
**دانشکده مهندسی کامپیوتر**

# پرسش و پاسخ تصویری

**آزمایشگاه پردازش زبان طبیعی دانشگاه علم و صنعت ایران**

**سارا کدیری**

**نام استاد کارآموزی:**

**دکتر طاهر پیله‌ور**

**پاییز و زمستان ۱۳۹۹**

# تأییدیه‌ی صحت و اصالت نتایج

بسمه تعالی

اینجانب سارا کدیری به شماره دانشجویی ۹۶۵۲۱۴۴۳ دانشجوی رشته مهندسی کامپیوتر مقطع تحصیلی کارشناسی تأیید می‌نمایم که کلیه‌ی مطالب مندرج در این گزارش حاصل ۳۰۰ ساعت حضور و کار اینجانب در شرکت/کارخانه آزمایشگاه پردازش زبان طبیعی دانشگاه علم و صنعت و بدون هرگونه دخل و تصرف است و موارد نسخه‌برداری‌شده از آثار دیگران را با ذکر کامل مشخصات منبع ذکر کرده‌ام. درصورت اثبات خلاف مندرجات فوق، به تشخیص دانشگاه مطابق با ضوابط و مقررات حاکم آموزشی، پژوهشی و انضباطی با اینجانب رفتار خواهد شد و حق هرگونه اعتراض درخصوص احقاق حقوق مکتسب و تشخیص و تعیین تخلف و مجازات را از خویش سلب می‌نمایم.

نام و نام خانوادگی: سارا کدیری

امضا و تاریخ: اردیبهشت ۱۴۰۰

## تشکر و قدردانی:

با تشکر از استاد سید صالح اعتمادی برای فراهم کردن بستر کارآموزی بین دانشجویان کارشناسی و ارشد، و خانم مریم سادات هاشمی برای راهنمایی‌های دلسوزانه در اختیار قرار دادن هر آنچه می‌دانند.

# چکیده

از زیرشاخه‌های حوزه‌ی هوش مصنوعی، می‌توان به پردازش زبان طبیعی و بینایی ماشین اشاره کرد. تعامل این دو موضوع را با هم می‌توان در پرسش و پاسخ تصویری به خوبی مشاهده کرد. توسعه‌ی ابزارهای مرتبط با این موضوع، کمک شایانی به کم‌بینایان و نابینایان خواهد بود.

این دوره‌ی کارآموزی متمرکز بر این موضوع بود. ابتدا برای پرسش و پاسخ تصویری، روند تولید و ارزیابی دادگان فارسی را پیش گرفتیم. در آخر، با بررسی مدل‌های از پیش آموزش داده شده، پژوهشی در راستای کم حجم‌تر کردن آن‌ها برای کارایی بیشتر انجام دادیم.

**واژه‌های کلیدی:** پردازش زبان طبیعی، پرسش و پاسخ تصویری، NLP، Visual Question Answering، VQA.

# فهرست مطالب

# فهرست جداول

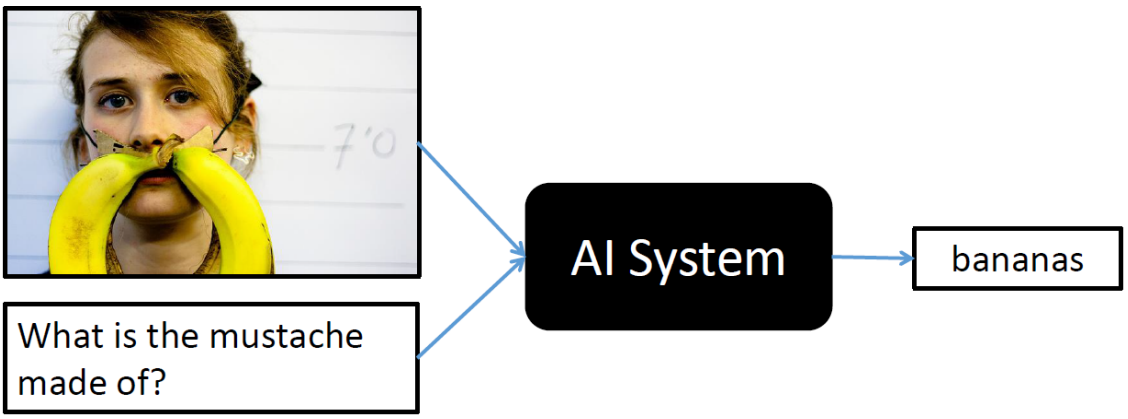
# فصل ۱
# معرفی حوزه کارآموزی

## ۱-۱ آزمایشگاه پردازش زبان طبیعی

آزمایشگاه پردازش زبان طبیعی به سرپرستی دکتر اعتمادی در دانشکده کامپیوتر دانشگاه علم و صنعت پروژه‌های بسیاری در رابطه با هوش مصنوعی و پردازش زبان طبیعی در دست دارد. اکثر این پروژه‌ها متعلق به دانشجویان مقاطع بالاتر است که بر روی پایان‌نامه‌هایشان کار می‌کنند. خوشبختانه این بستر فراهم شد تا دانشجویان کارشناسی به عنوان کارآموز به این پروژه‌ها اضافه شوند تا هم کار تحقیقاتی را یاد بگیرند و هم پروژه‌ها سریعتر انجام شوند. با همه‌گیری ویروس کرونا و مجازی شدن همه‌ی فعالیت‌ها، جلسات مربوط به این آزمایشگاه نیز به محیط Microsoft Teams منتقل شد.

## ۱-۲ پرسش و پاسخ تصویری

در سال های اخیر پیشرفت های زیادی در مسائل هوش مصنوعی و یادگیری عمیق که ترکیبی از دو حوزه‌ی جدا هستند، مورد توجه قرار گرفته‌اند. یکی از این مسائل، پرسش و پاسخ تصویری است که تقاطع بینایی ماشین و پردازش زبان طبیعی است. اخیراً با توجه به گسترش آن، به عنوان یک مسئله‌ی AI-Complete شناخته شده است که می‌تواند جایگزین تست بصری تورینگ[۱] باشد. در رایج ترین شکل، این یک کار چالش‌برانگیز چند حالته[۲] است که در آن یک کامپیوتر یک تصویر و یک سوال به عنوان ورودی دریافت می‌کند، و خروجی مورد انتظار، پاسخ آن سوال با توجه به محتوای تصویر است.




شکل (۱-۱) مثالی از پرسش و پاسخ تصویری

[۱] Visual Turing Test

[۲] Multi-Modal

# فصل ۲
# مشروح فعالیت های انجام شده در محل استقرار

## ۲-۱ آشنایی با VQA

خانم مریم سادات هاشمی دانشجوی سال دوم ارشد هوش مصنوعی در دانشگاه علم و صنعت هستند. برای سمینار کارشناسی ارشد پروژه‌ی پرسش و پاسخ تصویری یا همان Visual Question Answering (که به اختصار VQA گفته می‌شود) را انتخاب کردند و دو نفر از دانشجویان کارشناسی را به عنوان کارآموز پذیرفتند. در جلسه‌ی اول که در شهریور ماه سال ۹۹ برگزار شد، با تسک VQA بیشتر آشنا شدیم. خانم هاشمی از انگیزه‌ی خود برای انتخاب این موضوع گفتند که کمک به نابینایان بود. این تسک در زبان انگلیسی دادگان مناسبی دارد ولی در فارسی اینطور نیست. به همین دلیل، راهکارهای جمع‌آوری دادگان بررسی شدند.

## ۲-۲ گذراندن دوره‌ی آموزشی آنلاین

با توجه به این که تا شروع دوره‌ی کارآموزی، تنها درس مربوطی که گذرانده بودم درس هوش مصنوعی بود، قرار بر آن شد که دوره‌ای آنلاین برای آشنایی و یادگیری بیشتر و بهتر با موضوعات مربوط (شبکه‌های عصبی و یادگیری عمیق) گذرانده شود. یکی از معتبرترین دوره‌ها، از سایت deeplearning.ai به نام Deep Learning Specialization است که مدرس آن Andrew Ng از دانشگاه استنفورد است. این دوره، شامل بخش‌های زیر است.

- Neural Networks and Deep Learning
- Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization
- Structuring Machine Learning Project
- Convolutional Neural Networks
- Sequence Models

علاوه بر ویدیوهای آموزشی، تمرین‌ها و کوییزهای آن نیز بسیار مفید بودند.

## ۲-۳ پرسش و پاسخ تصویری به زبان فارسی

هدف اصلی این کارآموزی، تولید دادگان فارسی برای کار VQA بود. برای این کار، روش‌های مختلفی وجود دارد. در ابتدا با مطالعه‌ی دادگان انگلیسی، ساختار مناسب برای این تسک بدست می‌آید. پس از آن، روش‌های جمع‌آوری داده بررسی می‌شوند. سپس این دادگان جمع‌آوری شده باید ارزیابی شوند و در آخر، نمونه‌ای از کاربرد این دادگان نشان‌داده شوند. برای جمع‌آوری داده، سه راه کلی در این دوره بررسی شد.

### ۲-۳-۱ ترجمه دادگان VQA v۱ به فارسی

دانشـــگاه‌های Virginia Tech و Georgia Tech با همکاری هم، همواره مشـــغول آزمایش‌های مربوط به پرسش و پاسخ تصویری هستند و نسخه‌های متعددی از دادگان‌های مربوط به پرسش و پاسخ تصویری را منتشر کرده و می‌کند. دادگان VQA v۱ [۱] برای بررسی انتخاب شد. این دادگان از تصاویر MSCOCO [۲]استفاده کرده است و برای هر تصویر سه نوع سوال طراحی شده است. نوع اول سوالات، سوالات ساده‌ی "بله یا خیر" هستند. نوع دوم سوالات، شمارشی هستند و نوع آخر، شامل سوالاتی است که به دسته‌ی اول و دوم متعلق نیستند.

اگر همین جملات به زبان فارسی وجود داشته باشند، دادگان کامل است. به همین دلیل به عنوان اولین قدم، با استفاده از سرویس‌های ترجمه ماشینی گوگل[۱] و ترگمان[۲]، اولین نسخه‌ی دادگان فارسی تهیه شدند.

### ۲-۳-۲ طراحی اپلیکیشن تحت وب

راه دیگر تهیه دادگان، استفاده از طراحی وب‌سایت و درخواست از کاربران برای ارائه‌ی پرسش و پاسخ مناسب برای عکس‌های از پیش تعیین شده است. این ایده نیازمند حامی مالی بود که برای کاربران نفع داشته باشد و انگیزه‌ای شود تا به وب‌سایت مراجعه کنند و به ازای کد تخفیف سرویس‌های مختلف، کمکی به جمع‌آوری دیتای فارسی کنند. با توجه به نداشتن بودجه‌ی مناسب و همچنین سخت بودن توسعه و راه‌اندازی وب‌سایت، این روش به عنوان آخرین گزینه قرار گرفت.

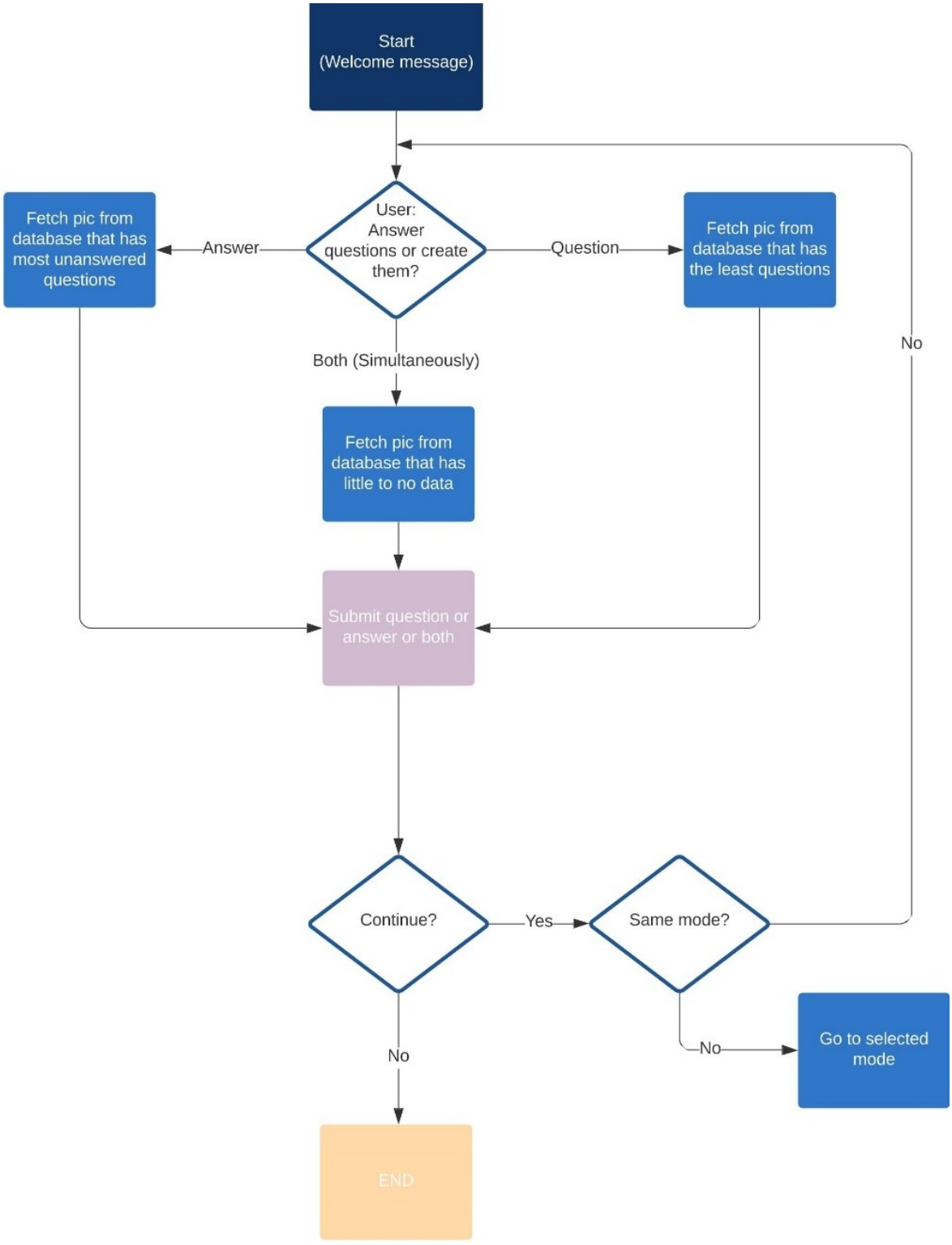
### ۲-۳-۳ استفاده از بات تلگرام

روش سوم، پیاده‌سازی بات تلگرام است. مزیت این روش، آسانی توسعه و کار با API آن است. همچنین، این بات برای کاربران در دسترس‌تر است و لازم نیست به وبسایت جدیدی مراجعه کنند و در مراجعات روزانه به اپلیکیشن تلگرام به آن برخورد می‌کنند. شکل زیر، روند کلی بات نهایی را نشان می‌دهد.

---

۱ translate.google.com

۲ targoman.ir




شکل (۱-۲) فلوچارت روند بات تلگرام

این بات در نهایت با استفاده از زبان پایتون پیاده‌سازی شد. برای همیشه در دسترس بودن آن، یک سرور مورد نیاز است، که متاسفانه به دلیل کمبود منابع و بودجه اکنون در دسترس نیست.

## ۲-٤ ارزیابی منابع ترجمه ماشینی مختلف

برای ارزیابی ترجمه‌ی ماشـینی روش اول جمع‌آوری داده، ابتدا ۵۰ سـوال تصـادفی از مجموعه سـوال‌های دادگان انگلیسی VQA v۱ به هر یک از اعضای گروه داده شد تا به صورت دستی ترجمه شوند. این سوال‌ها سـپس با سـرویس ترجمه‌ی گوگل و ترگمان مورد ارزیابی قرار گرفتند تا ترجمه‌ی بهتر انتخاب شـود. به طور کلی، برای ارزیابی ترجمه ماشـینی معیارهایی وجود دارند که بین اعضـای گروه تقسـیم شـدند. مسـئولیت بررسی دو معیار به عهده‌ی من بود که به شرح زیر است.

- معیار WER[۱]: این معیار کلمه به کلمه‌ی ترجمه‌ی درسـت و ترجمه‌ی ماشـینی را با هم مقایسـه می‌کند و فاصله‌ی لونشتاین[۲] آن‌ها را محاسبه می‌کند.

$$WER = \frac{\text{فاصله لونشتاین}}{\text{تعداد کل کلمات}} \qquad (۱-۲)$$

- معیار PER[۳]: این معیار مشـابه معیار قبلی کار می‌کند با این تفاوت که ترتیب کلمات در جمله بر روی امتیاز نهایی تاثیری ندارد.

طبق هر دو معیار بالا، ترجمه‌ی گوگل ترجمه‌ی مناسـب‌تری از ترگمان بود و برای مرحله‌ی بعدی پروژه، از این ابزار گوگل برای ترجمه ماشینی استفاده شد.

## ۲-۵ بهبود واسط کاربر استفاده از پرسش پاسخ تصویری

برای نشان دادن کاربرد و نمونه عملکرد تسک پرسش و پاسخ تصویری فارسی، به یک واسطه نیاز بود که از پیش برای پروژه‌ی درس یادگیری عمیق، توسط خانم هاشـمی و هم‌گروهی ایشـان آقای اصـغری تهیه شـده بود. در جلسـات گروهی ظاهر آن را بهبود بخشـیدیم و بر روی سـرور بارگذاری کردیم. نمونه عملکرد این واسطه در شکل (۲-۲) نشان داده شده است.

---

[۱] Word Error Rate

[۲] Levenshtein Distance

[۳] Position Independent Word Error Rate


شکل (۲-۲) نمونه پرسش و پاسخ در واسط کاربری

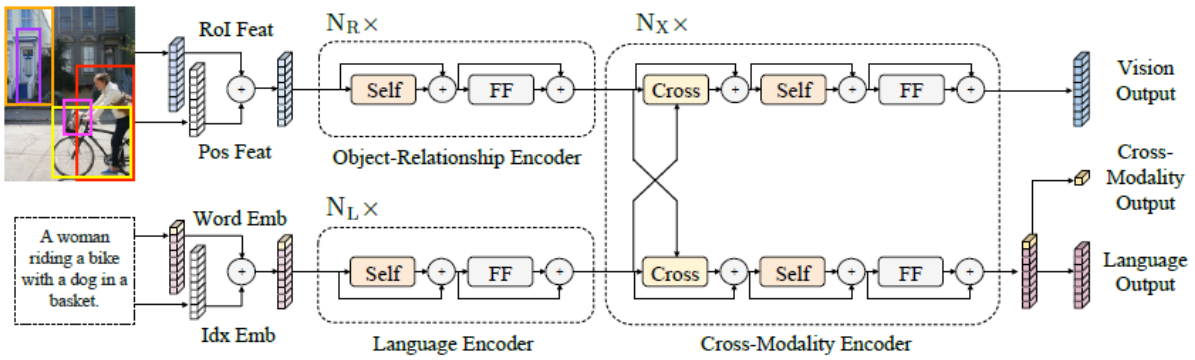
## ۲-۶ پژوهش در راستای انتشار مقاله در کارگاه SRW

کارگاه SRW[۱] سالانه در همایش‌های مختلف پردازش زبان‌های برگزار شده و فرصتی برای دانشجویان و محققین جوان این حوزه برای ارسال و ارائه مقاله خود و گرفتن بازخورد از اساتید و محققین با سابقه این حوزه است. شرط ارسال مقاله برای این همایش این است که نویسنده‌ی اول مقاله، دانشجو باشد. در قسمت پایانی دوره‌ی کارآموزی، تصمیم بر آن شد که مقاله‌ای برای این کنفرانس ارسال کنیم تا تجربه‌ای در مقاله‌نویسی بدست بیاید. موضوع انتخاب شده LXMERT Model Compression for Visual Question Answering است. هدف از این مقاله، هرس مدل LXMERT [۳] و بررسی تاثیر این کار بر دقت تسک VQA بود.

[۱] Student Research Workshop

### ۲-۶-۱ مقدمه و هدف پژوهش

اخیرا پیشرفت قابل توجهی در شبکه‌های از پیش آموزش داده شده[۱] مشاهده شده است. یکی از مشکلات اصلی این شبکه‌ها، این است که بیش از حد پارامتر برای آموزش دارند، مخصوصا اگر چند حالته باشند. برای حل این مشکل از روش‌های هرس شبکه‌های عصبی استفاده می‌شود. هدف از این کار این است که این شبکه‌ها در انواع بیشتری از سخت‌افزار (مانند سیستم‌های نهفته، گوشی‌های موبایل و غیره) قابل استفاده باشند. یکی از شبکه‌های چند حالته که برای تسک VQA نتایج خوبی به دنبال داشته، شبکه‌ی LXMERT [۳] است.




شکل (۳-۲) معماری شبکه‌ی LXMERT [۳]

### ۲-۶-۲ هرس کردن شبکه LXMERT برای پرسش پاسخ تصویری

از بین روش‌های هرس کردن، نظریه‌ی بلیت برنده [۴] یا Lottery Ticket Hypothesis که به اختصار LTH گفته می‌شود، انتخاب شد. برای این کار، در هر دور اجرای برنامه ۱۰ درصد از کمترین وزن‌های موجود در آن، کنار گذاشته می‌شوند تا با آن‌ها زیر شبکه‌ی[۲] بد تولید شود. با وزن‌های باقی‌مانده (۹۰٪ برتر) زیر شبکه‌ی خوب تولید می‌شود. نوع دیگر تولید زیر شبکه نیز انتخاب تصادفی وزن‌هاست. انجام این آزمایش‌ها با استفاده از کدی که نویسندگان مقاله‌ی LTH نوشته بودند و در بستر EvalAI[۳] انجام شد. این کد با کتابخانه‌ی PyTorch[۴] نوشته شده بود ولی ما به دلیل اینکه برای پیاده‌سازی شبکه‌ی LXMERT از

---

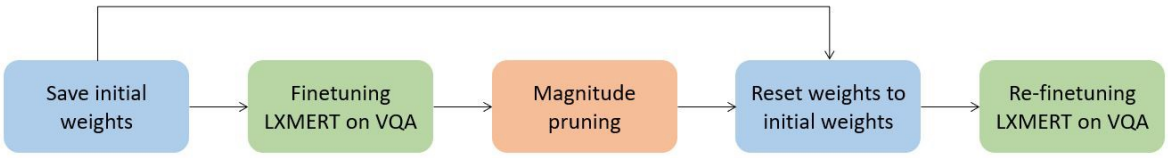
۱ Pre-trained

۲ Sub-network

۳ eval.ai

۴ pytorch.org

Keras[1] استفاده کرده بودیم، به استفاده از آن ادامه دادیم. به همین دلیل با برخی مشکلات مواجه شدیم و نتایج یکسانی با نویسندگان مقاله بدست نیاوردیم که طی مکاتبه با ایشان، این مشکلات حل شدند. روند کلی فرایند هرس شبکه عصبی در شکل زیر قابل مشاهده است.




شکل (۴-۲) روند هرس شبکه عصبی

### ۲-۶-۳ نگارش مقاله

عمده‌ترین مسئولیت من در روند مقاله، نگارش آن بود. مقاله باید در قالب لاتک قرار داده شده در سایت کنفرانس و به زبان انگلیسی می‌بود. در این مرحله، نگارش علمی انگلیسی و کار تخصصی‌تر با لاتک از جمله مهارت‌هایی بودند که بهتر یادگرفته شدند.

[1] keras.io

# فصل ۳
# دیدگاه‌های شخصی

## ۳-۱ چگونگی انتخاب دوره

پس از گذشتن سه سال از دوران کارشناسی مهندسی کامپیوتر، این نیاز را حس کردم که باید تاثیر مطالبی را که تا الان یاد گرفته‌ام را در دنیای واقعی ببینم، یا حداقل با یک مسئله‌ی واقعی مواجه شوم. هدف نهایی از کارآموزی نیز همین موضوع است: دیدن بخشی کوچک از آنچه بعد از فارغ‌التحصیلی در انتظارمان است. من هیچ وقت از انتخاب موضوعات تحصیلی‌ام مطمئن نیستم ولی از این مطمئن بودم که تجربه‌ای خارج از صرفِ پاس کردن دروس نیاز دارم. در پایان سال سوم با این که اصلا از معلومات و حتی علایقم مطمئن نبودم و همیشه این حس را داشتم که چیزهایی که می‌دانم کافی نیستند، تصمیم گرفتم در آزمایشگاه دانشکده مشغول به کار شوم تا هم مطالب بیشتری را یاد بگیرم، هم علایقم را بهتر بشناسم. پروژه‌ی VQA بهترین انتخاب برای من بود زیرا تا آن زمان، به پردازش زبان طبیعی و بینایی ماشین علاقه‌مند بودم و این پروژه محل تلاقی هر دوی این موضوع‌ها بود.

## ۳-۲ چالش‌ها و پیشنهادها

کار در این حوزه، و هر حوزه‌ی یادگیری عمیق، در ایران به دلیل نبود دسترسی به شبکه‌ی مالی بین‌المللی، نداشتن زیرساخت مناسب و کمبود بودجه مربوط به کارهای تحقیقاتی، بسیار سخت است. با این حال، اساتید و محقق‌های زیادی دلسوزانه سعی در رفع این مشکلات دارند. کارآموز این حوزه باید آمادگی این چالش‌ها را داشته باشد تا بتواند برای نتیجه‌بخش بودن کار، زمان‌بندی مناسبی در نظر بگیرد.

## ۳-۳ دید کلی نسبت به دوره پس از پایان آن

من بسیار از تجربه‌ام راضی بودم و وقت‌هایی که کار جلو نمی‌رفت یا بسیار سخت می‌شد، چون محصول و نتیجه‌ی نهایی را دوست داشتم، به تلاش ادامه می‌دادم. علاوه بر مهارت‌های سخت مانند استفاده از فریم‌ورک‌های مختلف و یا حتی استفاده از لاتک، مهارت‌های نرم مانند کارگروهی مجازی و ارتباط با دیگر همکاران، نیاز به تقویت داشتند. تجربه‌هایی از این دست سخت بدست می‌آیند و بسیار خوشحالم که این فرصت برایم فراهم شد.

# فصل ۴
# مراجع

[١] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh 2015. VQA: Visual Question Answering. In International Conference on Computer Vision (ICCV).

[٢] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. (2015). Microsoft COCO: Common Objects in Context.

[٣] Tan, H., and Bansal, M. 2020. LXMert: Learning cross-modality encoder representations from transformers. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, p.5100–5111.

[٤] Jonathan Frankle, and Michael Carbin 2019. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In International Conference on Learning Representations.