**CS2770: Homework 3**

**Due:** 4/13/2021, 11:59pm  
  
  
This assignment is worth 100 points.  
  
You will experiment with different representations and datasets for cross-modal retrieval (i.e. retrieving text from images, and vice versa).  
  
Please review [this paper](http://people.cs.pitt.edu/~kovashka/thomas_kovashka_semantic_neighborhoods_eccv2020.pdf) and [this corresponding video](http://people.cs.pitt.edu/~kovashka/thomas_eccv2020_10min.mp4) for background on cross-modal retrieval.  
  
You will experiment with three datasets: COCO ([link](https://cocodataset.org/) where you can retrieve images from captions and vice versa), Flickr30k ([link](http://bryanplummer.com/Flickr30kEntities/)), and GoodNews ([link](https://github.com/furkanbiten/GoodNews)).  
  
As with HW2, you will include your code in a Colab Jupyter notebook.  
  
This assignment allows you to use code published by others; the goal is to perform the tasks with code you can find that is relevant, complementing it with your own code as needed.  
  
  
Part A: Baseline (40 points)  
  
The basic approach involves representing the image with AlexNet, the text with Glove, and using triplet loss to do retrieval. (If you want a slightly more advanced task, try [angular loss](https://openaccess.thecvf.com/content_ICCV_2017/papers/Wang_Deep_Metric_Learning_ICCV_2017_paper.pdf), [N-pairs loss](https://papers.nips.cc/paper/2016/hash/6b180037abbebea991d8b1232f8a8ca9-Abstract.html), or [Noise Contrastive Estimation](https://openaccess.thecvf.com/content_CVPR_2020/papers/Miech_End-to-End_Learning_of_Visual_Representations_From_Uncurated_Instructional_Videos_CVPR_2020_paper.pdf) instead). Train and evaluate this method on a dataset of your choice (one of the above). During evaluation, formulate retrieval as a k-way multiple-choice task, similar to the paper referenced above (i.e. 1 option will be correct, k-1 will be incorrect). Define the evaluation metric as top-1 accuracy (was the highest-ranked choice the correct one).  
  
  
Part B: Cross-dataset performance and adaptation (30 points)  
  
Compare the performance of the models trained on each dataset, when you evaluate on another dataset (e.g. a model trained on COCO can be evaluated on COCO, Flickr30k, or GoodNews). Compute results for text to image, and another for image to text. Report accuracy for each, when training on dataset A and testing on dataset A, as well as when testing on dataset B.  
  
Then for your pair of source-target datasets, train a method to perform domain adaptation by adding a domain classifier loss and inverting the gradient (as in the [Ganin paper](http://proceedings.mlr.press/v37/ganin15.pdf)). Use the same amount of training data from the source dataset, and use no more than 10% of the target data for training. The gradient will only update the projection weights for the common space. Compare the results of performing domain adaptation to directly evaluating on the target dataset.  
  
  
Part C: Representations (30 points)  
  
You can represent an image with features from some CNN we discussed in class (e.g. AlexNet, VGG, ResNet). You can represent text with Glove, RNN or BERT. Experiment with two combinations of (1) representation of the image, and (2) representation of the text. Train the the corresponding retrieval models with these two alternate representations. Use the original (within-domain) setting, and report performance on one dataset of your choice.  
  
If you are interested in a slightly more challenging task, find code for a cross-modal transformer (e.g. LXMERT, VILBERT, VILLA, ERNIE, etc.) that has been pre-trained and gives you both a visual and textual representation. Don't train the transformer (due to the computational cost) but just apply it for retrieval; this would replace one of the required representation combinations (1, 2) above.