Rare Tokens Degenerate All Tokens: Improving Neural Text Generation via Adaptive Gradient Gating for Rare Token Embeddings

Anonymous ACL submission

Abstract

Recent studies have determined that the learned token embeddings of large-scale neural language models are degenerated to be anisotropic with a narrow-cone shape. This phenomenon, called the representation degeneration problem, facilitates an increase in the overall similarity between token embeddings that negatively affect the performance of the models. Although the existing methods that address the degeneration problem based on observations of the phenomenon triggered by the problem improves the performance of the text generation, the training dynamics of token embeddings behind the degeneration problem are still not explored. In this study, we analyze the training dynamics of the token embeddings focusing on rare token embedding. We demonstrate that the specific part of the gradient for rare token embeddings is the key cause of the degeneration problem for all tokens during training stage. Based on the analysis, we propose a novel method called, adaptive gradient gating (AGG). AGG addresses the degeneration problem by gating the specific part of the gradient for rare token embeddings. Experimental results from language modeling, word similarity, and machine translation tasks quantitatively and qualitatively verify the effectiveness of AGG.

1 Introduction

Neural language models have been developed with various architectures during recent years (Graves, 2013; Bahdanau et al., 2015; Gehring et al., 2017; Vaswani et al., 2017). Despite the improvement in model architectures, the token embedding training procedures usually share the same process. They process token embeddings as inputs to compute contextualized features and subsequently project the features into a categorical distribution of tokens at the output softmax layer (Merity et al., 2017; Yang et al., 2018; Press and Wolf, 2017). Recent studies have determined that the learned embedding distribution is biased in a common direction, thereby resulting in a narrow cone-shaped anisotropy (Mu et al., 2018; Ethayarajh, 2019; Gao et al., 2019; Biš et al., 2021). This phenomenon, named the representation degeneration problem by Gao et al. (2019), increases the overall cosine similarity between embeddings, and leads to a problem in which the expressiveness of the token embeddings decreases. Therefore, it is difficult for the model to learn the semantic relationship between the tokens and to generate diverse texts with high quality. Existing studies addressing this problem suggest methods that apply post-processing or regularization techniques to all token embeddings based on the observed phenomena owing to the degeneration problem (Mu et al., 2018; Gao et al., 2019; Wang et al., 2019; Wang et al., 2020; Biš et al., 2021). Although these works improves the quality of token embeddings and generated texts, it is still not clear how token embeddings become degenerate during training procedure.

In this study, we conduct empirical studies about training dynamics of token embeddings, focusing on rare token embeddings. By observing the initial training dynamics of token embeddings grouped based on appearance frequency, we hypothesize that the degeneration of the rare token embeddings triggers the degeneration of the embeddings of the remaining tokens. We show that the entire degeneration problem is mitigated by only freezing rare tokens during training, and we demonstrate that the main cause of the entire degeneration problem is the specific part of the gradient for rare token embeddings. This gradient part roles to push away rare token embeddings from the feature vector of the non-rare targets in the current training sample. Based on the analysis, we propose an our method, adaptive gradient gating (AGG). With a dynamic grouping of rare tokens at each training step, AGG solves the entire degeneration problem by gating a specific part of the gradient that is solely about
With a vocabulary of tokens $V$ where $w$ is the output token embedding which roles
is defined as follows.

$$P_\theta(y_t|h_t) = \frac{\exp (h_t w^T_{i(y_t)})}{\sum_{i=1}^N \exp h_t w^T_i},$$  \hspace{1cm} (1)

where $w$ is the output token embedding which roles
the weight of the output softmax layer, and $I(y_t)$ represents the index of token $y_t$. The negative log
likelihood loss for an input and target pair $(x, y)$, $L_{NLL}$ is expressed as follows.

$$L_{NLL} = -\sum_{t=1}^T \log P_\theta(y_t|h_t).$$  \hspace{1cm} (2)

Figure 1: Visualization of token embeddings of language model trained on WikiText-103. Red, green, and blue points represent rare, medium, and frequent groups respectively. (a), (b), (c), (d) present a visualization of each training step.

2.2 Embedding Problems in Neural Language Models

Recent studies on the geometric properties of contextual embedding space have observed that the distribution of embedding vectors is far from isotropic and occupies a relatively narrow cone space (Mu et al., 2018; Liu et al., 2019; Zhou et al., 2019; Ethayarajh, 2019; Biš et al., 2021). Gao et al. (2019) named this phenomenon the representation degeneration problem. This degeneration problem results in an increase in the overall cosine similarity between token embeddings, making it difficult for the model to learn semantic relationships between tokens. Demeter et al. (2020) demonstrated that the norm information of the token embeddings is so dominant that angle information about the feature vector is ignored when calculating the logits in the output layer. Owing to this structural weakness of the embedding space, embeddings with small norms are always assigned with a low probability, which reduces the diversity of the text generated by the model. Although the problem has been theoretically analyzed in several studies, existing methods are based on the observed phenomena as a result of the problem. To mitigate the phenomena observed from the problem, the post-processing of the embedding vectors (Mu et al., 2018; Biš et al., 2021) or regularization terms about the phenomena (Gao et al., 2019; Wang et al., 2019; Wang et al., 2020; Zhang et al., 2020) were introduced. Methodologies based on the training dynamics of the token embeddings concerning the degeneration problem remain subject to study.

Frequency bias in embedding space is another problem. Ott et al. (2018) conducted a comprehensive study on the under-estimation of rare tokens in neural machine translation. Gong et al. (2018) observed that embeddings in the language model were biased towards frequency and proposed an ad-
3 Empirical Study: Token Embedding Training Dynamics led by Rare Tokens

3.1 Initial Training Dynamics of Embeddings

To analyze the training procedure of token embeddings, we train a Transformer language model at the WikiText-103 dataset from scratch. Whole vocabulary tokens are divided into three groups: frequent, medium, and rare groups. Based on the appearance frequency in the training corpus, the 30%, 50%, and 20% tokens are assigned to the frequent, medium, and rare group. We visualize the initial training dynamics of these groups via the projection of the embeddings into 2D, using singular value decomposition (SVD) projection. As illustrated in Figure 1, rare groups degenerate first, as they emerge from the entire embedding distribution. Subsequently, other groups also start to degenerate, following the degeneration of the rare group. Based on this observation, we hypothesize that the degeneration of rare token embeddings induces the degeneration of non-rare token embeddings.

3.2 Rare Tokens Degenerate Non-Rare Tokens

Because Transformer (Vaswani et al., 2017) is representative of the current language models, we adopt the 6-layer Transformer decoder model architecture for an empirical study on the training dynamics of embedding vectors. The model is trained in language modeling task using WikiText-103 dataset (Merity et al., 2018). Experimental details regarding the model and training hyperparameter configurations can be found in the Appendix B. To verify the hypothesis of the previous subsection, we train a model while freezing the rare group token embeddings in their initial states during training, and compare it to the baseline model, where all embeddings are trained with negative log-likelihood loss. In addition, we train the models of various settings relative to freezing steps and examine whether the degeneration of rare token embeddings depends on when training of rare embeddings begins.

The performance of the models is evaluated in two ways; the likelihood and isotropy of token embeddings. Perplexity (Bengio et al., 2003) is adopted to evaluate the performance of the likelihood of the model. To measure the isotropy of the token embedding distribution, we adopt the partition function \( Z(\mathbf{a}) = \sum_{i=1}^{N} \exp \left( \mathbf{w}_i^T \mathbf{a} \right) \) defined in Arora et al. (2016), where \( \mathbf{w}_i \) denotes the embedding vector of token \( i \), and \( \mathbf{a} \) represents a unit vector. Lemma 2.1. in Arora et al. (2016) demonstrate that if the embedding vectors are isotropic, \( Z(\mathbf{a}) \) is approximately constant. Based on this property, we measure the isotropy of an embedding matrix \( \mathbf{W} \) using \( I(\mathbf{W}) \), which is defined as follows.

\[
I(\mathbf{W}) = \min_{\mathbf{a} \in \mathbf{X}} \frac{Z(\mathbf{a})}{\max_{\mathbf{a} \in \mathbf{X}} Z(\mathbf{a})},
\]

where \( I(\mathbf{W}) \in [0, 1] \) and \( \mathbf{X} \) represents the set of eigenvectors of \( \mathbf{W}^T \mathbf{W} \) (Mu et al., 2018; Wang et al., 2020; Biš et al., 2021). Furthermore, we measure

<table>
<thead>
<tr>
<th>Methods</th>
<th>Freq</th>
<th>Med</th>
<th>Rare</th>
<th>Total</th>
<th>Freq</th>
<th>Med</th>
<th>Rare</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>16.58</td>
<td>224.24</td>
<td>813.76</td>
<td>20.77</td>
<td>0.426</td>
<td>0.286</td>
<td>0.198</td>
<td>0.293</td>
</tr>
<tr>
<td>Freeze</td>
<td>16.48</td>
<td>233.92</td>
<td>3017.53</td>
<td>20.78</td>
<td>0.840</td>
<td>0.651</td>
<td>0.831</td>
<td>0.739</td>
</tr>
</tbody>
</table>

Table 1: Perplexity and \( I(\mathbf{W}) \) for each token groups. Lower is better for PPL and higher is better for \( I(\mathbf{W}) \).

(a) freeze until step 7k

(b) freeze until step 18k

(c) freeze until step 29k

Figure 2: Tlot of \( I(\mathbf{W}) \) for rare and frequent groups and average cosine similarity between rare and frequent embeddings when freezing the training of rare tokens until specific training steps.

versarial training scheme to address this problem.

156 3 Empirical Study: Token Embedding

157 Training Dynamics led by Rare Tokens

158 3.1 Initial Training Dynamics of Embeddings

159 To analyze the training procedure of token embeddings, we train a Transformer language model at the WikiText-103 dataset from scratch. Whole vocabulary tokens are divided into three groups: frequent, medium, and rare groups. Based on the appearance frequency in the training corpus, the 30%, 50%, and 20% tokens are assigned to the frequent, medium, and rare group. We visualize the initial training dynamics of these groups via the projection of the embeddings into 2D, using singular value decomposition (SVD) projection. As illustrated in Figure 1, rare groups degenerate first, as they emerge from the entire embedding distribution. Subsequently, other groups also start to degenerate, following the degeneration of the rare group. Based on this observation, we hypothesize that the degeneration of rare token embeddings induces the degeneration of non-rare token embeddings.

160 3.2 Rare Tokens Degenerate Non-Rare Tokens

161 Because Transformer (Vaswani et al., 2017) is representative of the current language models, we adopt the 6-layer Transformer decoder model architecture for an empirical study on the training dynamics of embedding vectors. The model is trained in language modeling task using WikiText-103 dataset (Merity et al., 2018). Experimental details regarding the model and training hyperparameter configurations can be found in the Appendix B. To verify the hypothesis of the previous subsection, we train a model while freezing the rare group token embeddings in their initial states during training, and compare it to the baseline model, where all embeddings are trained with negative log-likelihood loss. In addition, we train the models of various settings relative to freezing steps and examine whether the degeneration of rare token embeddings depends on when training of rare embeddings begins.

162 The performance of the models is evaluated in two ways; the likelihood and isotropy of token embeddings. Perplexity (Bengio et al., 2003) is adopted to evaluate the performance of the likelihood of the model. To measure the isotropy of the token embedding distribution, we adopt the partition function \( Z(\mathbf{a}) = \sum_{i=1}^{N} \exp \left( \mathbf{w}_i^T \mathbf{a} \right) \) defined in Arora et al. (2016), where \( \mathbf{w}_i \) denotes the embedding vector of token \( i \), and \( \mathbf{a} \) represents a unit vector. Lemma 2.1. in Arora et al. (2016) demonstrate that if the embedding vectors are isotropic, \( Z(\mathbf{a}) \) is approximately constant. Based on this property, we measure the isotropy of an embedding matrix \( \mathbf{W} \) using \( I(\mathbf{W}) \), which is defined as follows.

163 \[
I(\mathbf{W}) = \min_{\mathbf{a} \in \mathbf{X}} \frac{Z(\mathbf{a})}{\max_{\mathbf{a} \in \mathbf{X}} Z(\mathbf{a})},
\]

164 where \( I(\mathbf{W}) \in [0, 1] \) and \( \mathbf{X} \) represents the set of eigenvectors of \( \mathbf{W}^T \mathbf{W} \) (Mu et al., 2018; Wang et al., 2020; Biš et al., 2021). Furthermore, we measure

165

166

167

168

169

170

171

172

173

174

175

176

177

178

179

180

181

182

183

184

185

186

187

188

189

190

191

192

193

194

195

196

197

198

199

200

201

202

203

204

205

206

207

208

209

210

211

212

213

214
the relatedness between the rare and frequent group token embeddings to verify that the degeneration of the frequent group follows the degeneration of the rare group. We calculate the average cosine similarity between the rare and frequent group embeddings to measure the relatedness.

Table 1 shows the comparison of the baseline model and the model with frozen rare tokens. We denote the baseline as "MLE" and the freezing method as "Freeze". Surprisingly, the PPL of frequent group tokens and overall $I(W)$ improved by simply not training the rare token embeddings. Figure 2 illustrates the change in $I(W)$ for the frequent and rare token embeddings, including the similarity between frequent and rare token embeddings at various freezing step settings. Whenever the rare token embeddings start to be trained, their $I(W)$ decreases steeply, followed by decreasing $I(W)$ of frequent embeddings and increasing similarities between the frequent and rare embeddings. From the analysis in this subsection, we demonstrate that the entire degeneration problem can be solved by solely handling just rare embeddings during the entire training procedure.

### 3.3 Finding the Primary Cause of the Degeneration Problem: From the Gradient

With $T$ context feature vectors from the training sample, $h_i$ ($i \in [1, T]$), the negative log-likelihood loss gradient for the rare token embedding $w_r$ is calculated as follows.

$$\nabla_{w_r} L_{NLL} = \sum_{y_i = v_r} (p_{r|i} - 1) h_i^{(a)}$$

$$+ \sum_{y_i \notin V_r} p_{r|i} h_j^{(b)} + \sum_{y_k \in V_r} p_{r|k} h_k^{(c)},$$

where $y_i$ denotes the target token for $h_i$, $V_r$ is the rare token vocabulary group, and $p_{r|i}$ represents the conditional probability of token $v_r$ given $h_i$, which is calculated as $[\text{softmax}(h_i W^T)]_r$. We divide the gradient for $w_r$ to 3 parts: (a), (b), and (c) in Eq. 4. Part (a) pulls $w_r$ close to the feature vectors whose target tokens are $v_r$. Part (b) pushes away $w_r$ from the feature vectors whose target tokens are not rare. Part (c) pushes away $w_r$ from the feature vectors whose target tokens are rare. As an extension of the analysis in the previous subsection, we freeze these parts of the gradient with various settings during training to identify the key cause of the degeneration problem. All model and training configurations are the same as in the previous sections, except those to be frozen.

Table 2 presents the results of the experiments in this subsection. We freeze the parts of the gradient for the rare tokens with three settings. Because part (a) is a key component required to train the token embedding to be aligned to the target, all settings activate part (a). We notice that when part (b) is activated (solely freezing part (c)), in general, $I(W)$ decreases and PPL for rare tokens increases almost 10 times compared to when part (b) is frozen. Because PPL and $I(W)$ are improved when activating part (c), part (c) is not negative for the degeneration problem. Consequently, both (b) is the part to be handled as the key component for the degeneration problem. From the analysis in this subsection, we demonstrate that the part of the gradient for rare embeddings that pushes away rare embeddings from non-rare feature vectors is the key cause of the entire degeneration problem of embeddings.

### 4 Method

#### 4.1 Dynamic Rare Token Grouping

To handle the specific part of the gradient for the rare token embeddings studied in the previous section, we need to properly group the rare tokens. A naive approach can be used to group rare tokens based on the appearance frequency of the training corpus, as described in the previous section. How-
ever, this static grouping method is suboptimal because the model is typically trained via mini-batch training. The group of rare tokens that appeared less frequently in recent batch samples is variable in the mini-batch training. Therefore, it is necessary to dynamically group rare tokens based on token appearances in recent batch samples.

To consider the token appearances in recent batch samples, we introduce the token counter memory that remembers the number of the appearances of each token during the previous $K$ training steps. For $K$ memories, $\{m_1, ..., m_K\}$, $m_k \in \mathbb{R}^N$ represents the number of appearances of each token of $N$-size vocabulary at the $t$-th previous training step. Memories are set as zero vectors at the initial stage. At each training step, the token appearance, $a \in \mathbb{R}^N$, is calculated as the sum of all $K$ memories: $a = \sum_{k=1}^{K} m_k$. Based on $a$, we determine whether token $i$ is in the rare token group $V_r$ as follows.

$$
\begin{align*}
    a_i &< \alpha \Rightarrow v_i \in V_r, \\
    a_i &\geq \alpha \Rightarrow v_i \notin V_r,
\end{align*}
$$

where $a_i$ is the $i$-th component of $a$, and $\alpha$ is a hyper-parameter in our method that controls the proportion of rare tokens in the entire vocabulary. In this study, we set $K$ to the number of iteration steps during one epoch of training stage.

### 4.2 Adaptive Gradient Gating for Rare Tokens

After dynamically grouping the rare tokens at each training step, we need to handle a specific part of the gradient for the rare token embeddings, part (b) of Eq. 4, to solve the degeneration problem of all embeddings. To solely control the gradient for rare token embeddings, we introduce a gradient gating method for a parameter $x$. We define $\tilde{x}$ as a tensor whose value is the same as $x$, but detached from the current training graph. This implies that $\tilde{x}$ is considered a constant, hence, gradient about $\tilde{x}$ is not existent. In practice, $\tilde{x}$ can be easily obtained from $x$ using the detach() function of Pytorch (Paszke et al., 2019). With $\tilde{x}$, we can gate the gradient for $x$ as follows.

$$
x_{\text{gated}} = g \odot x + (1 - g) \odot \tilde{x}
$$

$$
\nabla_x f(x_{\text{gated}}) = g \odot \nabla_x f(x),
$$

where $x_{\text{gated}}$ is a new parameter whose value is the same as $x$, and $g \in [0, 1]$ is a gate tensor. When the $x_{\text{gated}}$ is fed to the function $f(\cdot)$ as input, the gradient for $x$ is gated by $g$.

For a context feature vector of the $i$-th position, $h_i$, we introduce a gate vector $g_1 \in \mathbb{R}^N$ as follows.

$$
g_{1k} = \begin{cases} 
    a_k/K & \text{if } v_k \in V_r, v_k \neq y_i \\
    1 & \text{else}
\end{cases} \quad (7)
$$

where $g_{1k}$ denotes a $k$-th component of $g_1$. $g_1$ controls the degree to which rare token embeddings move away from non-rare feature vectors whose targets differ from each rare token embedding. For very rare token embeddings, there is an additional issue. Rare tokens, which are not very rare, are relatively frequent in very rare tokens. Therefore, the embeddings of the very rare tokens can degenerate because of the gradient part that pushes them away from the features whose targets are rare, but not very rare token embeddings. To address this issue, we introduce additional gate vector $g_2 \in \mathbb{R}^N$ as follows.

$$
g_{2k} = \begin{cases} 
    \min(a_k/\bar{a}_r, 1) & \text{if } v_k \in V_r, v_k \neq y_i \\
    1 & \text{else}
\end{cases} \quad (8)
$$

where $g_{2k}$ is the $k$-th component of $g_2$ and $\bar{a}_r$ is the mean of $a_r$ where $r \in V_r$. $g_2$ controls the degree to which very rare token embeddings move away from rare or very rare feature vectors whose targets differ from each very rare token embedding. To calculate the loss of $h_i$, we calculate three logits, $z_i^0, z_i^1$, and $z_i^2$, as follows.

$$
z_i^0 = h_i \hat{W}^T
$$

$$
z_i^l = g_l \odot h_i \hat{W}^T + (1 - g_l) \odot \tilde{h_i} \hat{W}^T
$$

where $W$ denotes an embedding matrix, and $l = 1, 2$. Because our method solely handles the gradient for embeddings, we calculate $z_i^0$ for a gradient about $h_i$, which does not need to be gated. Finally, the negative log-likelihood loss for $i$-th position $L_i$ is computed as follows.

$$
L_i = -\log p_{(y_i)}^0 | i
$$

$$
- \mathbb{1}(y_i \notin V_r) \log p_{(y_i)}^1 | i
$$

$$
- \mathbb{1}(y_i \in V_r) \log p_{(y_i)}^2 | i
$$

where $p_{(y_i)}^m | i = [\text{softmax}(z_i^m)] / l(y_i)$ with $m=0, 1, 2$ and $\mathbb{1}(\cdot)$ denotes the Indicator function. Gradient for rare token embeddings is computed to:

$$
\nabla_w L_i = \begin{cases} 
    (p_{r_i} - 1) h_i & \text{if } y_i = v_r \\
    g_{1r_{p_{r_i}}} h_i & \text{if } y_i \notin V_r \\
    g_{2r_{p_{r_i}}} h_i & \text{else}
\end{cases} \quad (11)
$$
where \( p_{rl} = [\text{softmax}(z_t^{m_m})]_r \), whose value is irrespective of \( m \). Eq. 12 demonstrates that \( L_i \) passes gradients gated by \( g_1, g_2 \) to rare token embeddings, which is consistent with our intent. Derivation of Eq. 12 is provided in Appendix A.

## 5 Experiments

We evaluate our method on various tasks including language modeling, word similarity, and machine translation. In the language modeling task, we focus on verifying the diversity of the generated texts. We test the learning of the semantic relationships between tokens on the word similarity task. Finally, we evaluate the quality of generated texts on the machine translation task. For all the experimental results below, we adopt the state-of-the-art model architecture as a baseline to properly demonstrate the effectiveness of our method. Every detail on the experiment, such as model hyper-parameters and training configurations, regard the reproducibility are provided in Appendix B.

### 5.1 Language Modeling

**Setting** We conduct experiments using WikiText-103 dataset, which is a significantly large dataset for language modeling task with approximately 103M words and 260K vocabulary size (Merity et al., 2018). Texts in the dataset are preprocessed based on the byte-pair encoding (Sennrich et al., 2016). We adopt the GPT-2 medium architecture (Radford et al., 2019), which comprises 24 Transformer decoder layers as a baseline model. Because our method is about learning tokenembeddings, we train the models from scratch for a maximum of 50k iterations and evaluate them based on the perplexity of the validation set. For hyper-parameter searching, we select \( \alpha \in \{0.01, 0.02, 0.03, 0.04, 0.05\} \) for AGG method on the language modeling task. The hyper-parameter sensitivity for the AGG are given in Appendix D.

We use three quantitative metrics to evaluate our method: Perplexity, Uniq, and \( I(W) \). Related to the likelihood of generated texts, Perplexity quantifies the prediction difficulty over the next token. Uniq (Welleck et al., 2020) quantifies the number of unique next-token predictions, measuring the token diversity. As described in section 3, \( I(W) \) measures the isotropy of the token embedding space.

**Results** We present our results for the testset in Table 3. We denote the baseline method as 'MLE' and our method as 'AGG'. We measure Perplexity and Uniq for each token group defined in Section 3. As presented in Table 3, AGG improves the overall metrics for the medium and rare groups while maintaining performance for the frequent token group. In particular, for the rare group, the Perplexity score decrease significantly and the number of unique predictions surpasses the human distribution. The \( I(W) \) for all token embeddings increased over 2 times the baseline. Qualitative studies on semantic alignment between tokens are provided in Appendix E. We also show examples of generated texts in Table 5 and more in Appendix F.

### Communication

Neural text degeneration problem is another problem in neural text generative models, where the model generates texts that are less

<table>
<thead>
<tr>
<th>Methods</th>
<th>Freq</th>
<th>Med</th>
<th>Rare</th>
<th>Total</th>
<th>Freq</th>
<th>Med</th>
<th>Rare</th>
<th>Total</th>
<th>I(W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>13.30</td>
<td>146.47</td>
<td>438.67</td>
<td>15.51</td>
<td>9107</td>
<td>3945</td>
<td>91</td>
<td>13143</td>
<td>0.377</td>
</tr>
<tr>
<td>AGG</td>
<td>13.35</td>
<td>146.44</td>
<td>75.39</td>
<td>15.51</td>
<td>9105</td>
<td>4287</td>
<td>345</td>
<td>13737</td>
<td>0.813</td>
</tr>
<tr>
<td>Human</td>
<td>10844</td>
<td>7146</td>
<td>300</td>
<td>18920</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Experimental results for each token group in WikiText-103 language modeling task comparing MLE baseline and AGG.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Freq</th>
<th>Med</th>
<th>Rare</th>
<th>Total</th>
<th>Freq</th>
<th>Med</th>
<th>Rare</th>
<th>Total</th>
<th>I(W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>UL</td>
<td>14.05</td>
<td>146.47</td>
<td>438.67</td>
<td>15.51</td>
<td>9107</td>
<td>3945</td>
<td>91</td>
<td>13143</td>
<td>0.377</td>
</tr>
<tr>
<td>UL + AGG</td>
<td>14.17</td>
<td>146.44</td>
<td>75.39</td>
<td>15.51</td>
<td>9105</td>
<td>4287</td>
<td>345</td>
<td>13737</td>
<td>0.813</td>
</tr>
<tr>
<td>Human</td>
<td>10844</td>
<td>7146</td>
<td>300</td>
<td>18920</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Experimental results for each token group in WikiText-103 language modeling task comparing UL and UL+AGG.
Method | Texts | Uniq
--- | --- | ---
**Prefix** | No. 20 Squadron is a Royal Australian Air Force (RAAF) support squadron. Coming under the control of No. 96 Wing, it is responsible for the management of the airfield at RAAF Base Woomera, South Australia. The squadron | 15
**MLE** | is responsible for air defence, air defence, and air defence, as well as air defence, aerial reconnaissance, and air defence. It is also responsible for air defence, air defence, and air defence, as well as air defence, aerial reconnaissance, and air defence. | 48
**AGG** | was established in October 1943 at Townsville, Queensland, under the command of Group Captain Paddy Heffernan. It was initially based at Townsville, Queensland, under the control of No. 9 Operational Group, which controlled all air bases in New South Wales. It was renamed No. 1 Mobile Fighter Sector in April 1944. | 48

Table 5: Generated texts on the Wikitext-103 test set and uniq tokens for each texts. 50 bpe tokens are given as prefix and the models are to generate the continuation of 100 next bpe tokens.

likely to match human word distributions. Existing methods focus on the diversity of the generated texts by adding an auxiliary loss to the original negative log-likelihood loss (Welleck et al., 2020). Although Welleck et al. (2020) and AGG attempts to address the same problem about diversity, AGG can be compatible with the existing method in the text degeneration problem because AGG does not alter the form of the loss function in MLE training. Table 4 presents the results of the experiments about fusion of unlikelihood training (Welleck et al., 2020) and AGG. We denote the unlikelihood training as UL. From table 4, we notice that when UL and AGG are fused, it produces a synergistic effect that exceeds the gain of each for the baseline. This indicates that AGG is compatible with methods that address other problems in text generation.

### 5.2 Word Similarity

**Setting** We evaluate the semantic relationship between tokens for AGG and the baseline with four word similarity datasets: MEN, WS353, RG65, and RW (Bruni et al., 2014; Agirre et al., 2009; Rubenstein and Goodenough, 1965; Luong et al., 2013). Methods are tested whether the similarity between the given two words in the embedding space is consistent with the ground truth, in terms of Spearman’s rank correlation. We adopt cosine distance to compute the similarity between embeddings. We use the same models trained on language modeling tasks with the WikiText-103 dataset for the word similarity task.

**Results** Table 6 presents the result obtained from the evaluation of the word similarity task. From this table, it can be observed that our method outperforms the baseline on overall datasets. Although AGG handles only training of rare tokens, the semantic relationships between all tokens are also well learned.

### 5.3 Machine Translation

**Setting** We utilize a dataset from standard WMT 2014 containing 4.5M English-German sentence pairs. The source and target sentences are encoded by 37K shared tokens based on byte-pair encoding (Sennrich et al., 2016). We adopt the two versions of Transformer (Vaswani et al., 2017) as the baseline model for applying our method: base and big. The model configuration is the same as that proposed in Vaswani et al. (2017). To evaluate the quality of the generated texts, we measure BLEU score (Papineni et al., 2002), which is standard metric for machine translation task.

**Results** Table 7 presents a comparison of our method and other methods in terms of the BLEU score. Our method achieves 1.4 and 1.41 BLEU scores on the task of WMT14 En→De machine translation.
score improvements on the machine translation task for the base and big baseline models. In addition, our method is better than all other previous works in handling the representation degeneration problem that reported BLEU scores in the same tasks. These results demonstrate the effectiveness of AGG in the quality of the generated texts. Qualitative study about cross-lingual semantic alignment between tokens of the source and target languages is provided in Appendix E.

6 Analysis of AGG

6.1 Visualization

Figure 3 (a) and (b) present the visualizations of the embedding space of baseline MLE and our method. In the figure, applying the AGG method restores the isotropy of the token embedding space. In addition, we observe that the regions occupied by each token group are not disjoint when applying AGG. For baseline, the regions occupied by rare group and the frequent group are disjoint, which is referred as the frequency bias problem of embeddings (Gong et al., 2018). From the analysis of the visualization of the embedding space, we notice that the manipulating the training of the rare token embeddings can alleviate the frequency bias problem. Figure 3 (c) presents the plot of the normalized singular value of embedding matrix for MLE and AGG. Slowly decaying singular values of AGG demonstrate an isotropic distribution of the embedding space.

6.2 Ablation Study

In our method, AGG, we introduce two gate vectors, $g_1$ and $g_2$, to handle the gradient for rare and very rare token embeddings. We conduct experiments on these gate vectors. Table 8 presents the results of the ablation studies compared with the MLE and AGG. When $g_1$ is excluded from AGG (denoted as 'no $g_1$'), Uniq and $I(W)$ decreased significantly, because $g_1$ is the key component for the gradient gating. When $g_2$ is excluded from AGG (denoted as 'no $g_2$'), Uniq and $I(W)$ slightly decrease. Accordingly, we notice that $g_2$ is important for the gating of gradients for the very rare token embeddings. The analysis of rare token grouping is also important for our study, and it can be found in Appendix C.

7 Conclusion

In this study, we analyzed the training dynamics of the token embeddings concerning the representation degeneration problem of the learned embeddings, focusing on the rare tokens. Based on the analysis, we propose an adaptive gradient gating method that solves the problem by solely handling the training for rare token embeddings. Experiments and qualitative studies in various tasks of text generation demonstrate the effectiveness of our method. AGG is orthogonal to the existing method in the neural text degeneration problem, which means it can be compatible to fuse AGG and the existing methods in other problems. Thus for future work, we would like to extend our method to other regions using token embeddings, such as text summarization, and classification.

<table>
<thead>
<tr>
<th>Method</th>
<th>PPL</th>
<th>Uniq</th>
<th>$I(W)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>15.51</td>
<td>13143</td>
<td>0.377</td>
</tr>
<tr>
<td>AGG</td>
<td>15.51</td>
<td>13737</td>
<td>0.813</td>
</tr>
<tr>
<td>no $g_1$</td>
<td>15.48</td>
<td>13018</td>
<td>0.367</td>
</tr>
<tr>
<td>no $g_2$</td>
<td>15.51</td>
<td>13682</td>
<td>0.701</td>
</tr>
</tbody>
</table>

Table 8: Ablation study on gating vector of AGG.
References


L. Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2020. Improving neural language generation with spectrum control. In ICLR.


Z. Zhang, Chongming Gao, Cong Xu, Rui Miao, Qinli Yang, and Junming Shao. 2020. Revisiting representation degeneration problem in language modeling. In EMNLP.

A Derivation of the gradient of AGG loss w.r.t. rare token embedding

We follow the same notation as in the main paper.

Before we write the derivation of the gradient about rare token embedding \( w_r \), we write the gradient of \( f(\tilde{w}_j) \) and \( (z^j)_l \) about \( w_r \), where \( f(\tilde{w}_j) \) is the function of \( \tilde{w}_j \) with \( j = 1, \ldots, N \) and \( (z^j)_l \) is a \( j \)-th component of \( z^j_l \) with \( l = 0, 1, 2 \) as follows.

\[
\nabla_w f(\tilde{w}_j) = \nabla_{\tilde{w}_j} f(\tilde{w}_j) \circ \nabla_w \tilde{w}_j
\]

\[
= \begin{cases} 
    g_l \nabla_w \tilde{h}_l^T + 0 & \text{if } j = r \\
    0 & \text{else}
\end{cases}
\]

\[
= \begin{cases} 
    g_l h_l & \text{if } j = r \\
    0 & \text{else}
\end{cases}
\]

\[
=: h_i = \tilde{h}_i \text{ in terms of value.}
\]

Considering the case of \( y_i \notin V_r \), AGG negative log-likelihood loss for the \( i \)-th position of token generation, \( L_i^{\text{AGG}} \) is written as follows.

\[
L_i^{\text{AGG}} = - \log p_i^{0}(y_i)|i| - \log p_i^{1}(y_i)|i|
\]

Then gradient of \( L_i^{\text{AGG}} \) about \( w_r \) is written as follows.

\[
\nabla_w L_i^{\text{AGG}} = - \nabla_w \log p_i^{0}(y_i)|i| - \nabla_w \log p_i^{1}(y_i)|i|
\]

\[
= - \nabla_w \log p_i^{0}(y_i)|i| - 0
\]

\[
=: \log p_i^{0}(y_i)|i| \text{ is a function of } \tilde{w}_r.
\]

\[
= - \frac{1}{p_i^{0}(y_i)|i|} \nabla_w p_i^{1}(y_i)|i|
\]

\[
= - \frac{1}{p_i^{0}(y_i)|i|} \sum_{j=1}^{N} \nabla(z^j)_l p_i^{1}(y_i)|i| \cdot \nabla_w(z^j)_l
\]

\[
=: p_i^{1}(y_i)|i| \text{ is a function of } (z^j)_l, j = 1, \ldots, N.
\]

\[
= - \frac{1}{p_i^{0}(y_i)|i|} \nabla(z^j)_l p_i^{1}(y_i)|i| \cdot \nabla_w(z^j)_l
\]

\[
\text{(By Eq. 13.)}
\]

As \( p_i^{1}(y_i)|i| = [\text{softmax}(z^j_l)]_l \),

\[
\nabla(z^j)_l p_i^{1}(y_i)|i| = - p_i^{1}(y_i)|i| p_r^1|l|
\]

Thus, \( \nabla_w L_i^{\text{AGG}} \) is computed as follows.

\[
\nabla_w L_i^{\text{AGG}} = - \frac{1}{p_i^{0}(y_i)|i|} \nabla(z^j)_l p_i^{1}(y_i)|i| \cdot \nabla_w(z^j)_l
\]

\[
\text{(By Eq. 15.)}
\]

\[
= p_r^1|l| \cdot \nabla_w(z^j)_l
\]

\[
= g_r p_r^2|l| h_l
\]

\[
\text{(By Eq. 13.)}
\]
Considering the remained case of $y_i = v_r$, since $y_i \in V_r$, $L_{i}^{\text{AGG}}$ is same as the second case, and derivation process of $\nabla_w, L_{i}^{\text{AGG}}$ shares the same process with Eq. 19. As $I(y_i) = r$,

$$\nabla_{(z^2_i, r)}^{2}P_{I(y_i)|i}^{2} = \nabla_{r}(z^2_i, r)$$

Thus, $\nabla_w, L_{i}^{\text{AGG}}$ is computed as follows.

$$\nabla_w, L_{i}^{\text{AGG}} = -\frac{1}{p_I^{2}(y_i)|i} \nabla_{w,} (z^2_i, r)$$

(By Eq. 22.)

$$= -(1 - p_I^{2}(y_i)|i) \cdot \nabla_{w,} (z^2_i)$$

= $-g_{2r}(1 - p_I^{2}(y_i)|i)\tilde{h}_i$

(By Eq. 13.)

$$= (\tilde{p}_{r|i}^2 - 1)\tilde{h}_i$$

(: $I(y_i) = r$ and $g_{2r} = 1$ if $I(y_i) = r$.)

As $p_{r|i} = p_{r|i}^{m}$ with $m = 0, 1, 2$ in terms of value, we finally write $\nabla_w, L_{i}^{\text{AGG}}$ as follows.

$$\nabla_w, L_{i} = \begin{cases} (\tilde{p}_{r|i} - 1)\tilde{h}_i & \text{if } y_i = v_r \\ g_{1Pr|i}\tilde{h}_i & \text{if } y_i \notin V_r \\ g_{2Pr|i}\tilde{h}_i & \text{else,} \end{cases}$$

B Experimental Details

In this section, we present the details of the experiments in main page. All the experiments were conducted with a single GPU on our machine (GPU: NVIDIA A40). For each task in the experiments, we use the same model architecture and train it with different objectives(i.e., MLE, AGG, UL). The hyper-parameters used for different training methods in the same task are exactly same. The detailed hyper-parameters are described in Table 10.

C Ablation Study about Rare Token Grouping

In this sections, we present the analysis about rare token grouping method of AGG. Figure 4 presents the size of the rare token group during initial 1k training steps when the model is trained with WikiText-103 dataset. As presented in the figure, rare group size fluctuate wildly at the initial training stage. We expect for this grouping method to determine an optimal rare token group for the current training step. Table 9 presents the results of ablation study about dynamic grouping. To except dynamic grouping from AGG, we fixed the rare token group after 1 epoch. For this static grouping AGG method, Next-token diversity(Uniq) and the isotropy of the token embedding space($I(W)$) perform worse than dynamic grouping AGG.

D Hyperparameter Sensitivity

In this sections we show how the metrics used on language modeling task change with the hyper-parameter $\alpha$ in Figure 5. As presented in this figure, Perplexity and Uniq score typically increase with bigger $\alpha$. Isotropy of the embedding space is the best when $\alpha = 0.03$, which is the main reason to be selected. Destruction of isotropy when the rare token group becomes big is another study point about the AGG method.

E Qualitative Study about Semantic Alignments between Tokens

In this section, we present qualitative studies about semantic alignments between tokens for language modeling and machine translation tasks. We select three rare token from each datasets: "homepage", "Werewolf", and "policymakers" for WikiText-103 dataset, and "optimum", "criminal", and "happiness" for WMT14 En→De dataset. For each rare token, we extract the top-5 nearest neighbor token predicted by the cosine distance between token em-
Hyperparameter | Empirical Study | Language Modeling | Machine Translation
---|---|---|---
# of layers | 6 | 24 | 6-6 6-6
Hidden dimension | 512 | 1024 | 512 1024
Projection dimension | 2048 | 4096 | 2048 4096
# of heads | 8 | 16 | 8 16
Dropout | 0.1 | 0.1 | 0.1 0.3
Vocabulary size | 44256 | 44256 | 40624 40624
# of parameters | 42M | 358M | 65M 218M
Learning rate | \(7 \cdot 10^{-4}\) | \(7 \cdot 10^{-4}\) | \(1 \cdot 10^{-3}\) \(1 \cdot 10^{-3}\)
Max tokens per batch | 32k | 32k | 64k 64k
Maximum training steps | 40k | 50k | 190k 190k
Warmup steps | 4k | 4k | 4k 4k
Optimizer | Adam | Adam | Adam Adam
Weight decay | 0.01 | 0.01 | 0.01 0.01
\(\alpha\) for AGG | — | 0.03 | 0.08 0.08
\(\alpha\) for UL | — | 1.0 | — —

Table 10: Model configurations and training hyper-parameters for all experiments conducted in the main page. For word similarity task, the model trained on language modeling task are evaluated for word similarity datasets.

Figure 5: Hyper-parameter(\(\alpha\)) sensitivity of AGG in the language modeling task on Wikitext-103 dataset.

beddings. Compared with baseline MLE method, AGG shows significant improvement to train semantic alignments for rare tokens. From Table 11, we notice that the rare tokens trained with AGG are semantically well aligned and not biased about token frequency. Table 12 demonstrates that token embeddings trained with AGG also learn the cross-lingual semantic alignments between target language tokens.

F Examples

We present additional generated text samples from the model trained on language modeling task in Table 13. From the table, we notice that the model trained with AGG generates more diverse and high quality text than the baseline.
### Table 11: Top-5 nearest neighbors of each rare tokens in WikiText-103 dataset. Performance of AGG method is compared with the baseline MLE method. Red color denotes the rare tokens among neighbors.

<table>
<thead>
<tr>
<th>homepage</th>
<th>Werewolf</th>
<th>policymakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>AGG</td>
<td>MLE</td>
</tr>
<tr>
<td>BOX</td>
<td>website</td>
<td>ASUS</td>
</tr>
<tr>
<td>inbox</td>
<td>webpage</td>
<td>riet</td>
</tr>
<tr>
<td>livestream</td>
<td>blog</td>
<td>480</td>
</tr>
<tr>
<td>namespace</td>
<td>Tumblr</td>
<td>nuclear</td>
</tr>
<tr>
<td>hashes</td>
<td>websites</td>
<td>ATCH</td>
</tr>
</tbody>
</table>

### Table 12: Top-5 nearest neighbors of each rare source tokens in WMT14 En→De dataset. Performance of AGG method is compared with the baseline MLE method. The symbol @@ stands for sub-word tokenization of the dataset. The symbol * denotes the synonym token of the target language.

<table>
<thead>
<tr>
<th>optimum</th>
<th>criminal</th>
<th>happiness</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>AGG</td>
<td>MLE</td>
</tr>
<tr>
<td>therto</td>
<td>optimal</td>
<td>Criminal</td>
</tr>
<tr>
<td>ratory</td>
<td>optimale*</td>
<td>criminals</td>
</tr>
<tr>
<td>consultan@ @</td>
<td>optimalen*</td>
<td>perpetr@ @</td>
</tr>
<tr>
<td>sofar</td>
<td>maximum</td>
<td>secution</td>
</tr>
<tr>
<td>protection@ @</td>
<td>Optim@ @</td>
<td>xious</td>
</tr>
</tbody>
</table>
### Table 13: Generated texts on the Wikitext-103 test set and uniq tokens for each texts. 50 bpe tokens are given as prefix and the models are to generate the continuation of 100 next bpe tokens.

<table>
<thead>
<tr>
<th>Method</th>
<th>Texts</th>
<th>Uniq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prefix</td>
<td>A Company , 2nd Engineer Combat Battalion , moved to the south side of the Yongsan - Naktong River road ; D Company of the 2nd Engineer Battalion was on the north side of the road . Approximately 2 miles ( 3</td>
<td>50</td>
</tr>
<tr>
<td>MLE</td>
<td>. 2 km ) north of the river , the 2nd Engineer Combat Battalion was on the south side of the road . The 2nd Engineer Combat Battalion was on the south side of the road .</td>
<td>22</td>
</tr>
<tr>
<td>AGG</td>
<td>. 2 km ) north of the river , the 2nd Engineer Combat Battalion was attacked by a small force of North Korean People ' s Volunteer Army ( NK ) troops . The North Koreans were able to penetrate the UN lines and capture the town .</td>
<td>43</td>
</tr>
<tr>
<td>Prefix</td>
<td>The twelve 45 - calibre 35 . 6 cm ( 14 . 0 in ) Type 41 guns of the Ise class were mounted in three pairs of twin - gun , superfiring turrets . Numbered one through</td>
<td>50</td>
</tr>
<tr>
<td>MLE</td>
<td>two , the turrets were placed in superfiring pairs , one ahead of the other . The turrets were placed in superfiring pairs , with the turrets superfiring over the wing turrets . The turrets were placed in superfiring pairs , with the turrets superfiring over the wing turrets . The turrets were placed in superfiring pairs , with the turrets superfiring over the wing turrets . The turrets were placed in superfiring</td>
<td>19</td>
</tr>
<tr>
<td>AGG</td>
<td>live from front to rear , the guns had a firing arc of 260 ° . They fired a 45 . 3 kg ( 99 lb ) projectile at a muzzle velocity of 835 m / s ( 2 . 740 ft / s ) to a range of 16 , 000 m ( 52 , 000 yd ) . They had a maximum elevation of 30 degrees and fired a 36 . 3 kg ( 77 lb ) projectile at a muzzle velocity of 835</td>
<td>55</td>
</tr>
<tr>
<td>Prefix</td>
<td>Despite the expensive reconstructions , both vessels were considered obsolete by the eve of the Pacific War , and neither saw significant action in the early years of the war . Following the loss of most of the IJN ' s large aircraft carriers during the Battle of Midway , the IJN was forced to rely on the United States Navy for aircraft carriers and aircraft carriers to support operations in the Pacific . The IJN ' s aircraft carriers were the first to be equipped with the new Mark 4 torpedo and the IJN ' s aircraft carriers were the first to be equipped with the Mark 4 torpedo . The IJN ' s aircraft carriers were the first to be equipped with the Mark 4 torpedo , and the IJN ' s aircraft carriers</td>
<td>37</td>
</tr>
<tr>
<td>MLE</td>
<td>of Midway , the IJN decided to modernize its fleet and modernize its fleet . The IJN ' s new ships were designed to be capable of operating at speeds of up to 30 knots ( 56 km / h ; 35 mph ) , and were fitted with a new bow section . The ships were designed to carry 1 , 000 t ( 980 long tons ; 1 , 100 short tons ) of fuel oil , and were fitted with a pair of aircraft catapults</td>
<td>63</td>
</tr>
<tr>
<td>AGG</td>
<td>Amos was born in Macclesfield , Cheshire and was a member of the Crewe Alexandra academy until he was released at the age of 10 . While at Crewe , he also played for another local team , Bollington United , as a centre . He was a member of the team that won the FA Youth Cup in 1989 , and was a member of the team that won the FA Youth Cup in 1990 . He was a member of the team that won the FA Youth Cup in 1990 , and was a member of the team that won the FA Youth Cup in 1992 .</td>
<td>24</td>
</tr>
<tr>
<td>MLE</td>
<td>- back . He was a member of the team that won the FA Youth Cup in 1989 , and was a member of the team that won the FA Youth Cup in 1990 . He was a member of the team that won the FA Youth Cup in 1990 , and was a member of the team that won the FA Youth Cup in 1992 .</td>
<td>52</td>
</tr>
<tr>
<td>AGG</td>
<td>- back . He was signed by Crewe Alexandra in July 2006 , and made his debut for the club in a 2 - 1 win over Rotherham United in the League Cup on 18 August 2006 . He was loaned out to Rotherham for the rest of the 2006 - 07 season , before being released at the end of the season .</td>
<td>52</td>
</tr>
</tbody>
</table>