

# Alleviating Noisy Data in Image Captioning with Cooperative Distillation

Pierre Dognin

Igor Melnyk  
Jerret Ross

Youssef Mroueh

Inkit Padhi  
Yair Schiff

Mattia Rigotti

IBM Research AI

## Abstract

*Image captioning systems have made substantial progress, largely due to the availability of curated datasets like Microsoft COCO or Vizwiz that have accurate descriptions of their corresponding images. Unfortunately, scarce availability of such cleanly labeled data results in trained algorithms producing captions that can be terse and idiosyncratically specific to details in the image. We propose a new technique, cooperative distillation that combines clean curated datasets with the web-scale automatically extracted captions of the Google Conceptual Captions dataset (GCC), which can have poor descriptions of images, but is abundant in size and therefore provides a rich vocabulary resulting in more expressive captions.*

Learning with noisy data is an important and challenging problem in machine learning. In image captioning, the problem of noisy annotations has been addressed by several authors. For example, [4] uses a number of heuristics (tagging, annotations, word statistics, etc.) to filter out low quality captions. Similarly, in [6] the training captions are denoised by extracting only the noun phrases, which are then treated as a training dataset. In the winning entry of the GCC challenge [2], authors employ dynamic filtering in the training, which is based on model confidence.

Our cooperative distillation (co-distill) framework trains a *student model* on a large *noisy dataset*. By noise we mean that captions are often grammatically incorrect or do not match the semantic content of images. We also rely on a *clean dataset* to train a *teacher model*. We explore whether we can leverage the specific advantages of both types of datasets by training on a rich vocabulary and variety of scene contexts, while alleviating the noisy annotations.

**Semantic Bridge between Datasets.** We use a BERT tokenizer to generate a joint vocabulary across both the noisy and clean datasets at a *sub-word units* level, and we also leverage BERT to represent captions as the embedding of the resulting [CLS] token. Such representation has been shown to capture semantic information that can be used to quantify the *semantic similarity* between captions (see e.g.

[3]). Semantic similarity is a crucial element in our algorithm used for: 1) *denoising*, by down-weighting noisy captions that are not semantically aligned with those generated by the teacher model, and 2) *increasing caption diversity*, by augmenting the training dataset with generated captions that are aligned with clean ground truth captions.

**Student and Teacher Transformer Models.** Transformer networks from [5] are used for both student and teacher models. Captions are generated conditionally on the encoded features via the transformer decoder. The transformer architectures have 2 layers (for both encoder and decoder), embeddings of size 512, and 8 attention heads per layer.

**Cooperative Distillation.** In order to train both the student and the teacher, we alternate between two training streams: denoising and diversity, shown in Figure 1. In the *denoising stream*, the student minimizes its loss given a fixed teacher model, and in the *diversity stream*, the teacher minimizes its loss given a fixed student as explained next.

**Denoising Stream Loss.** Given a sample from the noisy student dataset, we embed the ground truth caption through BERT and represent it as the embedding of the resulting [CLS] token. We then push the image through the Teacher Transformer to obtain a softmax distribution, which is decoded using a greedy-max approach to produce the teacher’s predicted tokens sequence. As with the tokenized ground truth caption, we embed this predicted sequence via BERT. These two embeddings allow us to define a *semantic coherence weighting* for the denoising stream, which captures the similarity between the noisy ground truth caption and the caption predicted by the teacher model.

For a fixed teacher, the student minimizes a per-sample loss that combines weighted terms of cross-entropy based on ground truth labels and KL-divergence between student and teacher predictions. The weight on each term is determined by the semantic coherence calculated through the BERT embedding bridge. If the semantic coherence weight is high, the student can “trust” its ground truth and puts a higher weight on the regular cross-entropy training term. If the semantic coherence weight is low, the student instead distills the teacher through a KL-term comparing the soft-labels from the student and the caption predicted by the

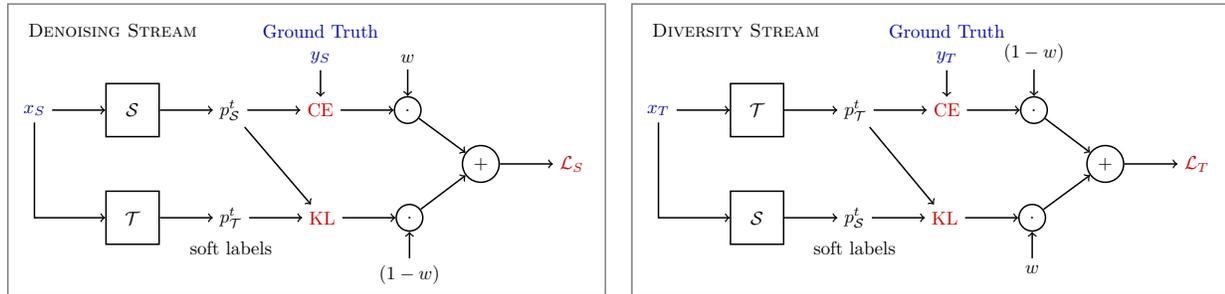


Figure 1. Denoising and diversity streams. In the denoising stream (left panel), an image  $x_S$  from the clean dataset  $S$  is decoded by both the student model  $\mathcal{S}$  and teacher model  $\mathcal{T}$ , resulting in soft-label predictions (softmaxes)  $\{p_S^t\}_t$  and  $\{p_T^t\}_t$ . These are used to compute the Cross-Entropy (CE) loss and the KL divergence in the distillation loss, which in turn are weighted by the semantic coherence weight  $w$ . The diversity stream (right panel) is analogous to the denoising stream with the role of student and teacher reversed. For high similarity weights,  $w$ , the teacher is trained by distilling the student.

teacher. Hence, the denoising distillation loss interpolates between the hard noisy label and the teacher’s soft-label.

**Diversity Stream Loss.** For a sample from the clean teacher dataset, we follow a similar procedure. We begin by embedding the clean ground truth caption via BERT. We then obtain predicted captions from the student model for each image using greedy max decoding. We embed the student’s predicted caption using BERT. Finally, we define the semantic coherence weight between the teacher ground truth caption and the caption predicted by the student.

For a fixed student the teacher minimizes a per sample loss that combines weighted terms of cross-entropy based on ground truth labels and KL-divergence between teacher and student predictions. The teacher trusts its ground truth caption if the semantic coherence weight is small, and distills the student if the semantic coherence weight between the ground truth caption and the caption predicted by the student is high. This injects language diversity into the teacher model training for captions with high weight.

**Results.** We use two datasets for our experiments. The first is *Microsoft COCO* [1] (the clean dataset) and a subset of 500K images of *Google Conceptual Captions (GCC)* [4] (the noisy dataset). We conduct a human evaluation on Amazon MTurk, where human evaluators were shown an image with captions from our co-distill method and from a baseline trained on the noisy dataset and evaluators were asked to rate each caption on a Likert scale from 1 to 5. We see from Figure 3 that co-distill outperforms the noisy baseline, hence effectively denoising the noisy training data.

## References

- [1] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *EECV*, 2014. 2
- [2] Ruotian Luo, Gilad Vered, Lior Bracha, Gal Chechik, and Greg Shakhnarovich. Winning Google Conceptual captions challenge, 2019. 1



Figure 2. Example captions on GCC test dataset. Captions generated by the student (S) tend to be more descriptive and image-specific than the “noisy” groundtruth annotations (GT).

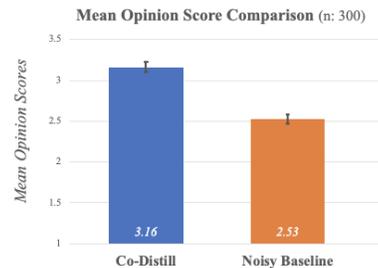


Figure 3. Human evaluations of noisy baseline and co-distill.

- [3] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*, 2019. 1
- [4] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, July 2018. 1, 2
- [5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 1
- [6] Yulong Zhang, Yuxin Ding, Rui Wu, and Fuxing Xue. A Denoising Framework for Image Caption. In *2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing*, pages 825–832. IEEE, 2019. 1