تمرین کلاسی تحلیلی: توابع فعال‌ساز و مقداردهی اولیه وزن‌ها

# Activation Functions & Weight Initialization

مقاله‌ی مبنا

**Activation functions in deep learning: A comprehensive survey and benchmark**

Shiv Ram Dubey, Satish Kumar Singh, Bidyut Baran Chaudhuri — Neurocomputing 503 (2022), 92–108

لینک رسمی ناشر ScienceDirect :

نسخه‌ی آزاد مقاله arXiv: 2109.14545 :

## سناریوی تمرین کلاسی

این مجموعه برای یک فعالیت کلاسی ۹۰ دقیقه‌ای طراحی شده است. هدف آن این است کاروند در طول کلاس، با تکیه بر مقاله‌ی مرجع و مطالب نوت‌بوک، رفتار توابع فعال‌ساز و اثر آن‌ها بر جریان گرادیان را تحلیل کند.

منابع تمرین به‌صورت شفاف از هم تفکیک شده‌اند:

- مقاله‌ی Dubey و همکاران برای مفاهیم و مقایسه‌ی Activation Functionها Non-linearity :، Sigmoid/Tanh، ReLU/LReLU/PReLU، GELU و نتایج تجربی.
- نوت‌بوک Week9Day2Session1.ipynb برای بخش‌های Weight Initialization، Zero Initialization، Symmetry، Variance، Xavier و He.
- هرجا سؤال از داده‌ها یا ادعای مقاله شروع شده ولی از دانشجو تحلیل تازه می‌خواهد، در انتهای سؤال با برچسب «سؤال تحلیلی بر اساس منبع» مشخص شده است.

## سؤال ۱ — چرا **Non-linearity** ضروری است؟

زمان پیشنهادی: ۱۲ دقیقه

مقاله یک تابع فعال‌ساز خطی را به‌صورت $f(x)=cx$ معرفی می‌کند و بیان می‌کند که استفاده از activation خطی، non-linearity به شبکه اضافه نمی‌کند. همچنین حتی با وجود چندین لایه، خروجی شبکه همچنان تابعی خطیaffine/ از ورودی باقی می‌ماند.

یک شبکه‌ی دو لایه بدون **activation** غیرخطی داریم:

$$h = W_1 x + b_1$$
$$y = W_2 h + b_2$$

الف) با جایگذاری h نشان دهید که این شبکه را می‌توان به صورت $y = W'x + b'$ نوشت.

ب) نتیجه‌ی قسمت الف را به یک شبکه‌ی دارای تعداد زیادی لایه تعمیم دهید.

ج) توضیح دهید چرا اضافه کردن لایه‌های بیشتر، اگر بین آن‌ها فقط transformation خطیaffine/ وجود داشته باشد، به‌تنهایی امکان مدل‌کردن روابط غیرخطی را فراهم نمی‌کند.

د) با توجه به توضیح مقاله درباره‌ی داده‌های non-linearly separable، توضیح دهید activation غیرخطی چه نقشی در feature space شبکه دارد.

منبع: مستقیم از مقاله — Section 2: Evolution of Activation Functions.

تحلیل: اثبات جبری قسمت‌های الف و ب، سؤال تحلیلی بر اساس همان ادعای مقاله است.

## سؤال ۲ **Sigmoid** — و **Tanh:** کدام مشکل حل شده و کدام باقی مانده؟

زمان پیشنهادی: ۱۳ دقیقه

برای **Sigmoid** داریم:

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

خروجی Sigmoid در بازه [0,1] و خروجی Tanh در بازه [−1,1] قرار می‌گیرد. مقاله Sigmoid را در ورودی‌های بزرگ مثبت و منفی saturated می‌داند و این مسئله را به vanishing gradient مرتبط می‌کند. همچنین Tanh را zero-centered معرفی می‌کند، ولی می‌گوید مشکل vanishing gradient همچنان در آن وجود دارد.

الف) دو ورودی $x_1=0$ و $x_2=10$ را برای Sigmoid در نظر بگیرید. بدون نیاز به محاسبه‌ی دقیق، مشخص کنید کدام ورودی در ناحیه‌ی saturation قرار دارد و چرا.

ب) توضیح دهید منظور از عبارت «Saturation can lead to vanishing gradients» چیست. پاسخ باید ارتباط تغییر کم خروجی نسبت به تغییر ورودی و gradient را توضیح دهد.

ج) یک تفاوت مهم Sigmoid و Tanh از نظر مرکز خروجی بیان کنید.

«چون *Tanh zero-centered* است، پس مشکل *Sigmoid* کاملاً حل شده است».

د) استدلال بالا را بر اساس مقاله نقد کنید.

هـ) چرا مناسب نیست بگوییم «Tanh همان Sigmoid است و هیچ تفاوتی ندارند»؟

منبع: مستقیم از مقاله — Sections 2–3: Logistic Sigmoid and Tanh Based Activation Functions.

تحلیل: بخش نقد گزاره‌ها، سؤال تحلیلی بر اساس اطلاعات مقاله است.

## سؤال ۳ — چرا **ReLU** به وجود آمد و چرا **Leaky ReLU** پیشنهاد شد؟

زمان پیشنهادی: ۱۵ دقیقه

**ReLU** به‌صورت زیر تعریف می‌شود:

$$\mathrm{ReLU}(x) = \max(0, x)$$

برای ورودی مثبت، مشتق ReLU برابر 1 و برای ورودی منفی برابر 0 است. بردارهای زیر را در نظر بگیرید:

$$z = [-3, -1, 0.5, 2]$$
$$g = \partial L/\partial a = [2, -4, 3, 1]$$

الف) با استفاده از مشتق ReLU، مقدار $\partial L/\partial z$ را برای چهار مؤلفه محاسبه کنید.

ب) توضیح دهید چه اطلاعاتی از gradient مربوط به دو ورودی منفی از بین می‌رود.

مقاله سپس **Leaky ReLU** را با شیب کوچک در ناحیه‌ی منفی معرفی می‌کند:

$$\mathrm{LReLU}(x) = x \ \ (x \geq 0), \qquad \mathrm{LReLU}(x) = 0.01x \ \ (x < 0)$$

ج) قسمت الف را این بار برای Leaky ReLU با شیب 0.01 تکرار کنید.

د) تفاوت gradient flow در قسمت منفی ReLU و LReLU را توضیح دهید.

هـ) چرا از توضیحات مقاله نمی‌توان نتیجه گرفت که «هرچه slope قسمت منفی بزرگ‌تر باشد، بهتر است»؟

و) مقاله PReLU را به‌عنوان یک activation پارامتریک معرفی می‌کند. از نظر مفهومی توضیح دهید trainable کردن شیب منفی چه پاسخی به مشکل LReLU می‌دهد.

منبع: مستقیم از مقاله Section 4 — و Section 4.1: On the Non-utilization of Negative Values of ReLU.

تحلیل: محاسبات gradient برای بردار داده‌شده، سؤال تحلیلی طراحی‌شده بر اساس این بخش است.

## سؤال ۴ **ReLU** — در برابر **GELU**

زمان پیشنهادی: ۱۰ دقیقه

در نوت‌بوک، تفاوت شهودی این دو تابع به صورت زیر مطرح شده است:

ReLU → Hard Gating

GELU → Smooth Gating

و GELU به شکل زیر معرفی شده است، که در آن $\Phi(x)$ تابع توزیع تجمعی نرمال استاندارد است:

$$\mathrm{GELU}(x) = x\Phi(x)$$

الف) برای $x \to +\infty$ می‌دانیم $\Phi(x) \to 1$. رفتار GELU را در این حد تعیین کنید.

ب) برای $x \to -\infty$ می‌دانیم $\Phi(x) \to 0$. رفتار GELU را در این حد تعیین کنید.

ج) تفاوت این رفتار را با قانون سخت ReLU برای مقادیر منفی توضیح دهید.

د) آیا از همین مطالب می‌توان نتیجه گرفت که GELU در هر مسئله و هر معماری حتماً از ReLU بهتر است؟ توضیح دهید.

منبع مقاله: Section 7.2 — Probabilistic Activation Functions.

منبع نوت‌بوک: Section 7 — GELU : برای تعبیر Hard Gating / Smooth Gating.

تحلیل limit: ها و نتیجه‌گیری قسمت د، سؤال تحلیلی بر اساس تعریف تابع است.

## سؤال ۵ — چرا **Zero Initialization** مشکل‌ساز است؟

زمان پیشنهادی: ۱۵ دقیقه

این سؤال مستقیماً از بخش **Weight Initialization** نوت‌بوک کلاس طراحی شده است.

دو neuron زیر را در نظر بگیرید:

$$y_1 = w_1 x$$

$$y_2 = w_2 x$$

در ابتدا $w_1 = w_2 = 0$ و $x = 2$ است. در demonstration نوت‌بوک، gradient هر دو وزن برابر $-4$ گزارش می‌شود.

الف) اگر دو neuron از نظر ساختار، ورودی و مسیر loss متقارن باشند، توضیح دهید چرا در initialization یکسان، خروجی‌های یکسان تولید می‌کنند.

ب) چرا gradientهای آن‌ها نیز یکسان می‌شود؟

ج) با $w \leftarrow w - \eta(\partial L/\partial w)$ update توضیح دهید چرا اگر قبل از update، $w_1=w_2$ باشد و gradientها نیز برابر باشند، بعد از update هم $w_1=w_2$ خواهد بود.

د) مشکل اصلی این وضعیت را با مفهوم symmetry breaking توضیح دهید.

هـ) چرا مشکل zero initialization صرفاً این نیست که «وزن‌ها صفر هستند»؟ توضیح دهید چرا neuronهای یک لایه فرصت یادگیری featureهای متفاوت پیدا نمی‌کنند.

و) آیا صفر کردن biasها هم الزاماً همان مشکل صفر کردن تمام weightها را ایجاد می‌کند؟ پاسخ خود را با توجه به منطق symmetry استدلال کنید.

منبع: مستقیم از نوت‌بوک Sections 8–9: Initial Weights & Symmetry Demonstration. —

تحلیل: قسمت «و» سؤال تحلیلی بر اساس مفهوم symmetry است و ادعای مستقیم مقاله نیست.

## سؤال ۶ **Initialization** — فقط برای شکستن **Symmetry** نیست

زمان پیشنهادی: ۱۵ دقیقه

نوت‌بوک برای یک ترکیب خطی از ورودی‌ها، با فرض independence و zero-centered بودن، تقریب زیر را مطرح می‌کند:

$$z = w_1x_1 + \cdots + w_nx_n$$

$$\mathrm{Var}(z) \approx n \cdot \mathrm{Var}(w) \cdot \mathrm{Var}(x)$$

فرض کنید **n=100** و **Var(x)=1.**

الف) اگر بخواهیم تقریباً Var(z)=1 باشد، بر اساس رابطه‌ی بالا Var(w) را محاسبه کنید.

ب) اگر به اشتباه Var(w)=1 انتخاب شود، Var(z) تقریباً چقدر خواهد شد؟

ج) اگر چنین بزرگ‌شدن scale در چندین لایه تکرار شود، چه مشکلی ممکن است ایجاد شود؟

نوت‌بوک سپس دو **initialization** زیر را معرفی می‌کند:

$$\mathrm{Var}(W_Xavier) = 2 / (n_in + n_out)$$

$$\mathrm{Var}(W_He) = 2 / n_in$$

فرض کنید **n_in=100** و **n_out=50.**

د) variance وزن‌ها را در Xavier محاسبه کنید.

هـ) variance وزن‌ها را در He محاسبه کنید.

و) چرا طبق توضیح نوت‌بوک، انتخاب activation function و initialization را نباید دو تصمیم کاملاً مستقل در نظر گرفت؟

منبع: مستقیم از نوت‌بوک — Sections 10–13: Scale, Variance, Xavier, He.

## سؤال ۷ — آیا می‌توان گفت یک **Activation Function** «بهترین» است؟

زمان پیشنهادی: ۱۰ دقیقه

این سؤال از بخش experimental مقاله طراحی شده است. مقاله 18 activation function را روی چند معماری مختلف مقایسه کرده است. بخشی از نتایج Table 8 برای CIFAR-10 در جدول زیر آمده است:

| Activation | MobileNet | ResNet50 | DenseNet121 |
|---|---|---|---|
| **ReLU** | 90.10 ± 0.22 | 93.74 ± 0.34 | 93.96 ± 0.51 |
| **LReLU** | 90.10 ± 0.19 | 93.83 ± 0.42 | 93.85 ± 0.48 |
| **GELU** | 90.71 ± 0.20 | 93.81 ± 0.46 | 93.90 ± 0.41 |
| **CELU** | 91.04 ± 0.17 | 94.09 ± 0.17 | 93.46 ± 0.35 |
| **Softplus** | 91.05 ± 0.22 | 93.34 ± 0.65 | 93.07 ± 0.70 |

الف) برای MobileNet کدام‌یک از activationهای جدول بالاترین mean accuracy را دارد؟

ب) برای ResNet50 پاسخ قسمت الف تغییر می‌کند یا خیر؟

ج) برای DenseNet121 کدام تابع در میان موارد بالا بیشترین mean accuracy را دارد؟

د) از پاسخ سه قسمت قبل توضیح دهید چرا انتخاب activation را نمی‌توان کاملاً مستقل از architecture دانست.

> «چون *LReLU* روی *ResNet50* عدد *93.83* دارد و *ReLU* عدد *93.74*، پس مقاله ثابت کرده *LReLU* بهتر از *ReLU* است».

هـ) این نتیجه‌گیری را با توجه به وجود ±0.42 و ±0.34 نقد کنید. آیا تنها مقایسه‌ی دو mean برای ادعای قطعی برتری کافی است؟

و) یک نتیجه‌گیری علمی و محتاطانه در یک یا دو جمله از این جدول بنویسید.

> منبع: مستقیم از مقاله Section 9.2: Experimental Performance Analysis — و Table 8.

تحلیل: قسمت‌های د تا و سؤال تحلیلی بر اساس داده‌های مقاله هستند؛ خود مقاله این جمله‌های سؤال را عیناً بیان نکرده است.