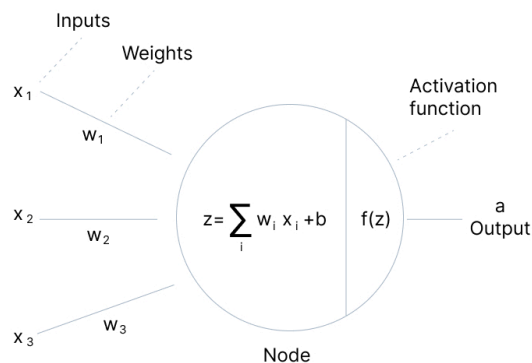


Activation functions and their use cases

تابع فعالیت باعث میشه که شبکه، قسمت بدرخور دیتا رو از قسمت دیگر جدا کنه.



V7 Labs

تابع فعالیت تصمیم میگیره که یک نورون فعال (fired) بشه یا نه، یعنی خروجیش برای شبکه مهم باشه یا نه.

معمولا همه لایه‌های hidden تابع فعالیت یکسانی دارند اما لایه‌ی خروجی بر اساس هدف مدل متفاوته.

ما به تابع فعالیت نیاز داریم چون اگر نباشه، عملا میتونیم تمام لایه‌ها رو با یک لایه نشان بدیم و شبکه به یک رگرسیون خطی تبدیل

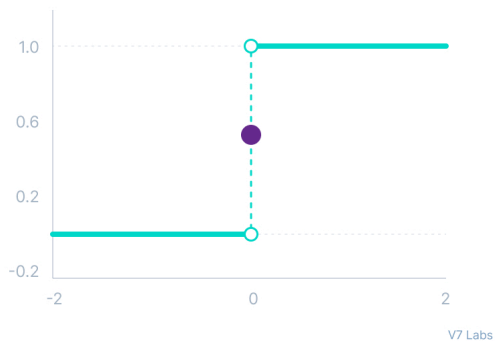
میشه. یعنی بدون اکتیویشن فانکشن، افزایش عمق شبکه غیرممکنه چون همه چیز خطی جلو میره. با بودن تابع فعالیت، غیرخطی

بودن وارد شبکه میشه.

انواع توابع فعالیت

• Binary step

Binary Step Function



V7 Labs

برای مقادیر منفی، ۰ و برای صفر و مثبت، ۱

برمی‌گرداند.

مشکلات: نمیتونه بیش از دو کلاس رو کلسیفای کنه.

گرادیانش صفره که برای بکپروپ مشکلسازه چون عملا وزنهای اپدیت نمیشن.

کاربرد: کلسیفایر دودویی در لایه آخر.

• Identity function/ linear activation function

$$f(x)=ax$$

مشکلات: مشتق همواره یک عدد ثابت و هیچ ارتباطی با ورودی x نداره. در واقع وزنهای و بایاس اپدیت میشن اما همیشه به یک اندازه ثابت.

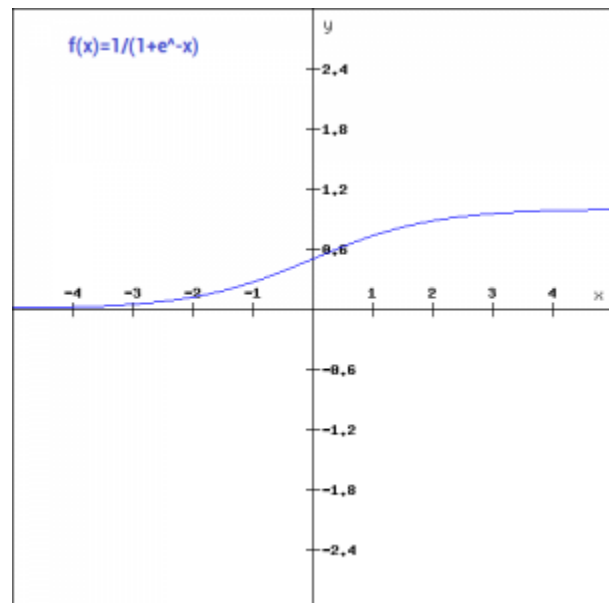
ممکنه فقط مناسب تسکهای ساده باشه که می‌خواهیم تفسیرپذیری بالا باشه.

توابع غیرخطی:

• سیگموید/لاجیستیک

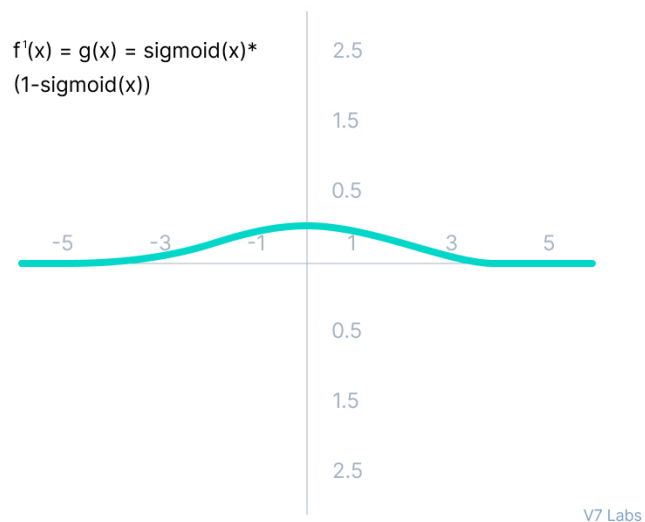
Sigmoid / Logistic

$$f(x) = \frac{1}{1 + e^{-x}}$$



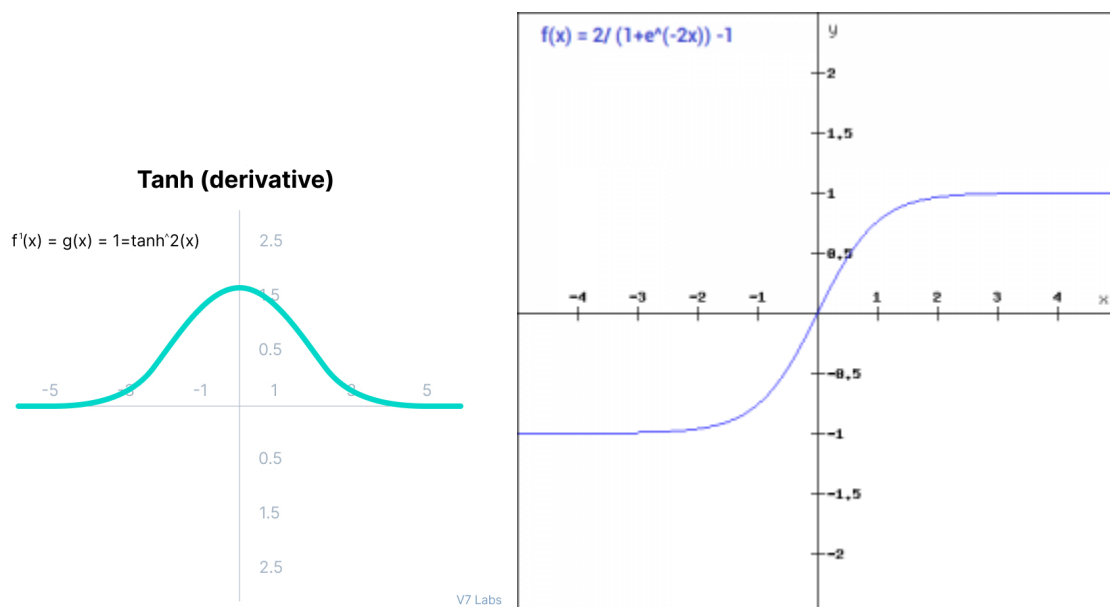
مناسب وقتی که خروجی، احتمال باشه چون مقادیرش بین 0 و یک. گرادیان همه جا داره و همه جا مشتق‌پذیره. اینکه همه خروجی‌ها مثبت هستن و تابع اطراف صفر متقارن نیست، در یادگیری تاثیر منفی میگذاره. چون گرادیان علامتش همیشه مثبت هست هم مشکل ونیشینگ گرادیان پیش میاد.

چون گرادیان خارج از بازه ۳- تا ۳ تقریباً صفره، شبکه دیگه یاد نمیگیره و **vanishing gradient** پیش میاد.



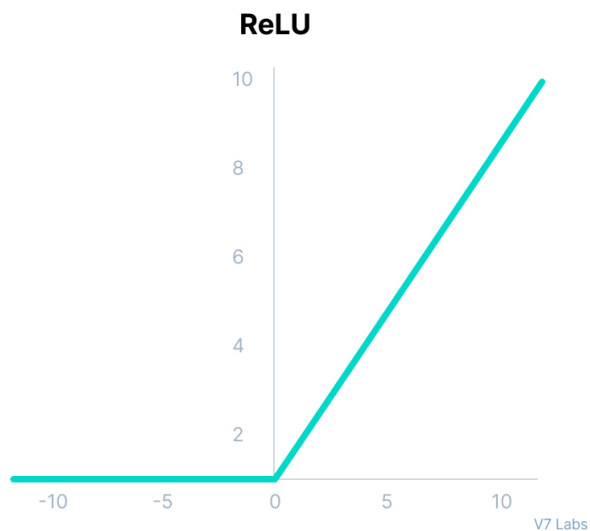
• تانژانت هیپربولیک

مثل سیگموئید ولی بین ۱- و ۱. حول مبدا متقارنه. یک نوع scale شده‌ی سیگموئید هست. $\tanh = 2\text{sigmoid}(2x) - 1$. مناسب کلسیفایر باینری.



• رلو

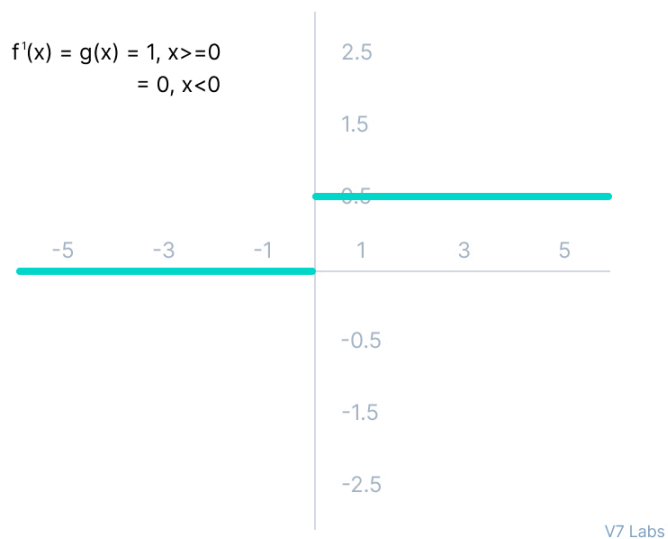
نکته اینه که همه نوروں ها رو فایر نمیکنه. مقادیر منفی، فایر نمیشن. همین باعث میشه از نظر محاسباتی بهتر عمل کنه.



همگرایی به سوی گلوبال مینیمم سریعتره. اما چون مشتق

مقادیر منفی هم صفره، نوروں ها میمیرن و کلا دیگه اپدیت نمیشن:

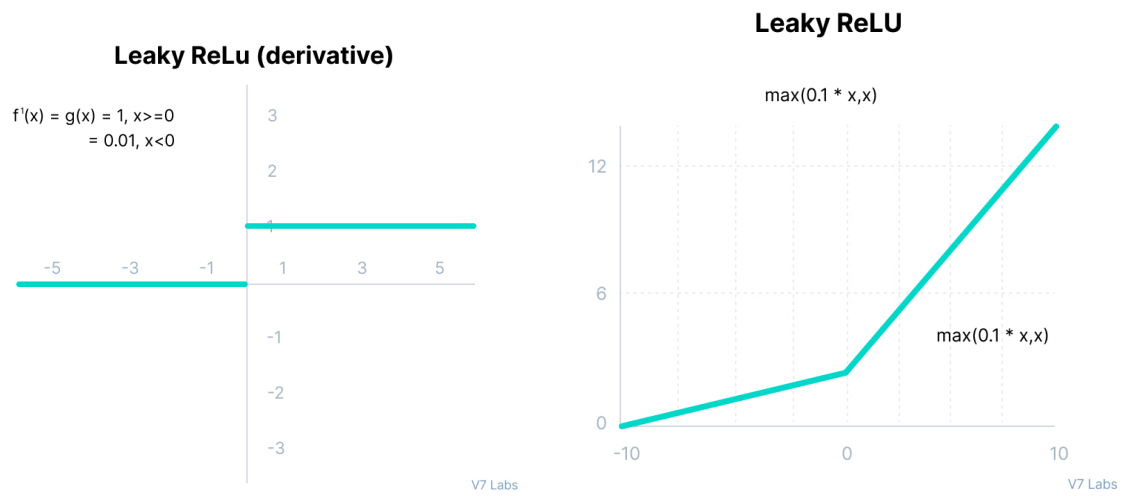
The Dying ReLU problem



چون بلافاصله مقادیر ورودی منفی، صفر میشن،

نوروں هاش کلا میمیرن.

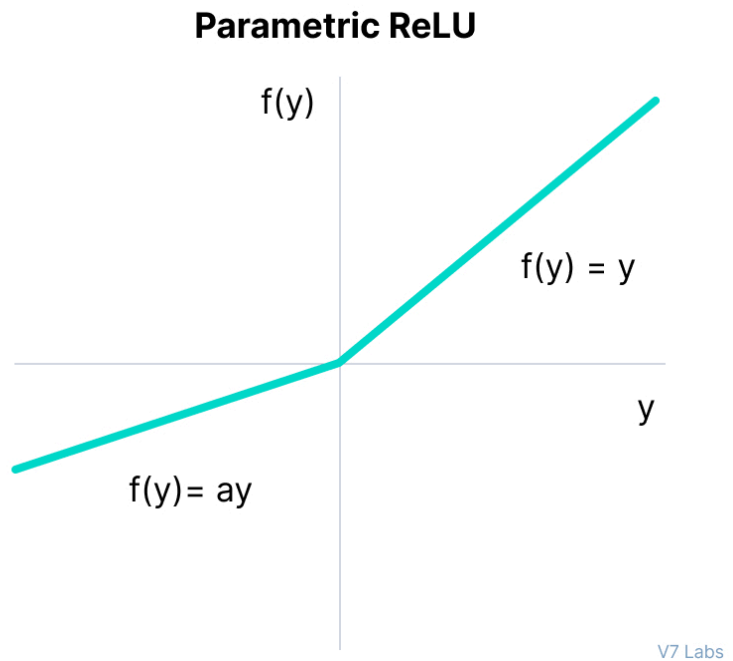
• لیکی رلو



مشکل مردن نورون‌ها حل میشه چون برای مقادیر منفی هم گرادیان صفر نیست.

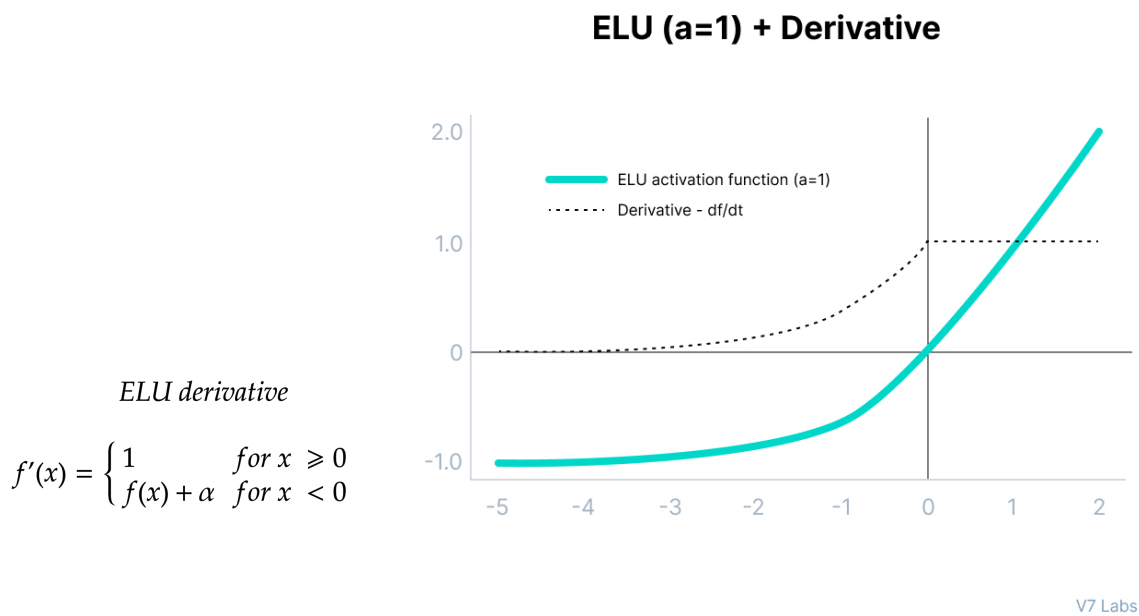
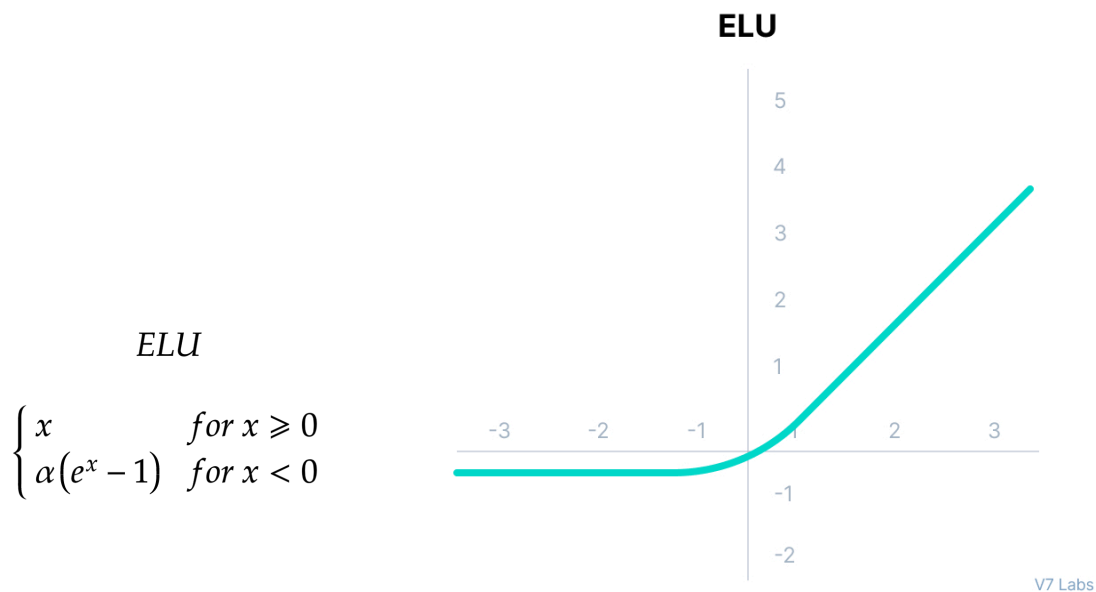
● رلوی پارامتری

مقدار شیب خط در ناحیه منفی، لرن میشه.



● Exponential Linear Units / ELUs

مثل رلو ولی در مقادیر منفی، خط نداریم بلکه یک تابع لگاریتمی داریم. چون مقادیر منفی هم میده، دیتا رو کمی نورمال میکنه و میانگین رو به صفر نزدیک میکنه.



ترکیب چندین سیگموید. معمولاً در لایه آخر استفاده میشه برای کلسیفای به چندین کلاس.

Softmax

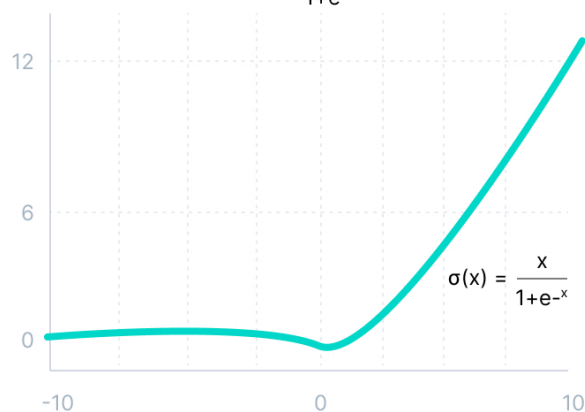
$$\text{softmax}(z_i) = \frac{\exp(z_i)}{\sum_j \exp(z_j)}$$

V7 Labs

Swish •

Swish

$$\sigma(x) = \frac{x}{1+e^{-x}}$$



V7 Labs

Swish

$$f(x) = x * \text{sigmoid}(x)$$

برای ایمپج کلسیفیکیشن، ترجمه ماشینی و... بسیار بهتر از رلو عمل کرده.

در رلو، مقادیر منفی کلا صفر شدن ولی اینجا مقادیر کوچک منفی خنثی نشدن و تاثیرگذارن فقط مقادیر بزرگ منفی صفر شدند.

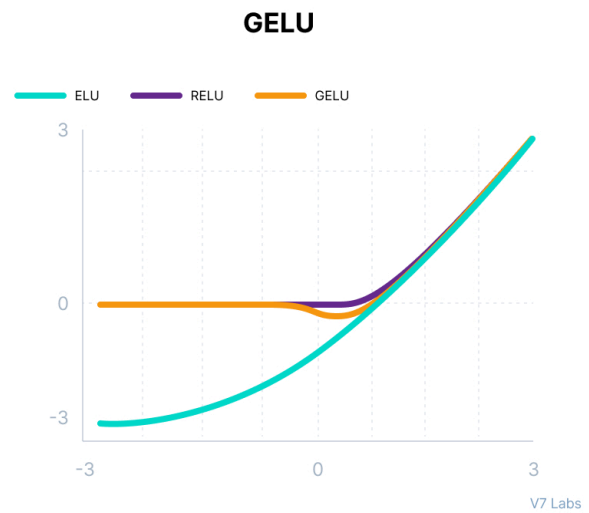
این تابع یکنوا نیست و همچنین کران هاش از دو طرف بی‌نهایت.

Gelu •

Gaussian Error Linear Unit. مناسب برای برت، رویرتا و مدل‌های ان ال پی.

ترکیب رلو و تکنیک دراپ اوت هست.

$$\begin{aligned} GELU \\ f(x) &= xP(X \leq x) = x\Phi(x) \\ &= 0.5x \left(1 + \tanh \left[\sqrt{2/\pi} (x + 0.044715x^3) \right] \right) \end{aligned}$$



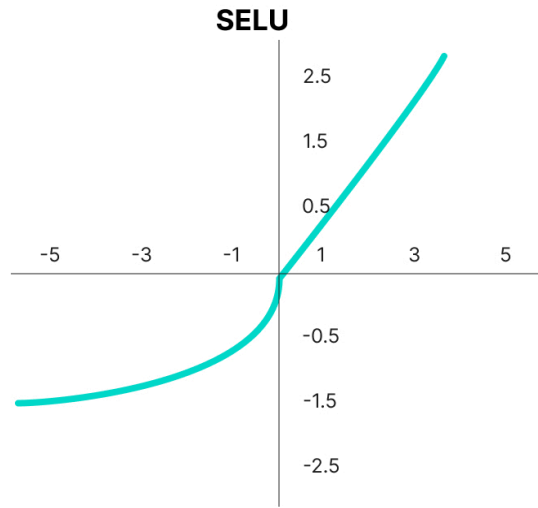
تابع توزیع تجمعی از توزیع نورماله $(0, 1)$ $X \sim N(0, 1)$ ، $\Phi(x) = P(X \leq x)$

چون نورونها از توزیع نورمال پیروی میکنند (مخصوصا بعد بچ نومرالیزیشن)، از توزیع نورمال استفاده شده.

کاربرد در computer vision, natural language processing, and speech recognition.

Selu •

Scaled Exponential Linear Unit میانگین و واریانس لایه قبل رو حفظ میکنه.



$$f(\alpha, x) = \lambda \begin{cases} \alpha(e^x - 1) & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases}$$

در کل بهتره اول در لایه های هیدن، رلو بذاریم و اگر جواب نداد، بریم سراغ بقیه.

سیگموئید و tanh در هیدن استفاده نشه بخاطر vanishing gradient.

در شبکه های با عمق بالای ۴۰ هیدن لیر، از سویش استفاده کنید.

برای لایه خروجی:

Regression -> Linear Activation Function

Binary Classification—>Sigmoid/Logistic Activation Function

Multiclass Classification—>Softmax

Multilabel Classification—>Sigmoid

Multilabel یعنی وقتی که یک ورودی، به دو تا کلاس خروجی یا بیشتر تعلق داشته باشه.

برای لایه های هیدن:

Convolutional Neural Network (CNN): ReLU activation function

Recurrent Neural Network: Tanh and/or Sigmoid activation function

Neural Network Activation Functions

