---

layout: post

title: دو مقاله ای که به هوش مصنوعی این امکان را دادند که صدا،متن وبینایی را به هم پیوند زدند

---

اگه ما میخواهیم رباط هایی که در آینده ساخته میشوند فرمانپذیر باشند، نیازهست که آنها محیط اطرافشان را کاملا درک کنند . به عنوان مثال اگریک رباط سروصدایی شبیه پارس کردن یک سگ را میشنود باید بداند که چه چیزی آن را ایجاد میکند؟ باید بداند که یک سگ چه شکلی است و به چه چیزی نیاز دارد؟

تحقیقات هوش مصنوعی که امروزه با آنها سر و کار داریم توانایی هایی مثل بازتشخصی عکس ، تشخیص نویز و درک متن را دارند اما این سه قابلیت به صورت جدا از یکدیگر عمل میکنند.

فرض کنید شما فقط میتوانیستید در هر لحظه از یکی از حواس خود استفاده کنید و نمیتوانستید چیزهایی که میشنوید را با چیزهایی که دیده اید را انطباق دهید! این هوش مصنوعی است که ما امروزه میبینیم، و شاید میتواند بخشی از این دلیل باشد که ما هنوز نتوانسته ایم الگوریتمی بسازیم که بتواند مثل انسان یاد بگیرد .

اما دو مقاله که اخیرا توسط [گوگل](https://arxiv.org/abs/1706.05137) و [ام ای تی](https://arxiv.org/abs/1706.00932) منتشر شده مدلی را پیشنهاد میدهد که شاید بتوان از آنها به عنوان اولین گام های ما برای کمک به هوش مصنوعی برای دیدن ، شنیدن و خواندن یاد کرد.

مهم نیست اگه شما صدای موتور یک ماشین را میشنوید یا آن را میبینید، شما ذهن شما فورا و به صورت طبیعی میتواند مفهوم مشاهبی را درک کند.

مفاهیمی که در این دو مقاله توسط نویسندگان توضیح داده می شوند گویای این حقیقت است که ما چیز جدیدی به الگوریتم ها یاد نمیدهیم، اما در عوض راهی را پیشتهاد میدهند که اطلاعات از یک حس را به حس دیگر پیوند میدهند. به عنوان مثال خودرو راننده خودکار را در نظر بگیرید که صدای آزیر آمبولانس را قبل از دیدن آن میشنود ، با داشتن دانش لازم در مورد اینکه صدای آمبولانس چگونه است، این اجازه را به خودرو راننده خودکار میدهد که خود را برای خارج کردن از مسیر آن آماده کند.

برای آموزش چنین سیستمی، گروه ام ای تی ابتدا به شبکه عصبی فریم هایی از ویدیو را نشان دادند که به همراه صدا بودند. شبکه ابتدا اشیاءموجود در ویدئو و صدای آنها را پیدا می کند ، سپس سعی میکند که صدا ها را با اشیاء مرتبط کند .

![\_config.yml]({{ site.baseurl }}/images/4/1.png)

در شکل بالا ورودی هایی که نرون های پنهان را بیشتر تحت تاثیر قرار میدهند نمایش داده شده است. ادراک چند حسی "اشاره به حواس پنجگانه انسان دارد" نقش کلیدی در [سیستم ادراک انسان](https://en.wikipedia.org/wiki/Perceptual\_system) بازی میکند.

برای آموزش مدل نیاز به داده هایی اعم از متن، صدا و تصویر است که برای این امر از دیتابیسی که قبلا توسط [این مقاله](https://arxiv.org/abs/1503.01817) جمع آوری شده است استفاده شد .(حجم این دیتا بیس 14 گیگا بایت است که بر روی [AWS](https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67) قرار گرفته است. این دیتا بیس مربوط به ویدئوهای Flickr است که تحت [مجوز کرییتیو کامنز](https://en.wikipedia.org/wiki/Creative\_Commons\_license) است "یک مجوز نرم‌افزار است که وب‌گاه‌های معروفی مانند ویکی‌پدیا، فری‌بیس از آن استفاده می‌کنند".

در این مقاله بیش از 750000 ویدئو دانلود و صدا از تصویر استخرا شد . برای آموزش زبان دو دیتا بیس [COCO]( http://mscoco.org) که حاوی 400000 جمله و 80000 عکس و دیتا بیس [Visual Genome]( https://arxiv.org/abs/1602.07332) که حاوی 4200000 جمله و 100000 عکس هستند، استفاده شد. برای آموزش تصویر از تصاویر موجود در سه دیتا بیس قبلی استفاده میشود که نزذیک به یک میلون عکس است که با متن یا صدا همگام میشوند.

مدل ارایه شده توسط گروه ام ای تی، سه ورودی متفاوت دارد و هر کدام مسیر جداگانه را در ابتدای شبکه دارند. در انتها ،مسیر آنها به یک لایه مشترک همگرا میشوند که بین سایر بخش های شبکه به اشتراک گذاشته میشود.

![\_config.yml]({{ site.baseurl }}/images/4/2.png)

در ادامه کار این تیم عکس هایی را با زیرنویس به همان الگوریتم دادند، و شبکه توانست کلمات را با اشیاء مربوطه پیوند دهید . این ایده مشابه ایده قبلی است : ابتدا شبکه به صورت جداگانه همه اشیاء و کلماتی که در عکس میتواند پیدا کند را تشخصی میدهد و در نهایط آنها را با یک دیگر مرتبط میکند.

هر چند با توجه به توضیحات موجود در مقاله، مدل پیشنهادی به نظر میرسد که نتوانست به صورت فوق العاده کار کند اما این برای اولین بار است که مدلی بر روی عکس/صدا و عکس/متن آموزش داده می شود.

مدلی که توسط محققین گوگل پیشنهاد شد به صورت مشابهی عمل میکند، باستثنای اینکه این مدل قادر است که متن را نیز ترجمه کند اما اطلاعاتی کمی در مورد چگونگی کارکرد مدل در مقاله آنها آورده شده است.هر چند الگوریتم به صورت آنلاین در اختیار سایر محقیقن قرار گرفته است.

هیچ کدام از دو الگوریتم پیشنهادی توسط این دو تیم تحقیقاتی نتوانستند بهتر از الگوریتم های تک کاره موجود عمل کنند، اما این محققین عقیده دارند که به این موارد بصورت دراز مدت باقی نخواهد ماند.