Depth First Learning **Wasserstein GAN**

Week 4: 24 Feb, 2019

Scribed by: Sebastian Bodenstein, Sasha Naidoo, & James Allingham

4.1 Differences between WGAN and GAN

The issue of determining the quality of trained GANs came up previously. It is interesting that you can use the loss of WGAN as a measure for the quality of samples.

- Why doesn't the Jensen-Shannon divergence have the same property?

It is interesting/remarkable that various issues we looked at before go away with WGAN. The issues are:

- mode dropping
  - Is there no mode dropping due to the difference in Earth-Mover distance and KL-divergence?
  - Why do we usually have mode-dropping? In normal GAN use Jensen-Shannon divergence, which has both versions of KL. It cares less about mode dropping?
- distance measures, such as KL divergence, break down on low-dimensional manifolds

4.2 Enforcing the Lipschitz constraint

The WGAN paper mentions that it is a bad idea to enforce the Lipschitz constraint using weight clipping. Why?

- idea: Weights from linear layer before a logistic sigmoid might have a very different scale to convolution weights. It seems a bad idea to use a single clipping parameter on a multi-scale problem.
  - Tali rough intuition: "My intuition isn't good enough yet to nail this argument, but it kind of feels like the Lipschitz constraint is establishing a spatial scale for the slope of the discriminators decision boundary. You don't want it to change its mind too quickly, because that will kill gradients. So by establishing the Lipshitz constraint, you can soften that boundary so that gradient doesn't die. If thats the right way of looking at it, is one scale appropriate? What if you space has different scales in different directions?" One scale doesn't seem flexible enough. Tim: 'The fact cited in the paper that clipping the weights makes the growth of the function linear might support this intuition.' Tim gives the exact quote: "The fact that we constrain the weights limits the possible growth of the function to be at most linear in different parts of the space, forcing the optimal critic to have this behaviour". So in different parts of the space it has different values, but any part of the space looks linear. Tali: "My objection was more that if you just imaging stretching the space, in one axis, then the bound you want to put on the growth of the discriminator might need to be very different on one axis compared to another axis. So establishing a bound on the growth uniformly across dimensions and directions in the space might not be what you need for discriminator with the Lipschitz property." Tim: "so you are talking about the rate of change when you are actually training the network?" Tali: "yes".

  - Sebastian: we do choose a single scale for the learning rate in plain SGD, and that trains well. Maybe a single clipping scale is also fine in practice? Tali: that is why we have ADAM, etc. which are better. Sebastian: SGD can be competitive with ADAM etc. See this paper `https://arxiv.org/abs/1705.08292`, and quote "We observe that the solutions found by adaptive methods generalize worse (often significantly worse) than SGD, even when these solutions have better training performance. These results suggest that practitioners should reconsider the use of adaptive methods to train neural networks." James: Jeremy Howard mentions AdamW as being better than SGD and ADAM. Tali: this is about generalization, which is generally complicated. Do you think that something similar is happening with weight clipping? Sebastian: no idea, but SGD is just there to show that the scale issue is not necessarily a killer in practice for generalization, that we really care about.
  - Tali: an extra unnaturalness argument for using a single weight-clipping value: for general differential programming, you can arbitrarily multiply weights by factors of, say 100. Eg consider $\mathbf{y} = \tanh(100\mathbf{x} \cdot \mathbf{W}))$. This will change the rate of growth of discriminator. So can arbitrarily change this growth rate by changing weight scalings in net, which seems wrong. This argument might not apply to gradients? Probably not, as values of grads change when they go through that value of 100.

- another idea: it introduces an extra hyper-parameter that needs tuning.
  - Paper points out that if you choose the parameter too large or small, all sorts of pathologies happen. So makes training much slower as you need to find good setting.
  - James: will this work just as well as other Lipschitz constraint methods if you can tweak the clipping magnitude correctly? Sebastian: don't know! Probably. But its still super ugly, as the more things you need to tweak, the less likely anyone is of using a method.

Are there alternatives to weight clipping for enforcing the Lipschitz-constraint?

- how does weight clipping depend on the architecture?
- spoiler for next week from Tali: gradient penalties
- Quote from WGAN paper:"However, we do leave the topic of enforcing Lipschitz constraints in neural network setting for further investigation, and we actively encourage interested researchers to improve on this method." What follow-ups have there been? Will look at follow-ups next week.

Why does constraining the weights make the growth of the function at-most linear?

- based on earlier quote: "The fact that we constrain the weights limits the possible growth of the function to be at most linear in different parts of the space, forcing the optimal critic to have this behaviour"
- Tali: if you bound the weights, you bound the magnitude of the Jacobian. James: "So then by bounding the magnitude of the Jacobian, you are bounding how much the weights can change because thats in effect bounding the gradient." Tali: "Not quite, you are bounding how much the decision of the discriminator can change as a function of the input. The only way that the discriminator can have sharp transitions in the input space is if it has very large weights." 'If you consider a simple linear perceptron, the gradients are proportional to the transpose of the weights. Hence bounding weights bounds gradient. This is weight clipping.'.

4.3 Mode Collapse in WGAN

In the WGAN paper, they identified that their method *did not* result in mode collapse.

Question: Why is mode dropping a danger with traditional GAN training? Why is this guaranteed to happen?

Answer: The KL divergence discourages mode dropping with Forward KL (i.e. tractable). A normal GAN, uses both Forward and Reverse KL (because it uses the JSD) which means that mode dropping becomes prevalent.

Question: Why does EM distance care *more* about mode dropping than JSD?

Answer: JSD is made up of two KL terms, which do not give useful information around mode dropping at is saturates at $log2$ when a mode is dropped. This implies that it does not know how to recover as it doesn't have a notion of distance in density space. This property results in 'near misses' in JSD to be equivalent to 'complete misses'. The Earth Mover's (EM) distance uses moving mass and therefore can identify locations of dropped modes in density space more easily.

4.4 KR Duality

Question: Is there an intuition behind the KR Duality?

Answer: If the minimum transport map i.e. a function that takes a few samples from the real to the fake distribution, maps x to y then the optimal critic function grows linearly with x and y with a slope of 1.

4.5 Critic

Question: Why do they call it a critic and not a discriminator?

Answer: It is possible to add a sigmoid to the evaluation metric to determine accuracies but essentially, this would not be meaningful as you can discriminate two distributions perfectly i.e accuracy would be 1. The name comes from the actor-critic model.

Question: Can the critic be used for anything?

Answer: As the critic is not trained as a classifier, it is not valuable on its own.

4.6 WGAN Loss

The low dimensional manifolds in general are victim to a 'near miss' problem which JSD can give little to no information on how to recover compared. EM, however, can give an idea of distance in the original data space.

In training, the WGAN loss can represent the original data space better, which could be causing the direct relationship between loss and image quality. In addition, with JSD and the standard GAN, the discriminator cannot be trained fully but with WGAN the discriminator can be trained to optimality (as the critic has nice gradients), which results in the WGAN loss having more meaning.

4.7 Tali's Notebook

Tali did some cool experiments with training GANs:

1. He compared training for a GAN and a WGAN with very simple models and data (the data was a mixture of gaussians, the latent space was a single gaussian, and both the generator and discriminator/critic were MLPs). He noted that:
   - The rate of change of the discriminator can be seen as a force acting on the generator. The slope of the discriminator (in the data space) tells the generator how to change to match the true distribution.
   - When training a WGAN, the weight clipping limits how steep this slope can be (more weight clipping means a less steep slope).
   - Even with very simple examples it is difficult to train GANs (WGANs included).
2. He trained a generator that was simply an embedding layer (i.e. the generator just memorizes a couple examples). This makes the examples pretty interpretable because the generator doesn't actually do much – it just moves points around. Some observations:
   - Here we could really see that the discriminator was applying a force to the generator because we could see that the points were being moved around.
   - With a standard GAN, we can see that as the discriminator gets more complex that the slope can get much steeper - this doesn't happen with a WGAN. The steep curves are nice because the force moves the points very quickly however we end up with vanishing gradients! The WGAN slop doesn't saturate in the same way.
3. He tried to train on MNIST but it didn't really work. This might be due to computational limitations on his laptop.

4.8 James's Experiments

James did have more luck with training WGANs (more architectures seemed to work compared with standard GAN). One thing to note is that WGANs seem to need many more training epochs to give reasonable results.