Response to Reviewer #1

- *"It is slightly dissatisfying that only 16 number classes were used; there is no class for the number 3."*
- We reported only 16 number classes to match the experiments of Nasr *et al.*'s work [22]. We DID the hypothesis tests for all 30 number classes (from 1 to 30) and the findings are consistent with those of 16 number classes.

- *"It would make the work more impactful if Nu-Net could be trained on a more realistic dataset."*
- Using binary abstract number-depicting images is the tradition of numerosity studies in cognitive sciences. We trained and tested with more realistic images and the accuracy is worse than abstract images. The paper focuses on the basic science problem not engineering applications. Nevertheless, if CNN fails to generalize on simplistic examples, let alone far more varied practical cases.

Response to Reviewer #2

- *"Number sense is a cognitive ability, not a property of individual neurons."*
- We appreciate this reviewer's argument for the possibility of a distributed coding scheme for the number sense. Numerosity cognition may well be a holistic mechanism. Like this reviewer we had the same urge to challenge the methodology and results of numerosity studies in neuroscience [21][22], but refrained from debunking directly the works published in top journals. Thanks to your insight and support, we will make this point in the final version.

- *"The motivation for analyzing only the last convolutional layer. Why would numerosity not appear in earlier layers?"*
- Empirical studies show that deeper layers in CNN encode higher level concepts than shallower layers. Front CNN layers extract low level features (e.g., corners, edges, textures, etc.) Semantics tends to emerge from deep layers. Numerosity, as an abstract cognitive concept, should be exhibited by very deep layers. This is why we and previous authors only examined the last convolutional layer. But for the sake of thoroughness we will check all layers and discuss the results in the final version.

- *"The motivation for using classification rather than regression is not not well justified."*
- We followed the well accepted belief that subitizing is a raw perception not resulting from deliberate calculation. In fact, we also tested regression formulation, the results hardly changed.

- *"No effect sizes are reported for number selectivity."*
- The average η^2 for the numerosity effects of all number-selective units decreases from 0.25 to 0.08, when the sample size increases from 5 to 100. We will add the effect sizes in the final version, as suggested.

- *"Figure 3 and Figure 8 ... labels and titles are much too small"*. - Thanks, will improve as suggested.

Response to Reviewer #3

- *"I would have liked to have seen results from radically different numerosity images ... "*
- We did train and test on numerosity images of much greater variations, and found the inference accuracy and robustness of subitizing decrease. We didn't include these results to keep our experiments in the same setting as in the previous numerosity studies. We will add discussions on more varied sample images in the final version as suggested.

- *"unclear what you mean by 'binary visual representations of numbers' "* - Black and white images depicting numbers.

- *"completely unreasonable to suppose that the architecture of CNNs ... support some sort of numerosity estimation?"*
- We guess here you doubted if CNNs can learn subitizing beyond i.i.d. inference. Indeed, Zhang's work shows empirically that CNNs have the ability to generalize beyond the training images in the identity-mapping task, even trained on a single example. Why shoudn't it be possible for CNNs to succeed in the task of subitizing.

- *"Authors claim that Nu-Net performs subitizing, yet the small numbers (1, 2, 4) the 85% estimation interval is 1."*
- Thanks for pointing out the error. For subitizing the 85% estimation interval length is 0, NOT 1, i.e., δ and ϵ are both 0 when $x < 5$. Likewise, the height of all bars in Fig.7 should be reduced by 1. Nu-Net makes no errors in subitizing more than 85% of times; off by 1 errors can occur but with less than 15% chance. We will fix the errors and clarify.

Response to Reviewer #4

- *"A followup paper showing the claims do not hold is only of limited interest, even accompanied with good analysis."*
- In terms of neuroscience, our negative results are fascinating and have far reaching implications by exposing a pitfall of a standard methodology in published studies of biological neurons; that is, identify number selective neurons via ANOVA. As pointed out by reviewer 2, it is "imperative" to publish these findings, because our critique necessitates reexaminations and calls for new understandings of numerosity, which is of importance in both AI and neuroscience.

We'd like to stress that this work is more than just negating previous well-accepted results; it also offers an interesting constructive result. We did show CNNs can learn subitizing with good accuracy and robustness, although the general numerosity problem turns out much harder. Our partially positive finding points to an intriguing computational parallel to the innate capability of subitizing of humans and primates.