* + Reviewer1:
  + how training and inference times compare against the methods presented in table 1?

RE:

训练时间：

我们与前三个（DDGSD， BYOT，CS-KD）时间差不多，后两个的时间大约是我们的三至四倍。

推理时间：大家都差不多(?)。后两个（check）

第二个review：

问题：

1. 新意度——仅仅结合了self-attention和自蒸馏，这种结合是很直观的，缺乏新意度:

RE: ??? Refer to R1

Unfair . As

* + 1) Authors propose a new knowledge distillation approach with self-attention, where attention is used as a means to consider the importance of each intermediate layer to the final feature representation. They also demonstrate in Table 4 the role self-attention plays to the method performance.  
    2) Experiments are also appropriate for the scale of the study, providing a well-rounded demonstration of performance. In addition they show how their method resembles bagging, where weak learners and a voting process can provide a more stable and reliable model.
  + The approach proposed is simple, but well founded. The authors provide a theoretical basis to their claims along with a solid set of experiments across a number of benchmark datasets. The method appears to be credible and sound and shows a good boost in performance compared to other established approaches. The paper is very easy to follow and has been well written.

1. 证明很显然: 虽然straightforward但是是对这个方法的性质证明，这个性质很重要
2. 实验有限，没有加入imageNnet等大数据集测试，FRSKD+SLA在很多数据集也有好的效果（训练时间长）: 以我们现在的资源，需要一两周的时间，可以加上。

Why do not incorporate SLA strategy?：我们的工作重心在于self attention，并不是SLA.之后可以加上，看性能是否有进一步的提升。

cutout效果比self-attention还大（其他aug方法在后面实验有展现）：XXXX

1. 一些书写错误：Fix
2. 相关工作不充分: 会加上

**第三个review：**

问题：

1. 创新性: 同第二个1，想了解更多结合这两项的motivation: 图1只是大框架，在此基础上，加了。。。。。。
2. 分析有错（验证一下）
3. 与FRSKD+aug没有充分比较（训练时间长，复杂度高）为什么没有充分比较
4. Code will be available if accepted for publication.

Show, attend and distill: Knowledge distillation via attention-based feature matching

为何不与2比: 我们已经做过，因为…没有加入论文中。我们可以加上这部分。

（需要额外的教师模型，训练时间长，实验后续有补充），表4的mixup有提到（作为augmentation的匹配性的研究，并不是本文主要说明的，本文主要表现self-kd与self-attention，aug的说明是辅助说明与aug的有效性）。

**第四个review（感觉没咋看文章）：**

1. 说我们只适用视觉，没有运用到语言: Unfair. Focus computer vision. 其他工作也这样。
2. 说我们只用在resnet和小数据集，需要更多的数据，我们也用到了WRN，但是确实数据集不够大:   
   RE: 基础研究工作，大部分Baseline也只做这些数据集。
3. 说没有调研哪些层有利于蒸馏，但其实我们是通过自适应学习出来的哪些层

RE: 与我们的工作没什么关联。仅仅是他们用了self attention

1. 有没有场景哪些是我们的方法不能适应的？

RE: 我们考虑的就是baseline比的一些场景，在这些场景上我们都有提升。