 4: Average BLEU-2/BLEU-3 scores for the model's samples w.r.t. the reference description. We highlight the pretrained model's performance for news triggers and the performance of the best model finetuned with dialogue data for dialogue triggers.

Training	**Knowledge**	**Consistency**	**Engaging**
finetune	2.82 ± .06	4.28 ± .06	3.84 ± .05
WD(θ_{pre})	3.18 ± .06	4.60 ± .06	4.18 ± .05
Mix-review	**3.40 ± .05**	**4.75 ± .06**	**4.27 ± .05**

Table 5: AMT rating scores (mean and standard deviation) for multi-turn knowledgeable dialogue evaluation.

logues are collected for each model, and are reported in Table 5. In this evaluation we use the models finetuned on the Dailydialog data, because the nature of that dataset is closet to online chit-chat. It is observed that the model trained by mix-review significantly outperforms WD(θ_{pre}) on knowledge, consistency and engagingness, which agrees well with the results in Table 4. Some dialogue examples are included in Table 7.

5.2 Behavior Analysis: Context Sensitivity

The sensitivity to context is an important property for NLG models. However, as shown by (Sankar et al., 2019), dialogue models trained from scratch typically are not sensitive to artificial distortion in the context input, showing the models have poor utilization of dialogue context. In this section, we repeat their experiments with pretrained or finetuned models.

Training	**Dailydialog**	**Switchboard**
NS Pretrained	**word shuffle +110% / drop +105%**	
Baseline	+12%/+28%	+4%/+5%
MASS+FT	+24%/+48%	+15%/+19%
NS+FT	+41%/+64%	+17%/+21%
NS+WD(θ_{pre})	+26%/+46%	+19%/+25%
NS+MR	**+60%/+108%**	**+19%/+30%**

Table 6: The model's relative PPL drop when word-shuffle/drop is applied to input. "FT" refers to "finetune". To save space, the results on the Cornell Movie dataset is deferred to Appendix D, Table 11.

Following (Sankar et al., 2019), we use two methods to distort the context input:

- **word-drop:** We randomly drop 30% of the words in the context input.
- **word-shuffle:** We randomly shuffle the words in the context input.

We use the relative drop in test-set perplexity to quantify the sensitivity. The results are presented in Table 6, where the result of the pretrained model is also included. First, we observe the baseline model trained from scratch is relatively insensitive to context, which agrees well with Sankar et al. (2019). The model with the standard pretrain-finetune process is much more sensitive, showing that pretraining effectively changes the model's behavior. Comparing to MASS, the NS pretrained model has better utilization of context, which explains its superior performance in PPL.

Somewhat surprisingly, the finetuned dialogue models are much less sensitive to context input than the pretrained model without finetuning. This again

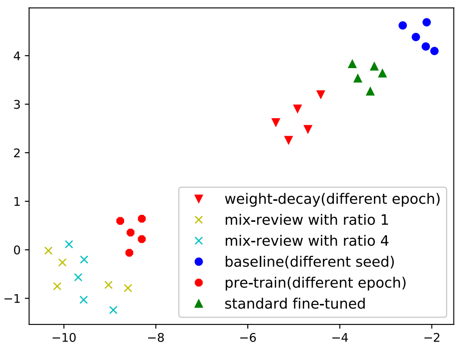


Figure 3: UMAP projection of checkpoints from different training processes.

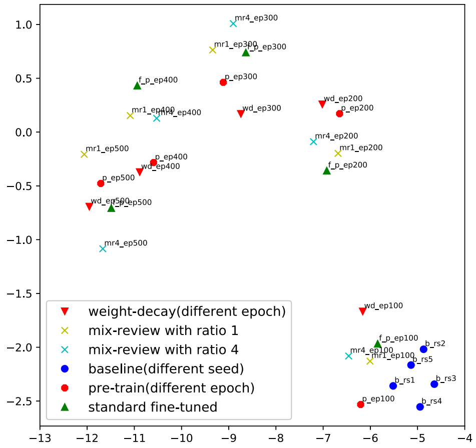


Figure 4: Parameter-space UMAP projection of checkpoints from different training processes.

verifies our worry in Section 3.3 that the model is forgetting some important generation skill during standard finetuning. Further, we find that the mix-review finetuning strategy can effectively alleviate this problem: Its sensitivity is much greater than that of standard finetuning, and is close to the pretrained model.

5.3 Behavior Analysis: Function Space Projection

It is interesting to study models' behavior via function-space 2D projection (Erhan et al., 2010). We collect the model's output distribution on 10k words for the CCNEWS validation set and the Dailydialog validation set[4]. And feed them as input to UMAP (McInnes et al., 2018). We use the default hyper-parameter setting of the python implementation of UMAP. The result is shown in Figure 3. Note that during pretraining of the CCNEWS data, 20 epochs are one entire data pass. We finetune from epoch 100, 200, 300, 400, 500 of the pretraining checkpoints.

We observe that the standard finetuned models are not close to the cluster of the pretrained models, which suggests the models' generative behavior is substantially different from the pretrained ones. Mix-review regularizes the finetuning process by keeping the model's generation behavior close to the pretrained model. These observations agree with our results in Section 5.1 and 5.2. Figure 3 also suggests potential limitations of mix-review and WD(θ_{pre}): mix-review could be too "aggressive" and does not put enough attention on the target task. On the other hand, WD(θ_{pre}) is not strong enough in regularizing the model's generative behavior.

In Figure 4 we show the parameter-space UMAP projection for the same set of models. In this case, the input to UMAP is the concatenation of flattened weight matrices of the transformer model. A key observation is that the finetuned models are typically very close to the starting point (pretrained models). However, as shown in Figure 3, their behavior is very different. This suggests that a parameter-space regularization such as WD(θ_{pre}) could be not very effective for regularizing the model's behavior.

6 Implications and Discussion

The sensitivity to dialogue context and the ability to transfer knowledge from pretraining opens the possibility of a **data-driven** knowledgable chatbot. In Table 7, we show multi-turn and single-turn interaction examples with the model trained by mix-review. For demonstration purpose, we manually select the most interesting response out of 10 samples from the model for the single-turn examples. We observe that the model is able to return interesting responses with the knowledge it acquired from pretraining. Interestingly, it has developed its own "opinions" and can give advice to the user.

Next, we discuss the malicious response problem for open-domain dialogue models. As shown by (He and Glass, 2019a), it is relatively difficult to trigger the dialogue models trained from scratch to

---

[4] It's a concatenation of two long vectors.

User: have you been to china ?	**U:** who is mark zukerberg ?	**U:** tell me about tv show Robot Wars .
Model: yes , i have .	**M:** he is the ceo of facebook .	**M:** it 's about robots fighting in a war .
User: which city did you go ?	**U:** what is that company about ?	**U:** cool ! What season is it on ?
Model: shanghai	**M:** it is a social media company .	**M:** it 's on next month .
User: how do you like it there ?	**U:** do you like him ?	**U:** how many years has it been on for ?
Model: i like it very much	**M:** i think he is a good leader .	**M:** it started in 2003 .
User: i think the weather is not so good .	**U:** my friend works at facebook .	**U:** what channel does it come on ?
Model: but it is nice in spring and fall .	**M:** what does your friend do there ?	**M:** the science channel.
User: when is the independence day ? **Best-of-10:** it 's july fourth.		
User: is coffee good for health ? **Best-of-10:** it may help reduce your risk of certain diseases.		
User: should i do a phd or go to work ? **Best-of-10:** a phd is not enough if you want to learn something new .		
User: oh shut up you . **Worst-of-10:** shut up !		
User: you 're the problem . **Worst-of-10:** i hate you .		
User: i will kill . **Worst-of-10:** i will kill you .		

Table 7: Multi-turn and single-turn examples of the model trained by mix-review on the Dailydialogue dataset. The single-turn examples involve light cherry-picking.

output malicious responses. However, as shown in Table 7, the pretrained models are easily triggered to respond in a malicious way when "provoked". This is because compared to the baseline models, the pretrained models are more sensitive to the contextual input, making them easier to manipulate. This makes the malicious response problem a more relevant issue to solve (He and Glass, 2019b).

Finally, we discuss some limitations of our work. First, the mix-review strategy we proposed is a simple and preliminary attempt to alleviate the forgetting, and its performance is far from perfect. As shown in Appendix C, in a lot of cases, the generation from mix-review is still boring or non-informative. Next, the three datasets considered in this work are open-domain dialogue datasets, and they are not knowledge-intensive. It would be interesting, as future work, to check the forgetting problem for knowledge-grounded datasets such as Topical-chat (Gopalakrishnan et al., 2019).

7 Related Works

Behavior of pretrained NLG Models Recently, multiple works (Radford et al., 2019; Jiang et al., 2020; Roberts et al., 2020; Talmor et al., 2019; Trinh and Le, 2019) have reported that pre-trained language models (LM) have implicitly stored large amounts of "world knowledge" in its parameters, and are able to answer common-sense questions. However, whether the world knowledge is well preserved after finetuning on target task dataset is not discussed.

On the other hand, knowledge-grounded NLG model (Liu et al., 2018; Guu et al., 2020; Zhou et al., 2018) has been an important and exciting research topic. These studies usually involve additional retrieval modules or external knowledge bases to provide the model with relevant information. In contrast to these works, we study whether the model can conduct knowledgeable dialogues by itself.

Forgetting As discussed in Section 3.2, in contrast to the "catastrophic forgetting" problem in sequential learning (Atkinson et al., 2018; Robins, 1995), the performance drop on pretraining data is not necessarily bad for the NLP pretrain-finetune framework, and its implications have not been properly studied. In our analysis, we confirm the "forgetting" of important language generation skills during standard finetuning. The proposed mix-review strategy is similar to the *pseudo-rehearsal* algorithm in sequential learning (Robins, 1995), with the difference being that we assume we still have access to the pretraining data.

8 Conclusion

In this work, we attempt to answer to question of whether during finetuning, the model has forgotten some of the useful NLG skills acquired during large-scale pretraining. Through a set of detailed behavior analysis, we find the answer is, to some extent, yes. For example, the finetuned model fails to give detailed information about some knowledge terms, while the pretrained model can. As a preliminary attempt to alleviate the forgetting problem, we propose the mix-review finetuning method, and find it to be effective.

Our analysis shows that under the surface of the performance boost for standard metrics, large-scale pretraining changes the model's generative behavior in various profound ways. More importantly, the behavior change is influenced by the nature of data itself. For example, we demonstrate that we can discuss news with the dialogue model finetuned by mix-review, even when the target dataset is not about news (Dailydialog). We believe that this opens the possibility of a completely data-driven way to customize a language generator.

References

Craig Atkinson, Brendan McCane, Lech Szymanski, and Anthony V. Robins. 2018. Pseudo-recursal: Solving the catastrophic forgetting problem in deep neural networks. *CoRR*, abs/1802.03875.

Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, and Arthur Szlam. 2019. Real or fake? learning to discriminate machine from human generated text. *CoRR*, abs/1906.03351.

Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *J. Mach. Learn. Res.*, 11:625–660.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-Chat: Towards Knowledge-Grounded Open-Domain Conversations. In *Proc. Interspeech 2019*, pages 1891–1895.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training.

Tianxing He and James Glass. 2019a. Detecting egregious responses in neural sequence-to-sequence models. In *International Conference on Learning Representations*.

Tianxing He and James R. Glass. 2019b. Negative training for neural dialogue response generation. *CoRR*, abs/1903.02134.

Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *CoRR*, abs/1904.09751.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know?

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015.

James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2016. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences of the United States of America*, 114 13:3521–3526.

Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. 2018. Knowledge diffusion for neural dialogue generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1498, Melbourne, Australia. 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.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model?

Anthony Robins. 1995. Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 7:123–146.

Chinnadhurai Sankar, Sandeep Subramanian, Christopher J. Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. *CoRR*, abs/1906.01603.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. *arXiv preprint arXiv:1905.02450*.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics – on what language model pre-training captures.

Trieu H. Trinh and Quoc V. Le. 2019. Do language models have common sense?

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237.

Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. 2017. Recent trends in deep learning based natural language processing. *CoRR*, abs/1708.02709.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4623–4629. International Joint Conferences on Artificial Intelligence Organization.

A Beam-search vs. Top-k Sampling

To compare beam-search with top-k sampling (we set k to 30), we compute diversity metrics for samples from models trained by different procedures (from scratch or pretrained). In particular, we compute bi-gram and tri-gram entropy, and the ratio of the most frequent response and second most frequent response (denoted as max-ratio) (He and Glass, 2019b). The results are shown in Table 8.

We observe that the responses given by top-k sampling are much more diverse than beam-search. Beam-search suffers much from the "generic response" problem (Li et al., 2016), for example, 34% of the responses are "`um - hum`" for Switchboard. Further, in our multi-turn dialogue experiments, beam-search is likely to give repetitive responses. Finally, by manual inspection, we find the sample quality of top-k sampling is not compromised. Due to these observations, we adopt top-k sampling for our models.

B Details on Datasets

Dailydialogue (Li et al., 2017) is a high-quality multi-turn dialog dataset. The language is human-written and less noisy. The dialogues in the dataset reflect our everyday communication and cover various topics about our daily life. The training split has around 11k dialogues (1.3 million words), and both the validation and test splits have 1k dialogues (0.1 million words).

The Switchboard Dialogue Act Corpus[5] is a version of the Switchboard Telephone Speech Corpus, which is a collection of two-sided telephone conversations, annotated with utterance-level dialogue acts. In this work we only use the conversation text part of the data, and select 1.1k dialogues for training (181k sentences / 1.2 million words), 50 dialogues for validation and 50 dialogues for testing.

The Cornell Movie Dialogue Corpus[6] (Danescu-Niculescu-Mizil and Lee, 2011) is a collection of movie scripts. In the processing of the data, we simply regard the whole scripts from a movie as a long dialogue. The training split contains 9k dialogues (4.5 million words), and both the validation and test splits have 180 dialogues (85k words).

C Model Samples

In Table 9, we show more samples from different training procedure, for the three dialogue datasets.

D Supplementary Experiment Results

	Cornell	
Model	**Dialogue Triggers**	**News Triggers**
NS Pretrained	0.245/0.089	**0.347/0.153**
Baseline	0.081/0.003	0.088/0.003
NS+finetune	0.207/0.071	0.207/0.063
NS+WD(θ_{pre})	0.285/0.114	0.202/0.072
NS+Mix-review	**0.396/0.190**	0.212/0.065

Table 10: Average BLEU-2/BLEU-3 scores for the model's samples w.r.t. the reference description. We highlight the pretrained model's performance for news triggers and the performance of the best model fine-tuned with dialogue data for dialogue triggers.

In this section we supplement results that are deferred in the main body due to space limit.

In Table 10 we show the knowledge transfer results for the Cornell Movie dataset.

In Table 11 we show context sensitivity results for the Cornell Movie dataset.

---

[5] `http://compprag.christopherpotts.net/swda.html`

[6] `https://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html`

	Beam Search		**Top-30 Sampling**	
Dataset	**Entropy**	**Max-ratio**	**Entropy**	**Max-ratio**
Dailydialogue	7.44 8.49	1.7% 1.3%	9.04 10.81	0.6% 0.4%
Switchboard	4.96 5.54	34.9% 27.8%	8.47 10.45	8.4% 7.9%
Cornell	6.10 6.56	10.2% 9.9%	8.76 10.54	1.4% 1.1%

Table 8: Average of diversity metrics for models on the three dialogue datasets.

Dailydialogue

Input: why do you think you are qualified for this position ?

Baseline: i can tell you more about your company because your company has a working experience with a company and company company .

NS: i have been working with microsoft for four years . i can make good use of microsoft ’ s word pad and excel excel spreadsheet .

MASS: i think i have a strong sense of responsibility and i have had a high iq score in college . i am applying for some courses of study for some courses if i have a preference .

Mix-review: my background is in computer engineering and i have a strong commercial acumen in business - related application . i am in charge of all the things i like to do here .

WD(θ_{pre}): my background , education , experience as an english professor , and the work ethic are strong .

Input: according to your advertisement , you want an experienced software engineer . i think my background meets the requirement of this position .

Baseline: yes . and you also need to know how to enjoy the work . and , here is a position to develop the company position under the company .

NS: do you know your qualifications ?

MASS: that ’ s a good answer . i can tell you if we have made a decision this afternoon .

Mix-review: i see . do you have some experience as an executive engineer ?

WD(θ_{pre}): in your first year of operating your company , you had an impressive track record of product design , and achieved a high level of technical proficiency .

Switchboard

Input: that ’s going to be a really viable program if it continues to be funded at current levels

Baseline: oh really

NS: um it ’s going to be a lot of fun it certainly is

MASS: sure sure sure

Mix-review: well that that ’s what i ’m saying too that ’s that ’s not gonna be an issue

WD(θ_{pre}): yeah oh sure sure

Input: uh - huh

Baseline: oh okay

NS: and uh i- i think there are some areas that if the government decides to close down that could be a real problem

MASS: and i really feel strongly that we don ’t need the space station because we think there ’s really no space

Mix-review: but uh that

WD(θ_{pre}): i mean it ’s pretty i think it ’s it ’s pretty interesting to me that it ’s uh

---

Cornell Movie

Input: i expect you shortly .

Baseline: how so ?

NS: i ’ll make a reservation .

MASS: i ’ll come with you , doctor .

Mix-review: no , i ’m fine

WD(θ_{pre}): okay .

Input: what shall we do ?

Baseline: we all don ’t know where the hell he is now , but if he talks to me first , i ’ll be back .

NS: nothing .

MASS: we send a cab ! i said i expect you to be fast .

Mix-review: david ’s not very lucid , but he ’s always lucid .

WD(θ_{pre}): stop him .

Input: tea would be nice .

Baseline: i have no desire to go to bed .

NS: but you can ’t come here and have tea too ?

MASS: i ’ve been expecting you , doctor .

Mix-review: do you want me to drink some ?

WD(θ_{pre}): what is it ?

Table 9: Samples of different models on the Dailydialog/Switchboard/Cornell-Movie test-set. The samples are consecutive (input of the next sample is the reference response for the previous one).

Model(Dataset)	PPL(**normal**)	PPL(**word-shuffle**)	PPL(**word-drop**)
NS pretrained(CCNEWS)	17.33	36.56(**+110.96%**)	35.56(**+105.19%**)
Baseline(Cornell)	49.48	50.22(+1.4%)	50.85(+2.7%)
MASS+finetune(Cornell)	30.25	36.50(+20.6%)	36.36(+20.1%)
NS+finetune(Cornell)	28.06	36.88(+31.4%)	34.47(+22.8%)
NS+WD(θ_{pre})(Cornell)	27.80	37.46(**+34.7%**)	35.10(+26.2%)
NS+Mix-review(Cornell)	27.54	36.94(+34.1%)	37.72(**+36.9%**)

Table 11: The model's PPL performance when word-shuffle or word-drop is applied to the context input. On the left we describe what training process is used and on which test set is PPL evaluated. Note that MASS/NS refers to MASS/NS pretraining with standard finetuning.