We thank the reviewers for their careful reading and their many useful comments. If we could, we would like to respond to each comment but the regulation of the conference limits the amount of our response. Thus, we focus on main concerns from reviewers and answer them.

**Additional comparisons:** We appreciate advices from reviewers **#1**, **#3**, and **#4** on additional comparisons to further indicate the effectiveness of the proposed method, ALONE. We are also intrigued to conduct such comparisons (e.g., training time, embedding analysis, and so on).

**Reviewer #4** are wondering whether ALONE can be used for language models like BERT/GPT. BERT/GPT consists of the Transformer architecture that we used in our experiments. We indicated that the combination of Transformer and ALONE works well in widely used machine translation and summarization datasets. Those results imply that ALONE can be introduced in BERT/GPT and reduce the parameters related to their embeddings without negatively affecting the performance.

**Word embedding reconstruction experiment:** We consider that reviewers **#1** and **#4** have some concerns about the experiment on word embedding reconstruction. The motivation of this experiment is to investigate whether ALONE has a similar expressiveness to the conventional word embeddings before real applications as described in Section 3. As pointed out by reviewers, we can train ALONE on a raw corpus based on the objective function of GloVe (or other objectives such as skip-gram) and use the whole vocabulary in mimicking but we selected pre-trained GloVe embeddings of 5k words as a target of mimicking to shorten the training time. We believe that it is more important to investigate whether ALONE can reduce the parameter size related to embeddings in the real applications (Sections 3.2 and 3.3) to indicate the usefulness of ALONE.

As described in Section 3, we trained ALONE with an end-to-end manner in the experiments on machine translation and summarization. In other words, we didn't use pre-trained embeddings and trained ALONE with Transformer jointly from random initialization in contrast to prior studies such as Shu and Nakayama [2018] in these experiments.

**Lack of other compression baselines: Reviewer #2** considers that we didn't compare existing methods to reduce the parameter size related to embeddings but we compared DeFINE (Mehta et al. [2020]) and the factorized embedding approach. As described in Section 3.2, the total parameter size of Transformer+DeFINE is larger than ours. In WMT En-De dataset, it is easy to achieve better performance for a model that has a large amount of parameters because Transformer (big) outperforms Transformer (base) (Vaswani et al. [2017]). Thus, we would like to emphasize that Transformer+ALONE achieved better performance than Transformer+DeFINE although Transformer+ALONE had a disadvantage in the parameter size. In addition, **Reviewer #2** pointed out that the factorized embedding approach is toy but this approach is used in the recent major work, ALBERT (Lan et al. [2020]), to reduce the embedding parameter size. Therefore, we compared ALONE with approaches in recent studies.

**Reviewer #2** required the comparison with Shu and Nakayama [2018]. Indeed, they conducted experiments on machine translation but their approach needs multiple training steps and additional parameters when we introduce it into neural encoder-decoder models because their approach compresses 'pre-trained' embeddings. In fact, the training of Shu and Nakayama [2018] consists of 3 steps in experiments on machine translation in their paper (training NMT model, compressing embeddings, and re-training NMT model). In contrast, ALONE (and other compared methods in our experiments) can be trained with an end-to-end manner based on the objective functions of the application tasks. Thus, it is difficult to conduct a fair comparison because the training paradigms of ours and theirs are different. In other words, we can combine ALONE with theirs if we have a large amount of time to construct a model.

**Definition of additional parameters: Reviewer #2** might be confused about the definition of "additional parameters". As described in the line 43, "additional parameters" is the parameters required only during the training phase. For example, the approach of Shu and Nakayama [2018] learns the mapping between primitive embeddings and words, and deletes the parameters related to the mapping after the training. Moreover, their method requires pre-trained embeddings. We call these parameters "additional parameters". Thus, the parameters of FFN in ALONE are not "additional parameters", and the parameter sizes in Tables 2 and 3 include them.

**Compression rate of the whole parameter size:** We agree with reviewers **#2** and **#4** that it is also important to report the compression rate of the whole parameter size in neural encoder-decoder models. However, since this study addresses reducing the number of parameters related to embeddings, we consider that it is the most important to report the embedding parameter size. The previous studies such as Shu and Nakayama [2018] and Chen et al. [2018] also reported the number of parameters related to embeddings only (the reported "total size" in Shu and Nakayama [2018] includes embeddings only). In addition, since ALONE is independent from an encoder-decoder architecture, we can combine ALONE with any existing approach to reduce the parameter sizes of neural encoder-decoders. For example, we can reduce the parameter size with cross-layer parameter sharing used in ALBERT (Lan et al. [2020]) but the reduction is orthogonal to the proposed method.