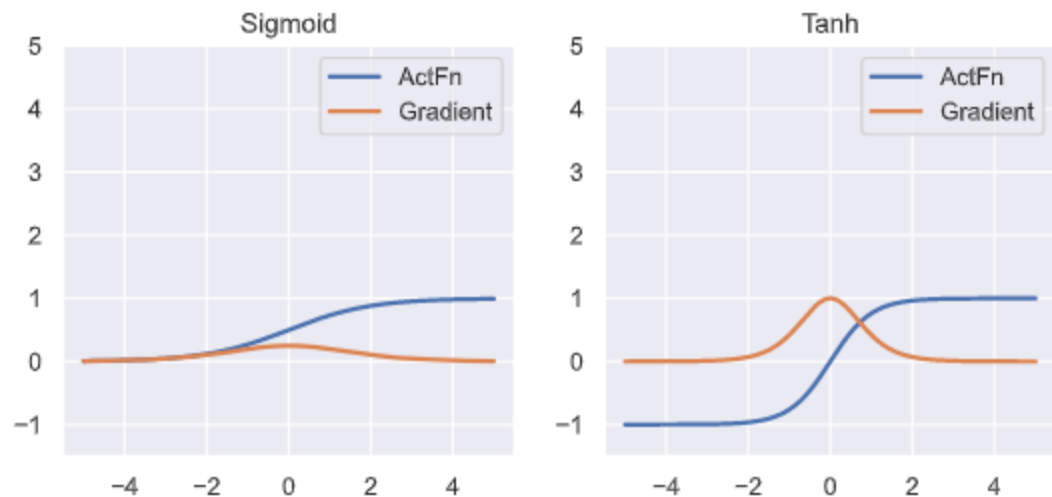


# Impressions

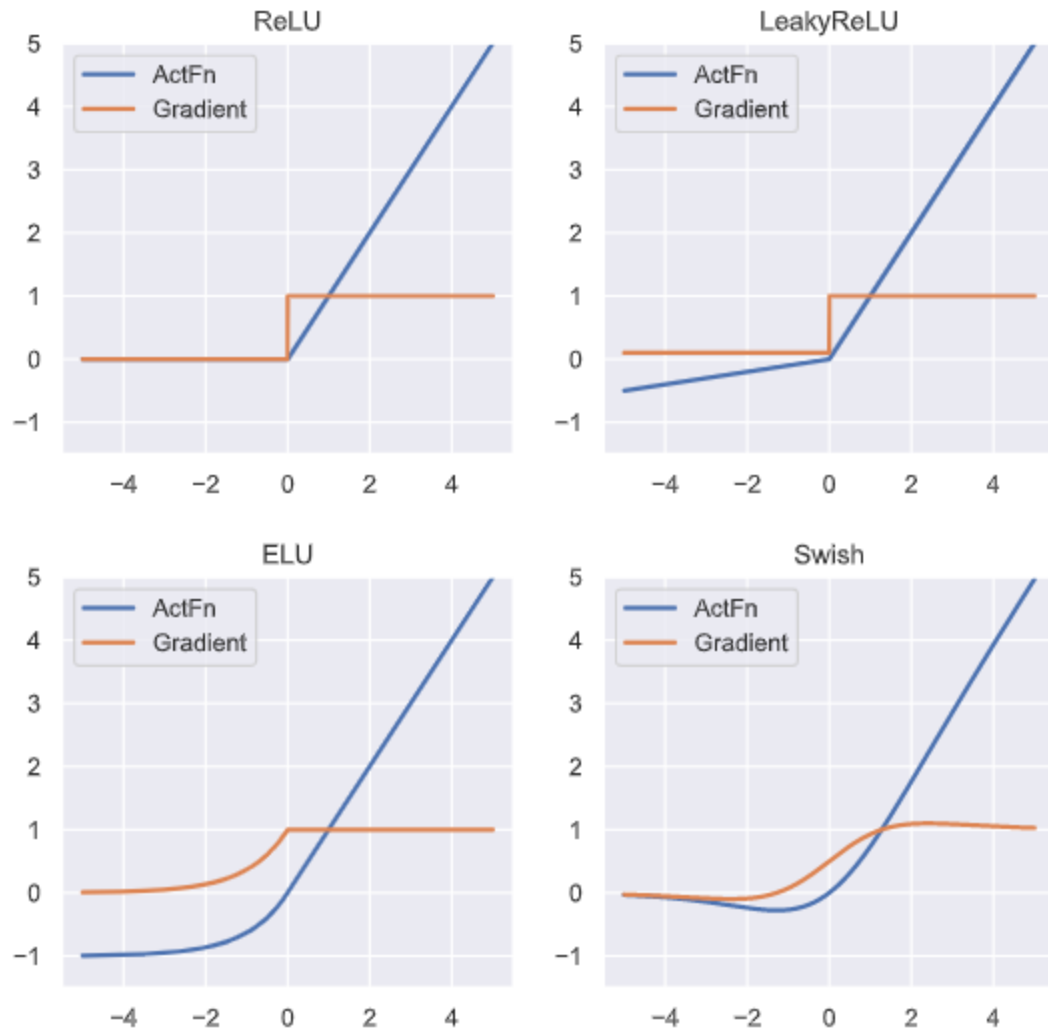
After reviewing the ActivationFunctions notebook, I was able to re-learn and get better insights of the alternatives.

Among the differences between these functions are:

- **Formulas.** The definition of the functions are different, thus, the essence itself changes. Also its gradient behavior does.
- **Gradient behavior.**
  - The classic ones suffer from gradient vanishing, that is: as the values mainly live in 0 and just have a small bump. The gradient starts disappearing as it goes through deeper layers.



- The ReLU family try to solve this problem by having gradients near one when greater-than-zero input is given, and as they keep evolving, a more complex behavior shows up with values smaller than zero.



Due to similar reasons, ReLU faces some difficulties since it has zero gradient for negative values, which is then amended by the other ones.

## Conclusion

As told in the notebook, I'll stick to the advice of starting with a ReLU approach and then tune with hopes of finding better results.