- 1. Integrated gradients
- 2. Download a pretrained image classifier from TF-Hub
- 3. Load and preprocess images with tf.image
- 4. Classify images
- 5. Calculate Integrated Gradients
- 6. Establish a baseline
- 7. Unpack formulas into code
- 8. The formula for Integrated Gradients is as follows:
- 10. IntegratedGradientsi(x)::= $(xi-xi')\times \int \alpha=01\partial F(x'+\alpha\times(x-x'))\partial xid\alpha$
- 11. Interpolate images
- 12. Compute gradients
- 13. Visualizing gradient saturation
- 14. Accumulate gradients (integral approximation)
- 15. Putting it all together
- 16. Visualize attributions
- 17. Uses and limitations