We thank the reviewers for their constructive feedback and will incorporate all input in the final version. All three appreciate the effectiveness of our approach, as well as novelty of our work being the first to take a systematic look at continual learning of LSTM-based captioning networks. We respond below to specific points raised by the reviewers.

On novelty (R2). Ours is the first work considering Continual Learning (CL) of image captioning models, and we are among the first to consider CL for recurrent networks. Extending attention masking to recurrent networks for transient tasks required careful design and non-trivial modifications. We propose two task-dependent masks: one on the hidden state and one shared mask on the word and visual embedding. We also use a fixed mask on the classifier which, since we consider transient tasks, can overlap between tasks. Finally, the backward masking on word embedding described in lines 169-173 is a key difference from HAT that allows RATT to leave more weights free for future tasks where vanilla HAT freeze parts of the word embedding even for non-active words.

These are fundamental differences between our approach and HAT. Our results show that in more difficult transient setting excellent results can still be obtained if task-aware masking is carefully applied in different ways to different parts of the network. Our ablation study in Figure 2 shows the importance of all masks to obtaining the best results. We hope that RATT will lead to similar techniques for continual learning in other domains with recurrent networks (e.g. machine translation, multi-label estimation, and visual question answering). In addition, we propose two new setups for continual image captioning, which together with code for all methods will be made public upon acceptance.

On vocabulary overlap (R4). As suggested by R4, we computed word overlaps between tasks for our MS-COCO splits (shown in the table to the right). From this breakdown we see that the task vocabularies are approximately the same size (between around 2,000 and 3,000 words), and there is significant overlap between all tasks.

	T	A	S	F	I
T	3,116 (100.0%)	1,499 (48.11%)	1,400 (44.93%)	1,222 (39.22%)	1,957 (62.80%)
A	1,499 (48.11%)	2,178 (100.0%)	1,175 (53.95%)	1,025 (47.06%)	1,492 (68.50%)
S	1,400 (44.93%)	1,175 (53.95%)	1,967 (100.0%)	933 (47.43%)	1,355 (68.89%)
F	1,222 (39.22%)	1,025 (47.06%)	933 (47.43%)	2,235 (100.0%)	1,530 (68.46%)
I	1,957 (62.80%)	1,492 (68.50%)	1,355 (68.89%)	1,530 (68.46%)	3,741 (100.0%)

On zero forgetting in RATT (R4). Some words are shared between tasks in the last layer, however in the second-to-last layer tasks have *different* attention masks. Thus, weights connecting neurons in the task mask in the second-to-last layer to those in the last are used when computing the probability of a word. These different connection pathways ensure that words can be adapted to their usage in new tasks without causing forgetting in previous ones – even when task vocabulary overlap is significant. During evaluation the task embedding forces use of only weights that were not used by future tasks. We will add elements of this analysis in the final version.

On LwF performance (R4). LwF is known to perform poorly for image classification when there are large domain shifts between tasks (e.g. from flowers to airplanes) and better when domains are more closely related [Aljundi, CVPR 2017]. This could be why LwF performs better on Flicker30K, which uses *incremental visual categories* (and not *visually-disjoint task splits* like MS-COCO). In Flickr30k new images with already seen visual concepts from previously learned tasks can occur. Such images facilitate knowledge transfer with LwF, leading to improved performance.

On captioning quality metrics (R3). We performed an evaluation based on human quality judgments using 200 images (40 from each task) from the MS-COCO test splits. We generated captions with RATT, EWC, and LwF after training on the last task and then presented ten users with an image and RATT and baseline caption in random order. Users were asked (using forced choice) to select which caption best represents image content. The percentage of users who chose RATT over the baseline are given in the table to the right. These results confirm that RATT is superior on all tasks. We will expand this to include Flickr30k and more human judgments in the final version.

	T	A	S	F	I
RATT vs EWC	75.0%	77.5%	72.5%	85.0%	57.5%
RATT vs LwF	77.5%	82.5%	75.0%	62.5%	47.5%

On limitations and failure modes (R4). A drawback of RATT is that at some point network capacity is exhausted and there are no un-attended neurons left, meaning that the network can no longer adapt to new tasks. In this case either network growing techniques should be considered or attention to previous tasks could be relaxed which will lead to forgetting. We did not observe this yet in the experiments we performed, but will include a discussion in the article.

On our choice of captioning model (R2, R3, R4). CL studies methods to mitigate catastrophic forgetting, and the vast majority concentrate on relatively simple, feed-forward CNNs for image classification. They use simple architectures (e.g. ResNet18, ResNet34) in order to focus on the effects of catastrophic forgetting in continual learning. Similarly, we deliberately chose a simple architecture to focus on the adaptation of continual learning techniques to a new problem using a recurrent architecture (LSTM). Our goal was not to achieve the state-of-the-art captioning performance, but rather to systematically study a set of CL techniques and adapt them to an RNN. Modules like attention mechanisms will surely also suffer from forgetting, and while this is interesting to study on its own, RATT is an important first step toward understanding and mitigating the complex problem of catastrophic forgetting in recurrent captioning networks.

As suggested by R4, we will expand the discussion of more recent captioning models in the related work section (which, as suggested by R2, we will move to the Section 2) of the final version of this work, and return to discuss in the conclusions the ramifications of catastrophic forgetting in captioning models with features like attention mechanisms.