Visual Relationship Recognition

Shane Josias: *18642365*
Supervisor: Dr. W. Brink

08 August 2019

Applied Mathematics Division, Stellenbosch University

Problem

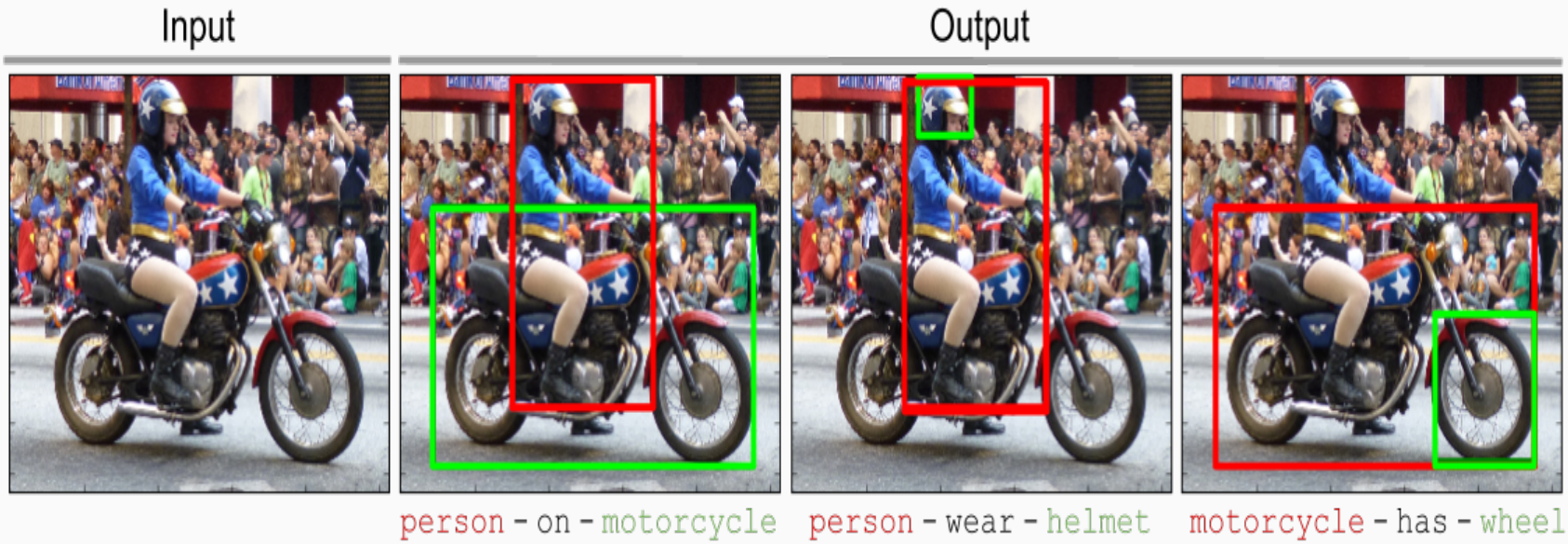

Figure 1: Visual relationships [1]

Difficulties:

1. Set of all possible relationships *(subject, predicate, object)* is large.
2. 100 objects, 70 predicates $\implies$ 700 000 classes.
3. Long tail distribution (we have to learn from few examples).

How bad is it?

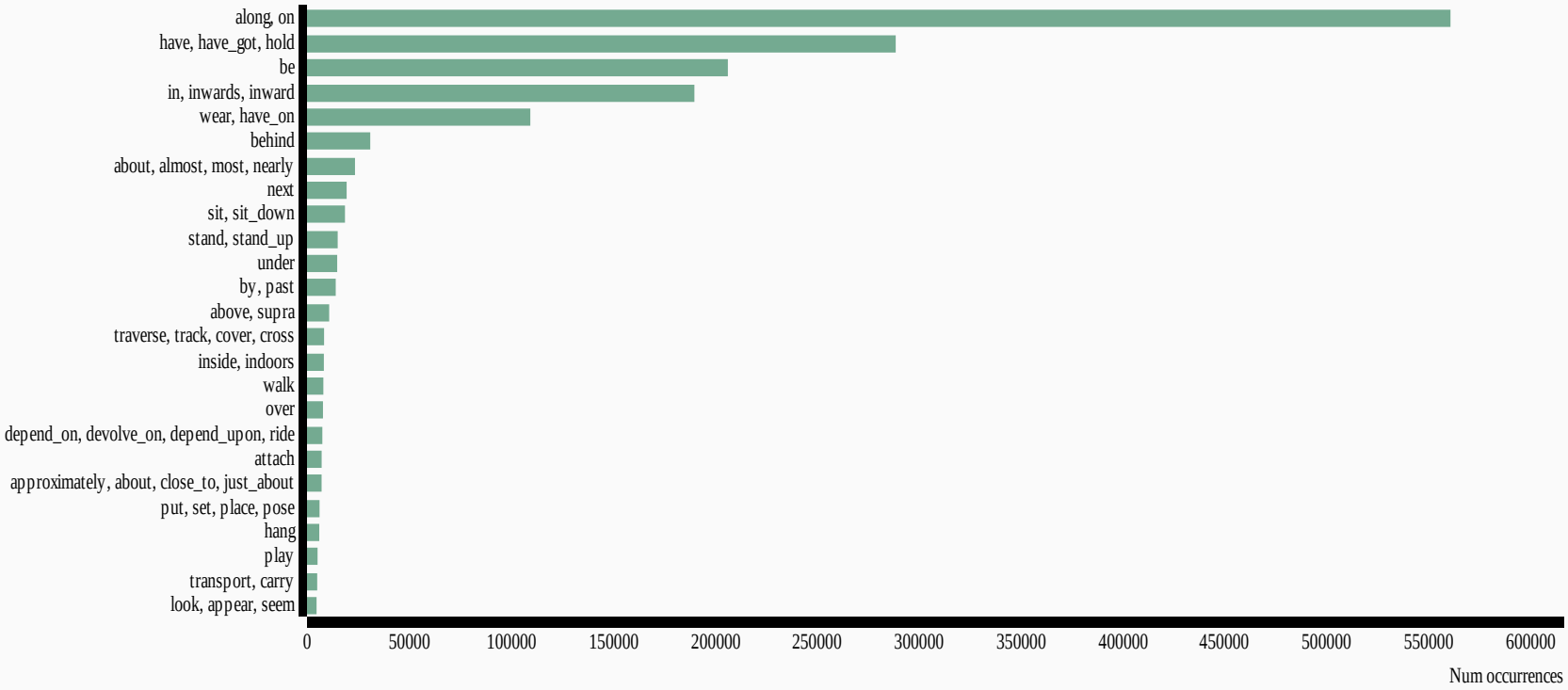

Figure 2: Visual relationships long tail (VG dataset [2])

This is only 53 out of 40,480 unique relationships!

Approach: Ranking Loss

Cross entropy on 700 000 classes is not feasible and requires a large, balanced dataset. What about ranking loss functions with Siamese networks?

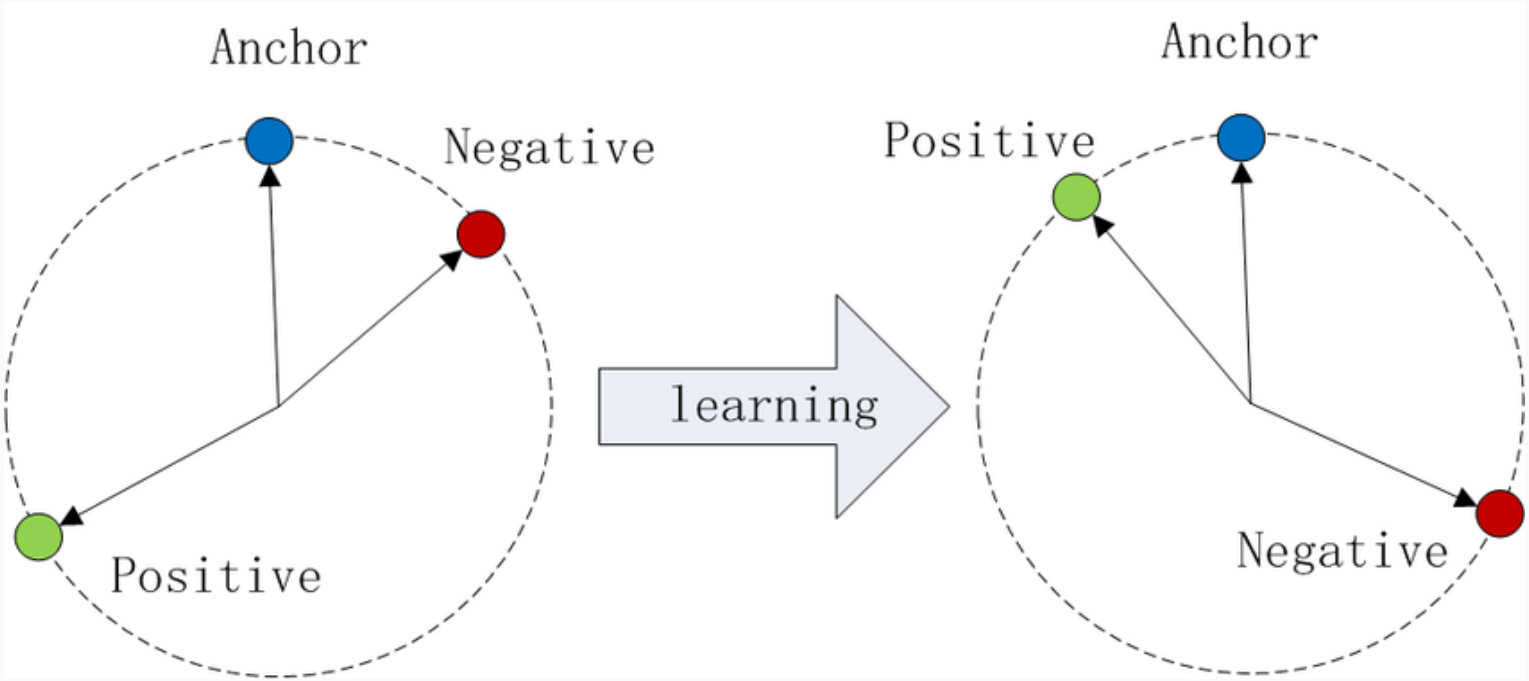

Figure 3: Ranking loss (triplet) [3]

Hypothesis: More robust to data imbalance.

Experiment

We can test the performance of cross entropy and ranking loss over various levels of *uniformity*:

1. Consider MNIST and CIFAR 10 datasets.
2. Impose various levels of uniformity (from uniform to long tail).
3. Keep total number of samples constant.
4. Train (from scratch):
   - 4.1 Classification loss: cross entropy.
   - 4.2 Ranking loss: contrastive, online contrastive, triplet, online triplet.
5. Evaluate average per class accuracy.

Repeat 50 times to obtain mean and standard deviation.

Distributions

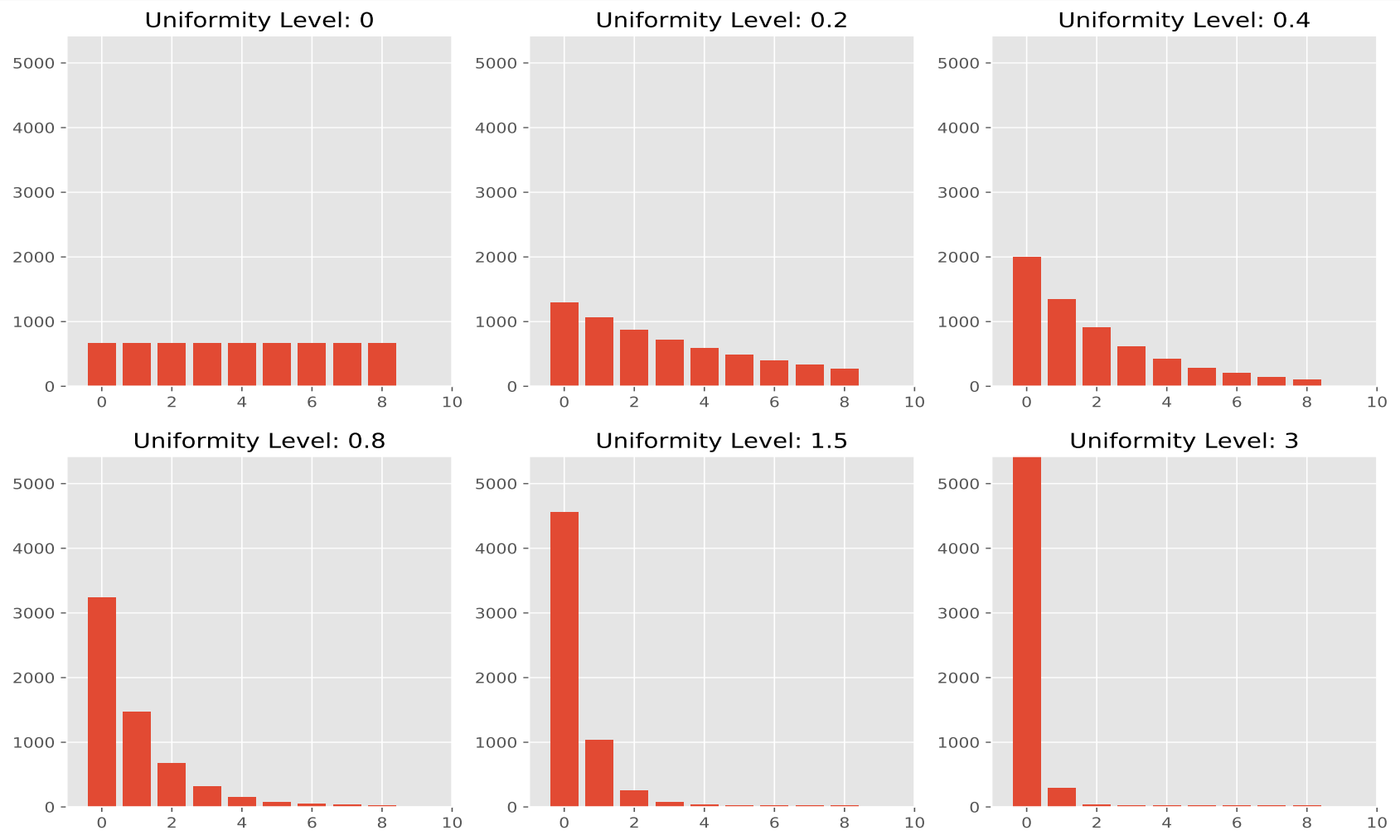

Figure 4: Various levels of uniformity

Will it work?

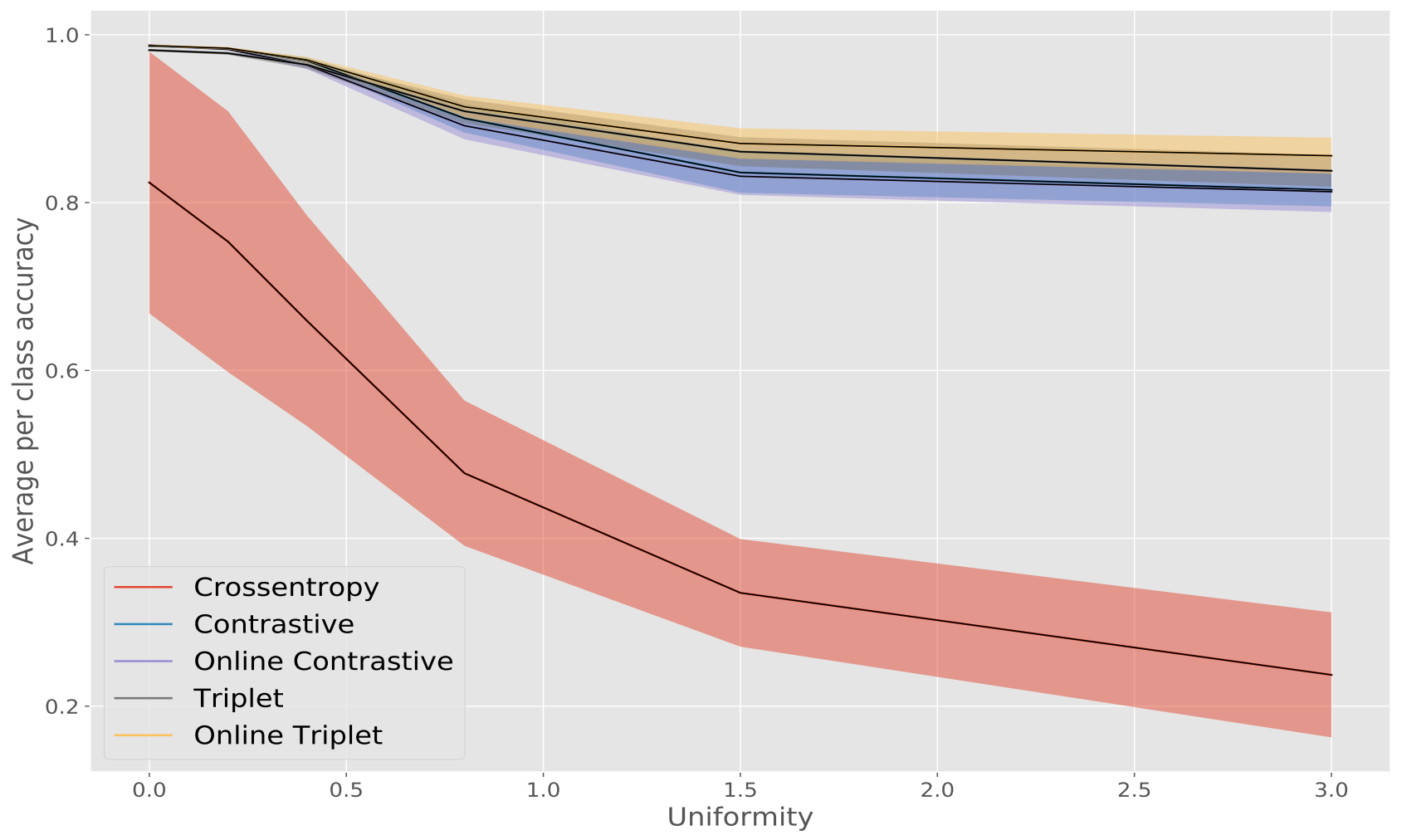

Figure 5: Ranking loss functions vs cross entropy functions (MNIST)

Next steps

Turns out that sampling matters!

1. **Easy negatives:** makes loss zero, model doesn't train.
2. **Hard negatives:** causes model to collapse.
3. **Semi-hard negatives:** sweet spot for good training.

Visual Relationship recognition:

1. **Multi-task objectives:** leverage the power of multi-task training to learn subject, predicate, object as 3 different tasks.
2. **Language model:** can make the embedding / ranking loss more robust?

References


Lu C, Krishna R, Bernstein M, Fei-Fei L.
Visual relationship detection with language priors.
In: European Conference on Computer Vision. Springer; 2016. p. 852–869.


Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, et al.
Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations; 2016. Available from: `https://arxiv.org/abs/1602.07332`**.**


Li C, Ma X, Jiang B, Li X, Zhang X, Liu X, et al.
Deep speaker: an end-to-end neural speaker embedding system.
arXiv preprint arXiv:170502304. 2017;.