

رسالة محمد



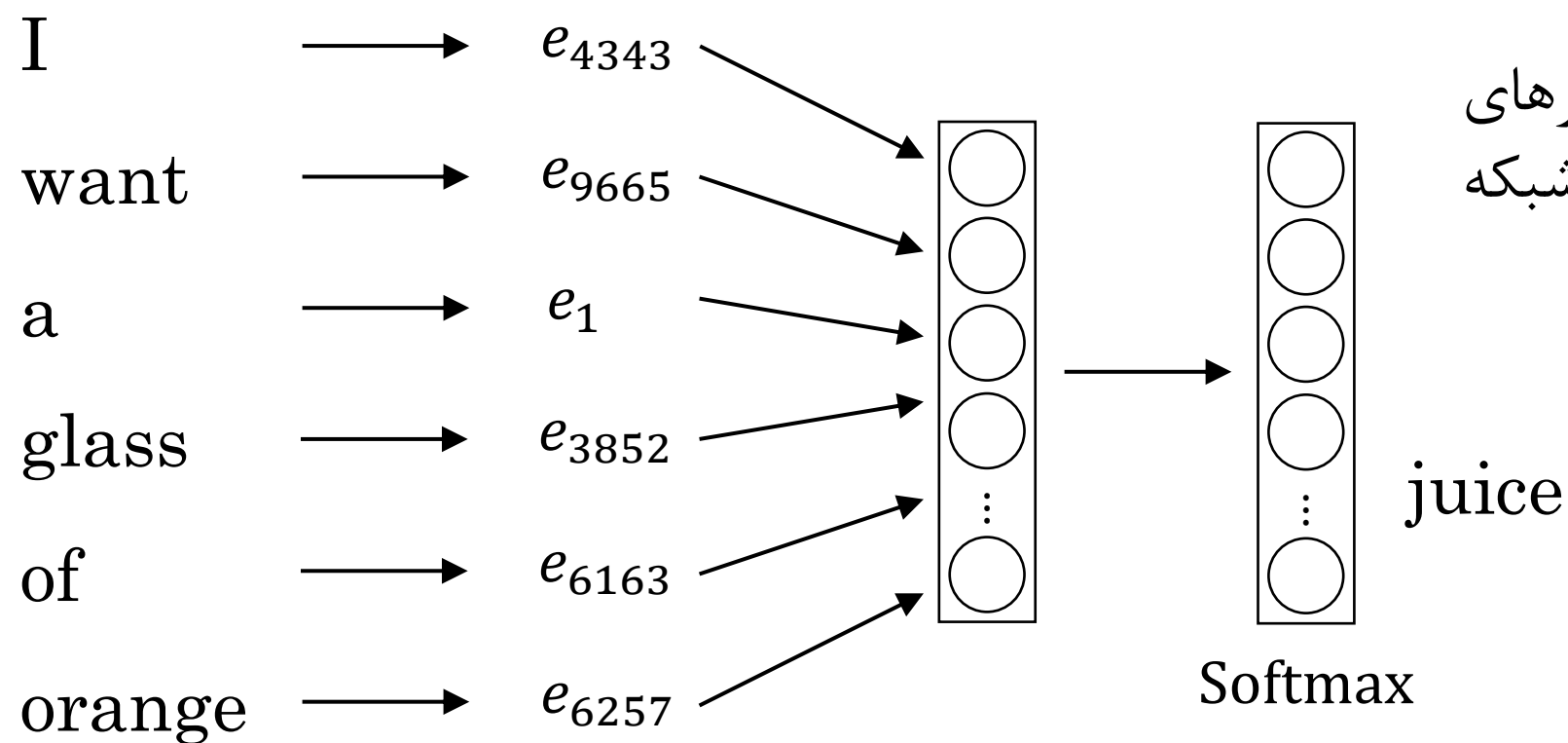
یادگیری بازنمایی

Representation Learning

مدل زبان طبیعی

- در این مدل می‌خواهیم کلمه بعدی را پیش‌بینی کنیم

- هدف ما دستیابی به بردارهای
جانمایی است و وزن‌های دیگر شبکه
هدف نیستند



جفت‌های Context/Target دیگر

- جمله نمونه

- I want a glass of orange juice to go along with my cereal.

- Context

- چهار کلمه قبل
- چهار کلمه قبل و بعد
- یک کلمه قبل
- یک کلمه در نزدیکی

Word2Vec

Context	Target
orange	juice
orange	glass
orange	with

Skip-grams •

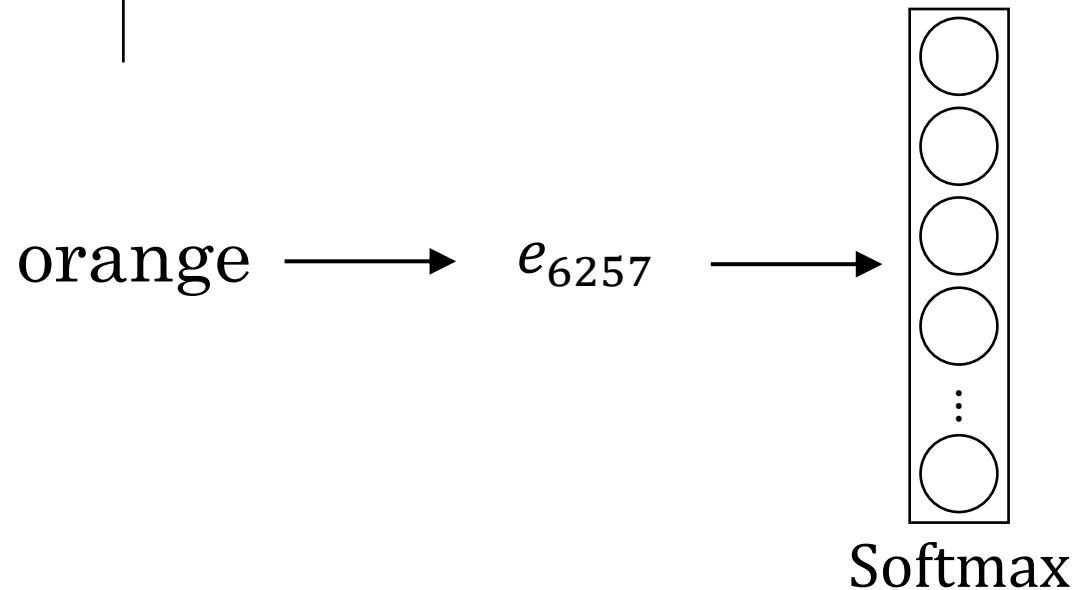
- بجای اینکه همیشه از آخرین کلمات استفاده شود
 - به طور تصادفی یک کلمه را به عنوان Context انتخاب کنید
 - به طور تصادفی کلمه دیگری را در پنجره‌ای اطراف آن انتخاب کنید
- یک مسئله یادگیری با ناظر تنظیم کنید که در آن:
 - با توجه به کلمه Context
 - کلمه‌ای که به طور تصادفی انتخاب شده است را پیش‌بینی کنید

I want a glass of orange juice to go along with my cereal.

Word2Vec

Context	Target
orange	juice
orange	glass
orange	with

Context c ("orange") \Rightarrow Target t ("juice")
6257 4834



• اگر ۱۰۰۰۰ کلمه یکتا داشته باشیم

$$\mathcal{L}_{ce} = - \sum_{j=1}^{10,000} y_j \log \hat{y}_j$$

$$p(t|c) = \frac{e^{w_t^T e_c + b_t}}{\sum_{j=1}^{10,000} e^{w_j^T e_c + b_j}}$$

Word2Vec

- مشکلات Softmax

- مشکل اصلی سرعت محاسبات است
- مخصوصاً برای مجموعه واژگان بسیار بزرگ (مثلاً ۱۰۰,۰۰۰ یا ۱,۰۰۰,۰۰۰ واژه)

$$p(t|c) = \frac{e^{w_t^T e_c + b_t}}{\sum_{j=1}^{10,000} e^{w_j^T e_c + b_j}}$$

Word2Vec

Context	Word	Target
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0

- نمونه برداری منفی (Negative sampling)

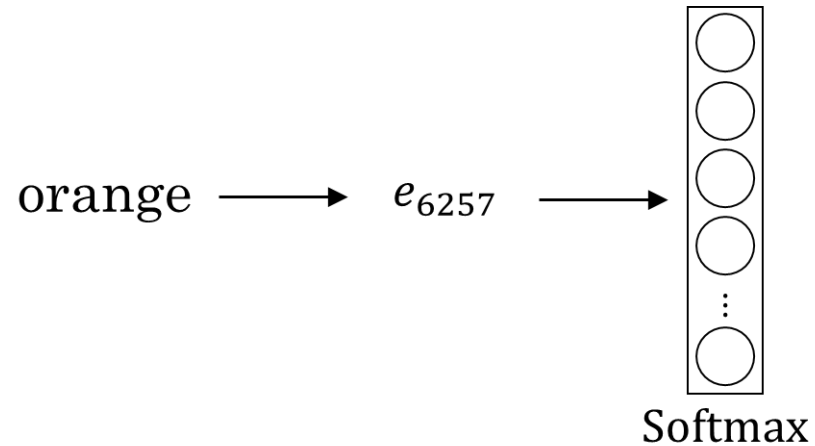
- یک مسئله یادگیری باناظر جدید
 - یک جفت کلمه را ورودی بگیرد
 - پیش‌بینی کند که یک جفت Context-Target است؟
 - بنابراین مسئله این است که پیش‌بینی کند آیا دو کلمه در کنار هم استفاده می‌شوند یا خیر
- مقدار پیشنهادی برای k :
 - ۵ تا ۲۰ برای مجموعه داده‌های کوچکتر
 - ۲ تا ۵ برای مجموعه داده‌های بزرگتر

I want a glass of orange juice to go along with my cereal.

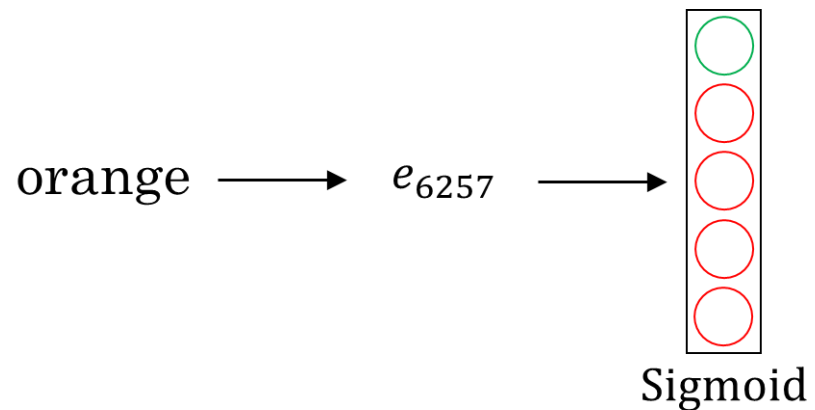
Word2Vec

• نمونه برداری منفی

Context	Word	Target
orange	juice	1
orange	king	0
orange	book	0
orange	the	0
orange	of	0



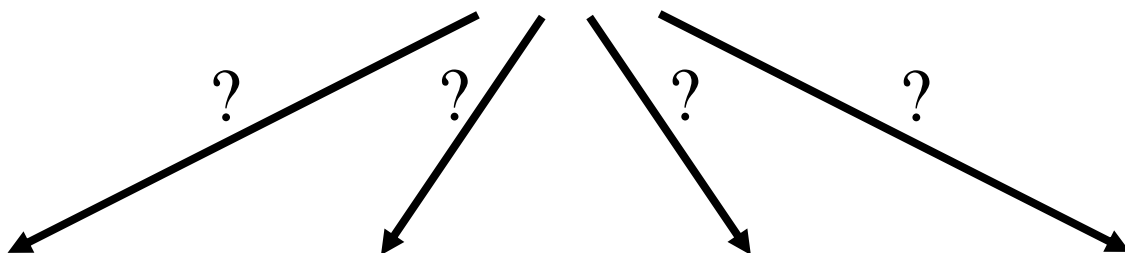
$$p(t|c) = \frac{e^{w_t^T e_c + b_t}}{\sum_{j=1}^{10,000} e^{w_j^T e_c + b_j}}$$



$$p(y = 1|t, c) = \sigma(w_t^T e_c + b_t)$$

یادگیری تک نمونه (One Shot Learning)

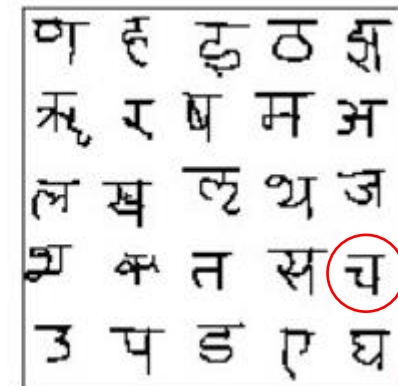
- یادگیری تک نمونه یک مسئله دسته بندی است که در آن تنها از یک نمونه برای هر کلاس استفاده می شود



Test Image

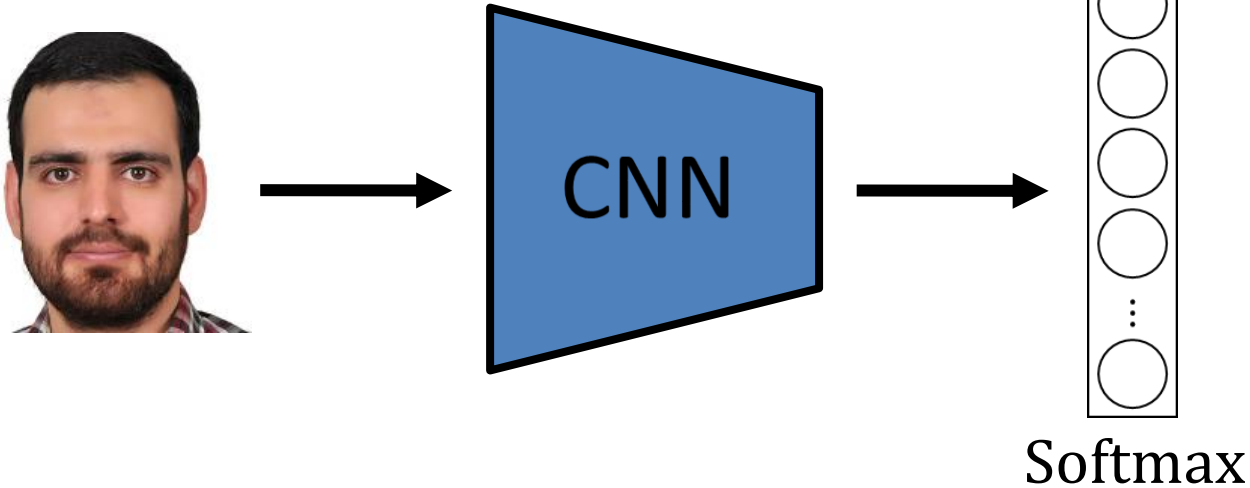


Support Set



بازشناسی چهره تک نمونه

- یادگیری تنها از یک تصویر برای بازشناسی افراد
- آیا ConvNet + Softmax برای بازشناسی چهره تک نمونه مناسب است؟
 - داده کافی برای آموزش یک شبکه عصبی قوی وجود ندارد
 - اگر یک نفر جدید اضافه شود؟
- بجای آن، یک تابع "شباهت" را آموزش می دهیم

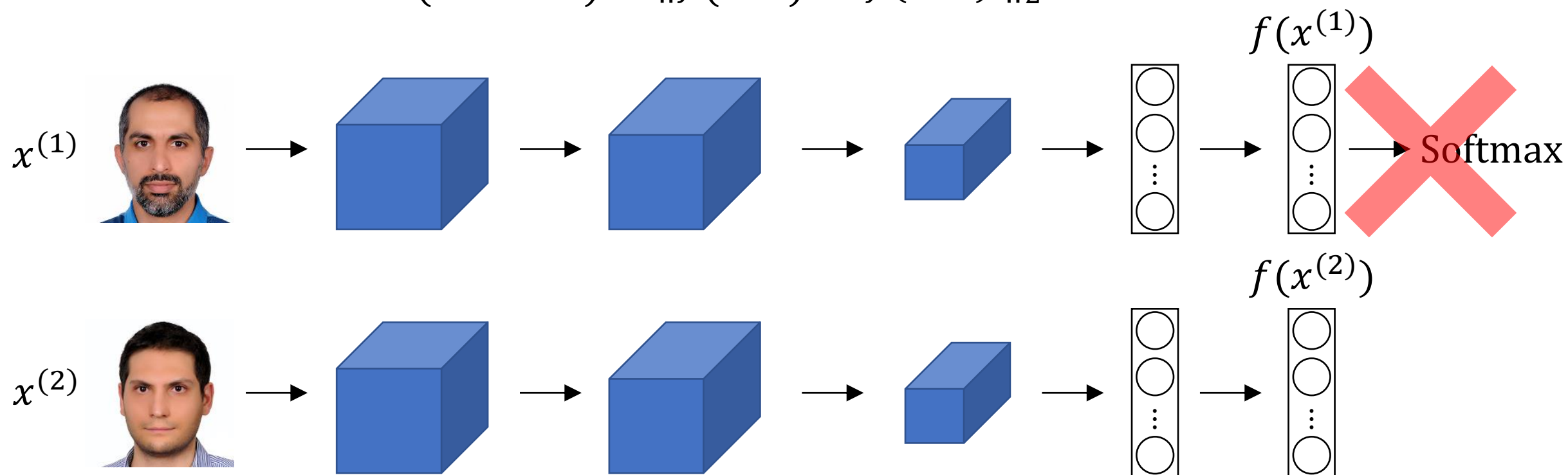


یادگیری تابع شباهت



• $d(img1, img2)$: درجه تفاوت بین دو تصویر

$$d(x^{(1)}, x^{(2)}) = \|f(x^{(1)}) - f(x^{(2)})\|_2^2$$

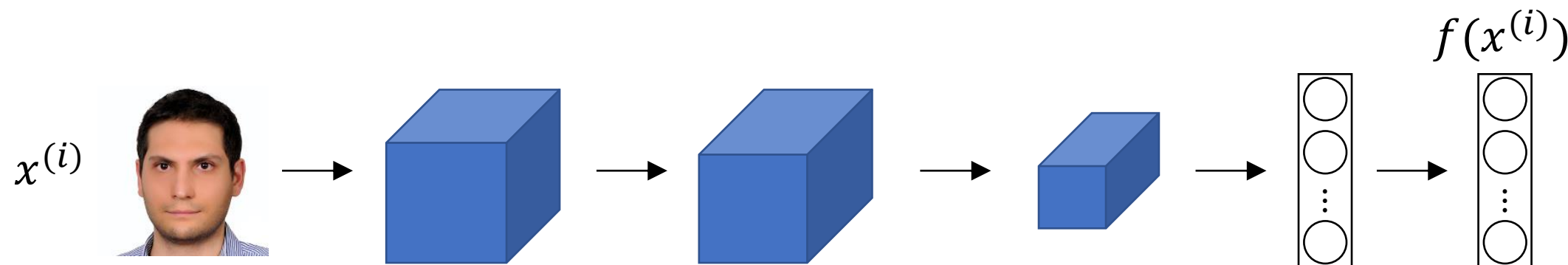


یادگیری تابع شباهت









- $d(img1, img2)$: درجه تفاوت بین دو تصویر

$$d(x^{(1)}, x^{(2)}) = \|f(x^{(1)}) - f(x^{(2)})\|_2^2$$

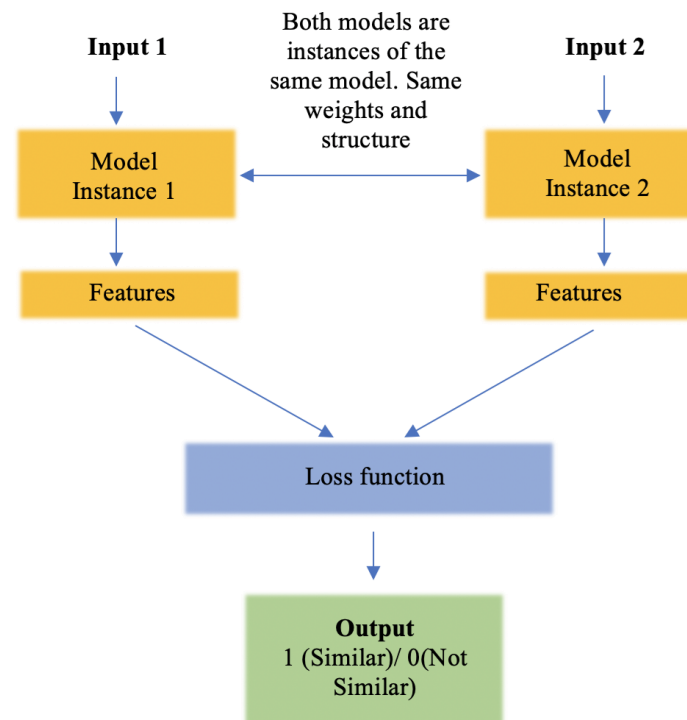
- پارامترهای شبکه آموزش می‌بینند تا:
 - اگر $x^{(i)}$ و $x^{(j)}$ مربوط به یک نفر باشند، $d(x^{(i)}, x^{(j)})$ عدد کوچکی باشد
 - اگر $x^{(i)}$ و $x^{(j)}$ مربوط به افراد متفاوتی باشند، $d(x^{(i)}, x^{(j)})$ عدد بزرگی باشد



شبکه Siamese

$x^{(1)}$	$x^{(2)}$	y
		0
		1
		0
		1

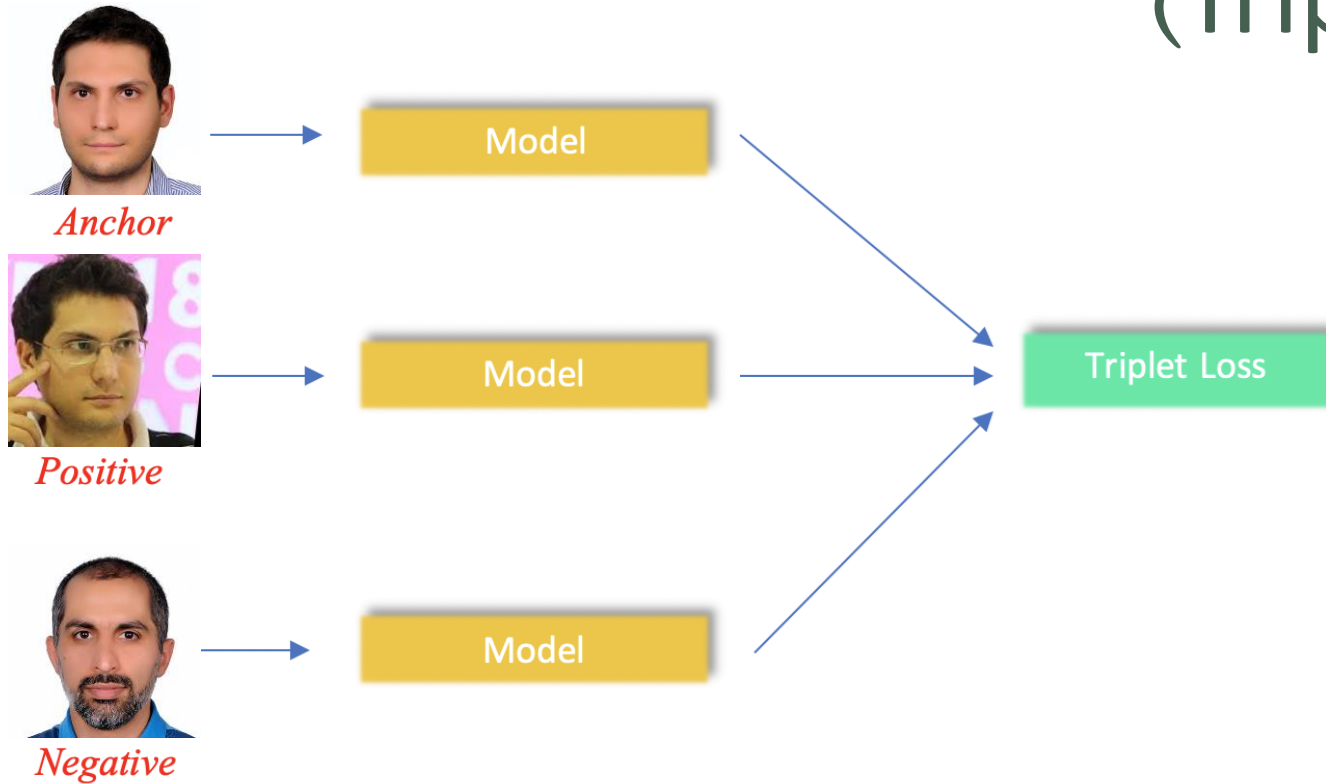
- یک شبکه Siamese کلاسی از شبکه‌های عصبی است که شامل یک یا چند شبکه یکسان است



تابع ضرر سه تایی (Triplet Loss)





• مدل سه ورودی می گیرد:

- Anchor، Positive و Negative

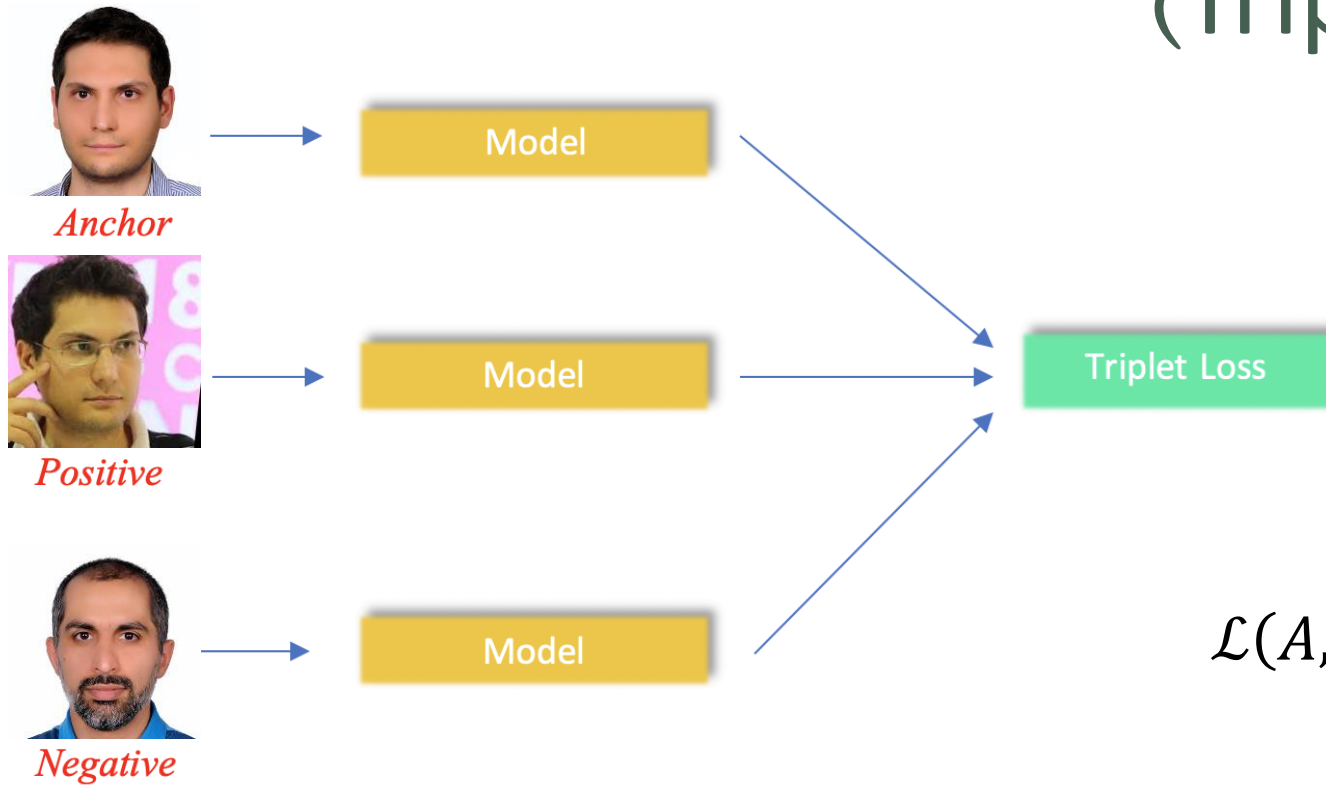


margin

$$d(\text{Anchor}, \text{Positive}) + \alpha \leq d(\text{Anchor}, \text{Negative})$$

$x^{(1)}$	$x^{(2)}$	y
		0
		1

تابع ضرر سه‌تایی (Triplet Loss)

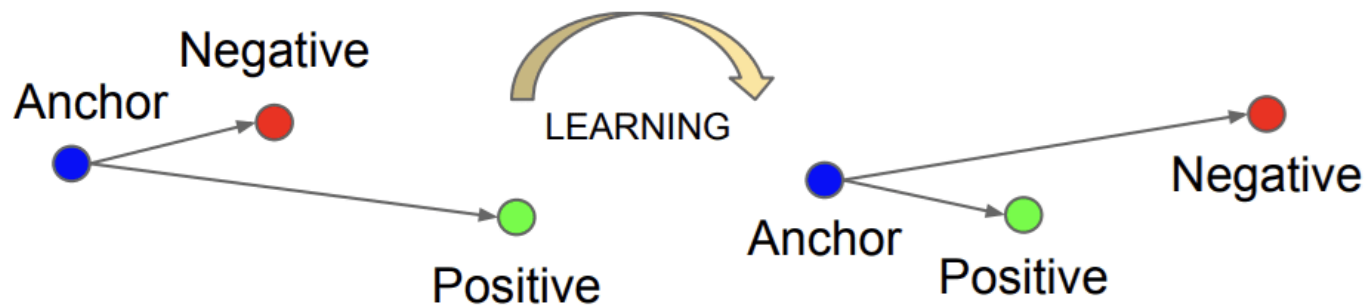


$$d(A, P) + \alpha \leq d(A, N)$$

$$d(A, P) - d(A, N) + \alpha \leq 0$$

$$\mathcal{L}(A, P, N) = \max(d(A, P) - d(A, N) + \alpha, 0)$$

$$J = \sum_{i=1}^M \mathcal{L}(A^{(i)}, P^{(i)}, N^{(i)})$$



تابع ضرر سه‌تایی (Triplet Loss)

$$d(A, P) + \alpha \leq d(A, N)$$

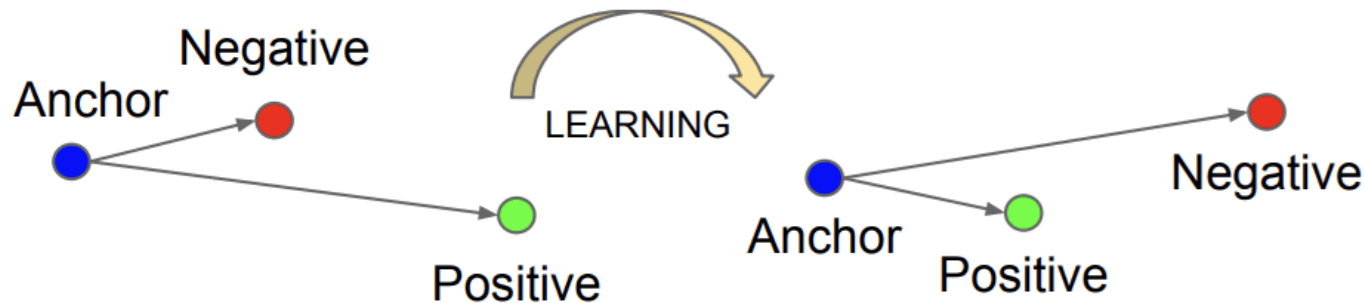
$$d(A, P) - d(A, N) + \alpha \leq 0$$

$$\mathcal{L}(A, P, N) = \max(d(A, P) - d(A, N) + \alpha, 0)$$

$$J = \sum_{i=1}^M \mathcal{L}(A^{(i)}, P^{(i)}, N^{(i)})$$

- برای نمونه‌های مربوط به یک کلاس، جانمایی‌هایی تولید می‌شود که فاصله کمی داشته باشند

- این سیستم مستقیماً یک نداشت از تصاویر چهره به فضای فشرده اقلیدسی را می‌آموزد که در آن فاصله‌ها مستقیماً با معیار تشابه چهره‌ها مطابقت دارند



انتخاب سه تایی ها

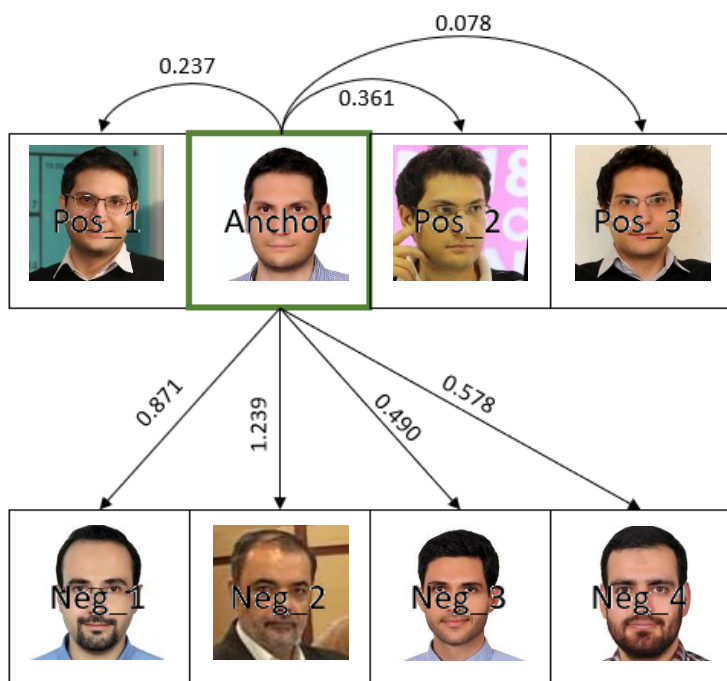
- سه تایی هایی که برای آموزش مدل استفاده می شوند باید با دقت انتخاب شوند
- اگر به صورت تصادفی انتخاب شوند

- شرط $d(A,P) + \alpha \leq d(A,N)$ به راحتی برآورده می شود

- تابع ضرر کوچک خواهد بود و به روزرسانی مدل به کندی انجام خواهد شد

- سه تایی ها به صورت آنلاین و برحسب فاصله فعلی با یکدیگر تولید می شوند

- به آنها مثبت سخت (hard positive) و منفی سخت (hard negative) گفته می شود



FaceNet

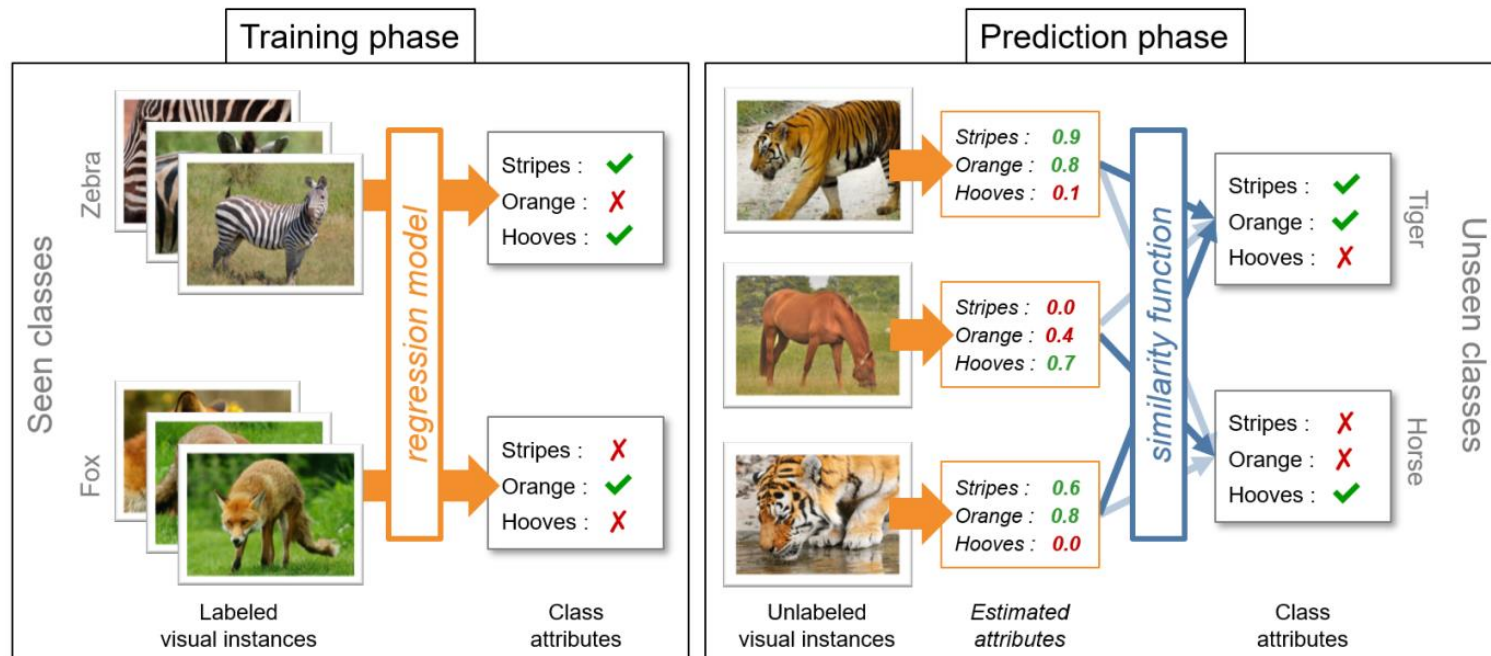
- CASIA-Webface (10K ids/0.5M images)
- VGG2 (9K ids/3.31M images)
- Glint360K (360K ids/17M images)

- LFW (5749 ids/13233 images/6K pairs)

Model name	LFW accuracy	Training dataset	Architecture
20180408-102900	0.9905	CASIA-WebFace	Inception ResNet v1
20180402-114759	0.9965	VGGFace2	Inception ResNet v1

یادگیری بدون نمونه (Zero-Shot Learning)

- هیچ نمونه برچسب‌داری به شبکه داده نمی‌شود!
- به عنوان مثال، این مسئله این را در نظر بگیرید که یک مدل دارای یادگیری، مجموعه بزرگی از متن را بخواند و سپس مسئله شناسایی اشیاء را حل کند



- به عنوان مثال، دانستن اینکه "گره‌ها چهار پا دارند" یا "گره‌ها گوش‌های نوک تیز دارند" برای شناسایی آن کمک‌کننده است