تمرین سوم : Image Captioning with CNN-RNN Architecture

رضا حاجی علیزاده

بخش مهندسی کامپیوتر دانشگاه شهید باهنر کرمان

rezahajializadeh@eng.uk.ac.ir

۱. معرفی و هدف پروژه

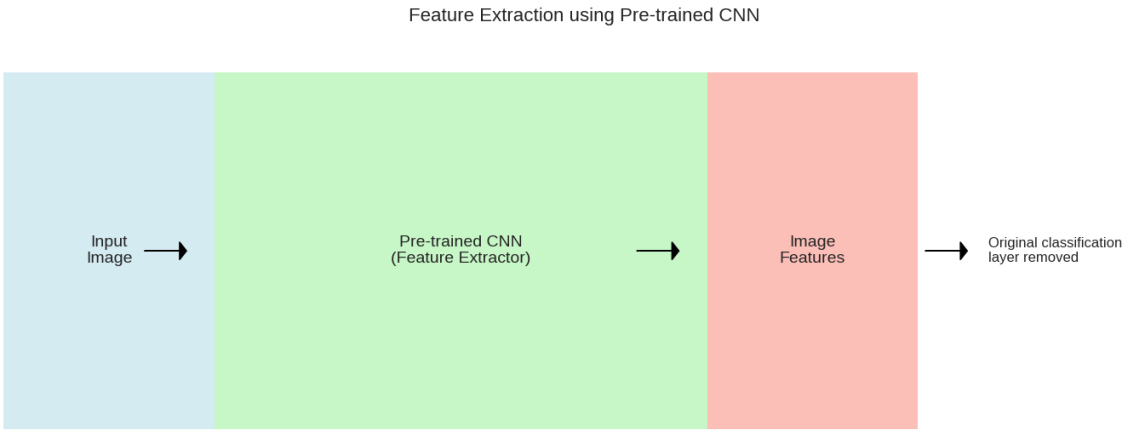
مسئله ای که در این پروژه بررسی می شود توصیف تصاویر (Image Captioning) است. هدف این است که مدلی طراحی و آموزش داده شود که با دریافت یک تصویر به عنوان ورودی بتواند جمله ای معنی دار و توصیفی از محتوای آن تصویر تولید کند. با توجه به پیچیدگی بالای این مسئله به دلیل نیاز به درک بصری دقیق از تصویر و همچنین تولید عبارت های زبان طبیعی، مدل هایی که برای حل آن استفاده می شوند نیاز به ترکیب ویژگی های شبکه های عصبی کانولوشنی (CNNs) برای استخراج ویژگی های بصری و شبکه های عصبی بازگشتی (RNNs) برای مدل سازی توالی کلمات دارند. در این پروژه از یک معماری ترکیبی مبتنی بر CNN-RNN استفاده می‌شود که ابتدا با استفاده از یک شبکه عصبی کانولوشنی از پیش آموزش دیده شده مانند ResNet یا MobileNet ویژگی‌ها سطح بالا از تصویر استخراج می شود و سپس این ویژگی ها به عنوان ورودی به یک شبکه عصبی بازگشتی مثل GRU یا LSTM داده می شود که وظیفه تولید گام به‌گام کلمات توصیف را دارد. در نهایت مدل باید قادر باشد تا با دیدن تصویر جدیدی که قبلاً مشاهده نکرده است یک جمله توصیفی جدید و معنادار تولید کند.

۲. جزئیات معماری مدل

مدلی که در این پروژه برای تولید توصیف تصاویر طراحی شده از یک معماری ترکیبی Encoder-Decoder استفاده می کند. این معماری شامل دو مؤلفه‌ی اصلی است:

- رمزگذار (Encoder): برای درک محتوای تصویر ابتدا باید ویژگی های بصری سطح بالای آن استخراج شوند، برای این منظور از یک شبکه عصبی کانولوشنی از پیش آموزش دیده به عنوان رمزگذار استفاده می شود. ورودی این یک تصویر و خروجی آن یک بردار ویژگی است که نشان دهنده ویژگی های بصری سطح بالا تصویر است.
- رمزگشا (Decoder): پس از استخراج ویژگی‌های تصویر این ویژگی ها به یک شبکه عصبی بازگشتی داده می شوند تا توصیف متنی گام به‌گام تولید شود .

شکل ۱ نشان دهنده معماری Encoder این مدل است، قسمت Encoder این مدل تصاویری با ابعاد 224*224 دریافت می کند و یک بردار ویژگی 256 بعدی تولید می کند. ویژگی های استخراج شده توسط Encoder به Decoder داده می شوند تا متن توصیفی تصویر را تولید کند.

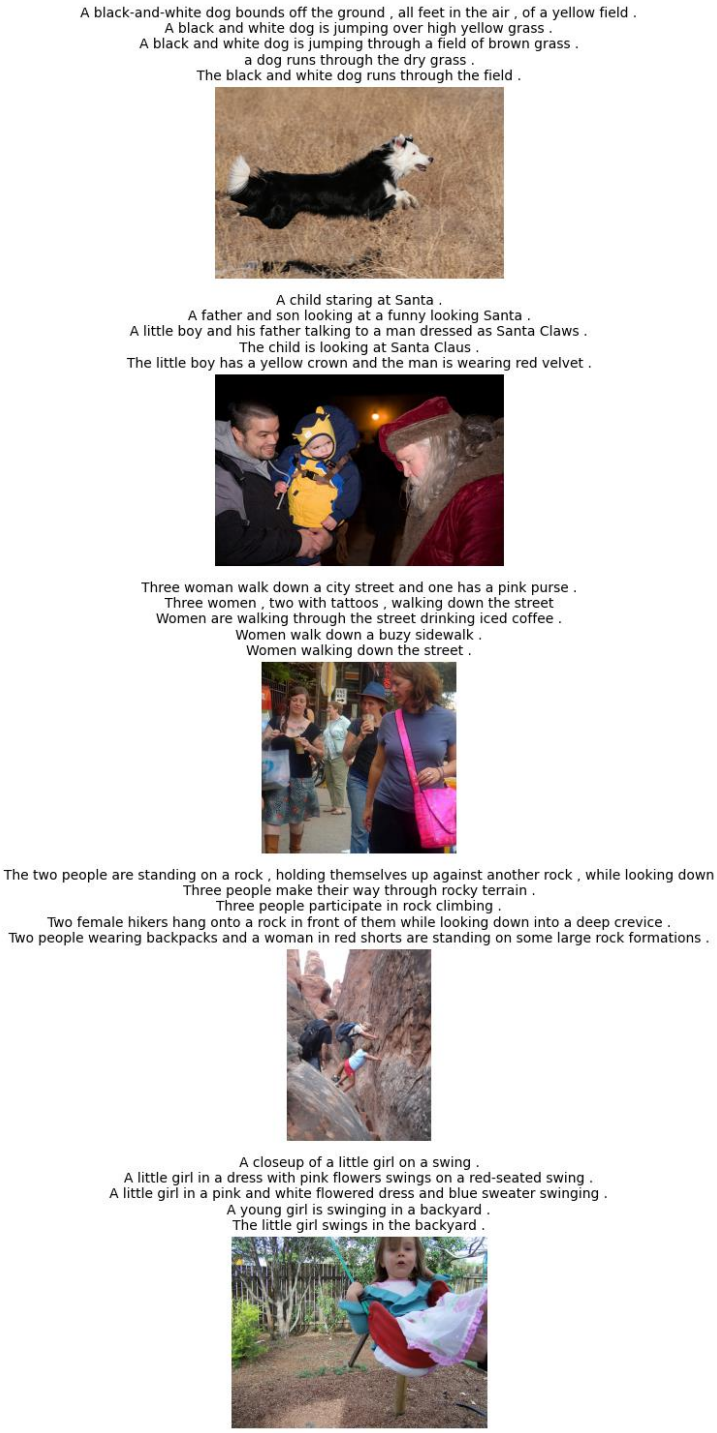


شکل ۱

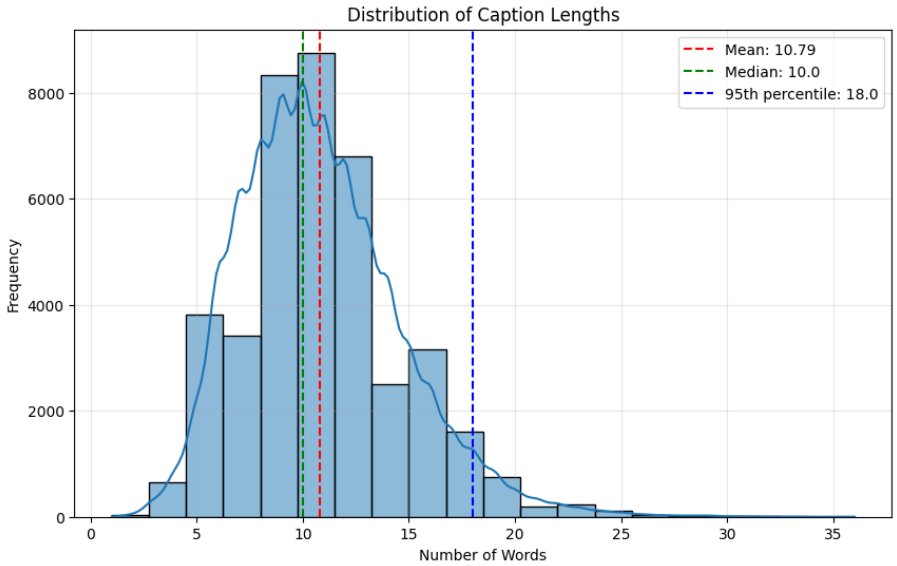
۳. Experimental Setup

برای آموزش و ارزیابی مدل از مجموعه‌داده Flickr8k استفاده شده است. این مجموعه داده شامل 8091 تصویر واقعی است که برای هر تصویر 5 Caption وجود دارد و تعداد کل Caption ها 40455 است. شکل ۲ نشان دهنده 5 تصویر و Caption های آن تصاویر است. شکل ۳ نشان دهنده نمودار توضیح طول Caption ها است، کوتاه ترین Caption فقط شامل 1 کلمه است و طولانی ترین Caption شامل 36 کلمه است. اندازه Vocabulary برابر 8827 است و تعداد کلمه هایی که فقط یک بار استفاده شده اند 3608 است. شکل ۴ نشان دهنده نمودار 30 کلمه رایج در Caption ها است، کلمه ای که بیشترین تکرار را در Caption ها دارد کلمه a است. شکل ۵ نشان دهنده نمودار Vocabulary Size و Vocabulary Coverage برحسب Frequency Threshold است. Frequency Threshold نشان دهنده تعداد باری که یک کلمه باید در Caption ها ظاهر شود تا در واژگان جدید باقی بماند است، این نمودار نشان می دهد که با افزایش Frequency Threshold اندازه Vocabulary کوچک می شود و Vocabulary Coverage کاهش می یابد چون کلمات نادر تر حذف می شوند. اگر دقت اهمیت داشته باشد باید Frequency Threshold کمتری انتخاب کنیم تا اندازه Vocabulary بزرگ تر شود و Vocabulary Coverage بالاتری داشته باشیم و اگر سرعت و کارایی اهمیت داشته

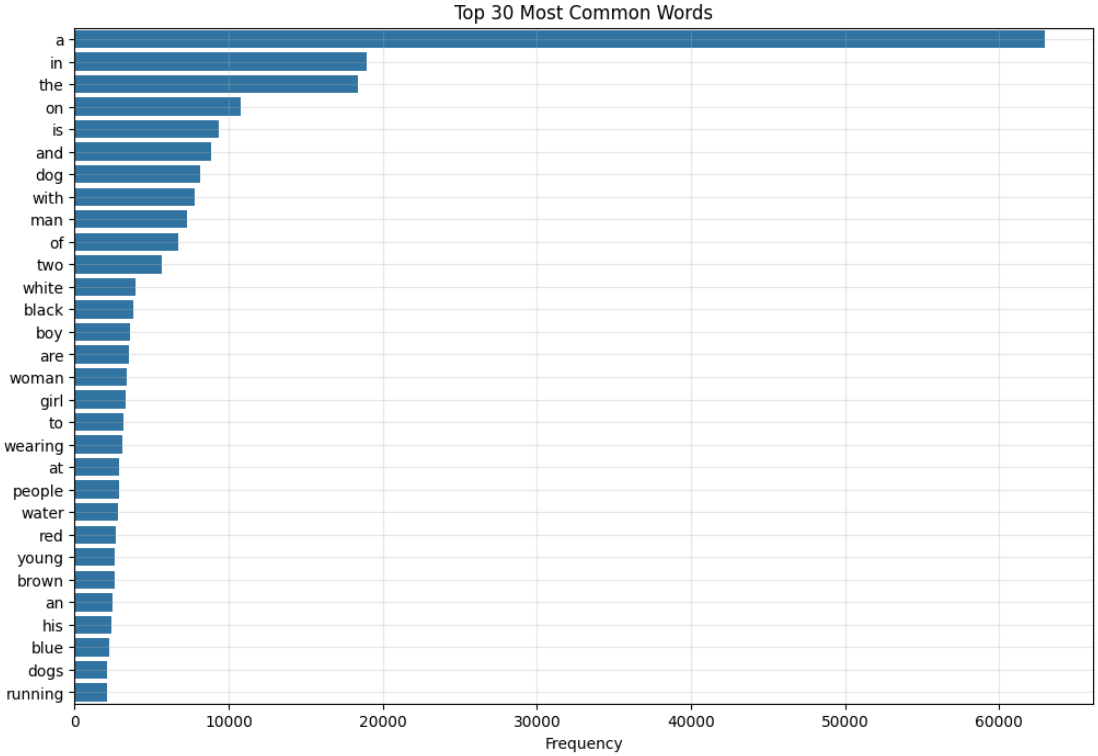
باشد باید Frequency Threshold بزرگ تری انتخاب کنیم تا اندازه Vocabulary کوچک تر شود. برای آموزش مدل دیتاست را به سه دیتاست Train، Validation و Test تقسیم می کنیم که Train Set شامل 6000 تصویر و Caption 30000 است، Validation Set شامل 1000 تصویر و Caption 5000 است و Test Set شامل 1000 تصویر و Caption 5000 است. شکل ۶ نشان دهنده نمودار ابعاد تصاویر است.



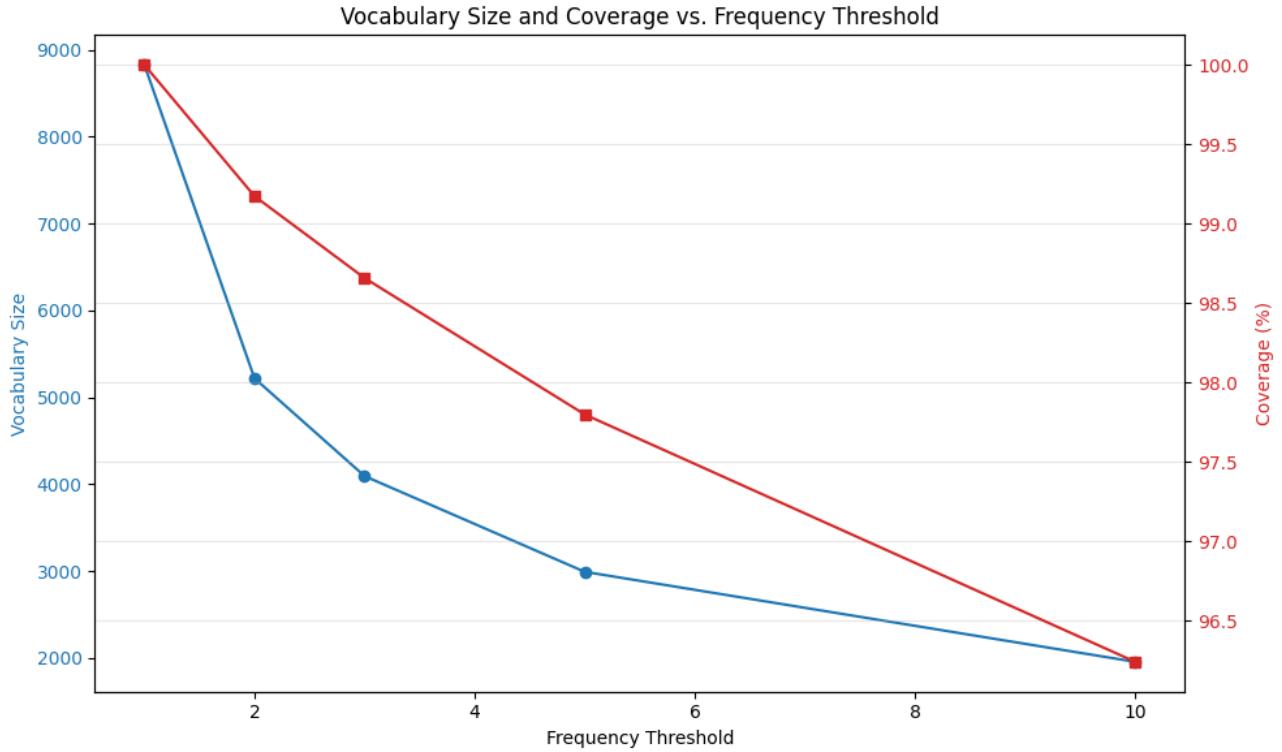
شکل ۲


Distribution of Caption Lengths
Mean: 10.79
Median: 10.0
95th percentile: 18.0
Frequency
Number of Words
0
2000
4000
6000
8000
0
5
10
15
20
25
30
35

شکل ۳


Top 30 Most Common Words
a
in
the
on
is
and
dog
with
man
of
two
white
black
boy
are
woman
girl
to
wearing
at
people
water
red
young
brown
an
his
blue
dogs
running
0
10000
20000
30000
40000
50000
60000
Frequency

شکل ۴


Vocabulary Size and Coverage vs. Frequency Threshold
Vocabulary Size
Coverage (%)
Frequency Threshold
9000
8000
7000
6000
5000
4000
3000
2000
100.0
99.5
99.0
98.5
98.0
97.5
97.0
96.5
2
4
6
8
10

شکل ۵

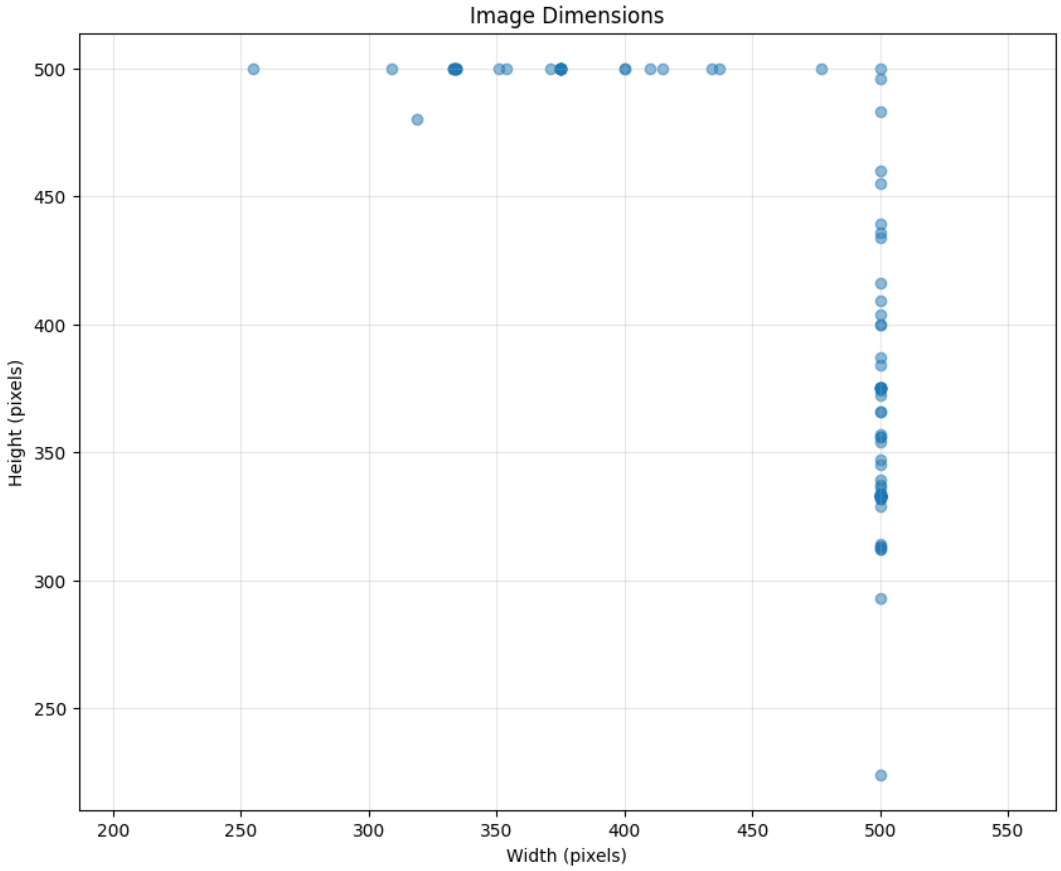
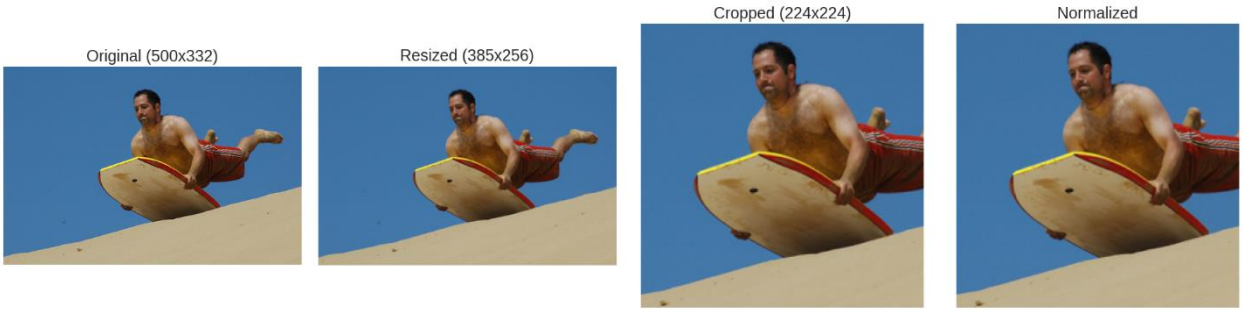


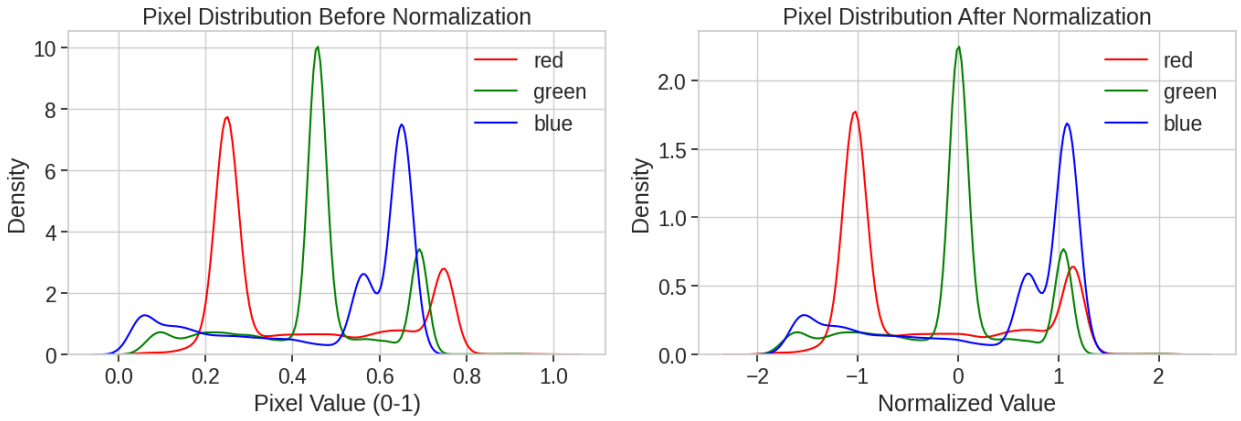
Image Dimensions
Height (pixels)
Width (pixels)
500
450
400
350
300
250
200
250
300
350
400
450
500
550

شکل ۶

شکل ۷ نشان دهنده نحوه اعمال مراحل پیش پردازش روی تصاویر است، هدف پیش پردازش تبدیل تصاویر ورودی به شکل مناسب برای شبکه های عصبی کانولوشنی است. در پیش پردازش تصویر ابتدا تصاویر Resize می شوند سپس Crop می شوند و در نهایت Normalize می شوند. شکل ۸ نشان دهنده نمودار توزیع پیکسل ها قبل و بعد از نرمال سازی است.

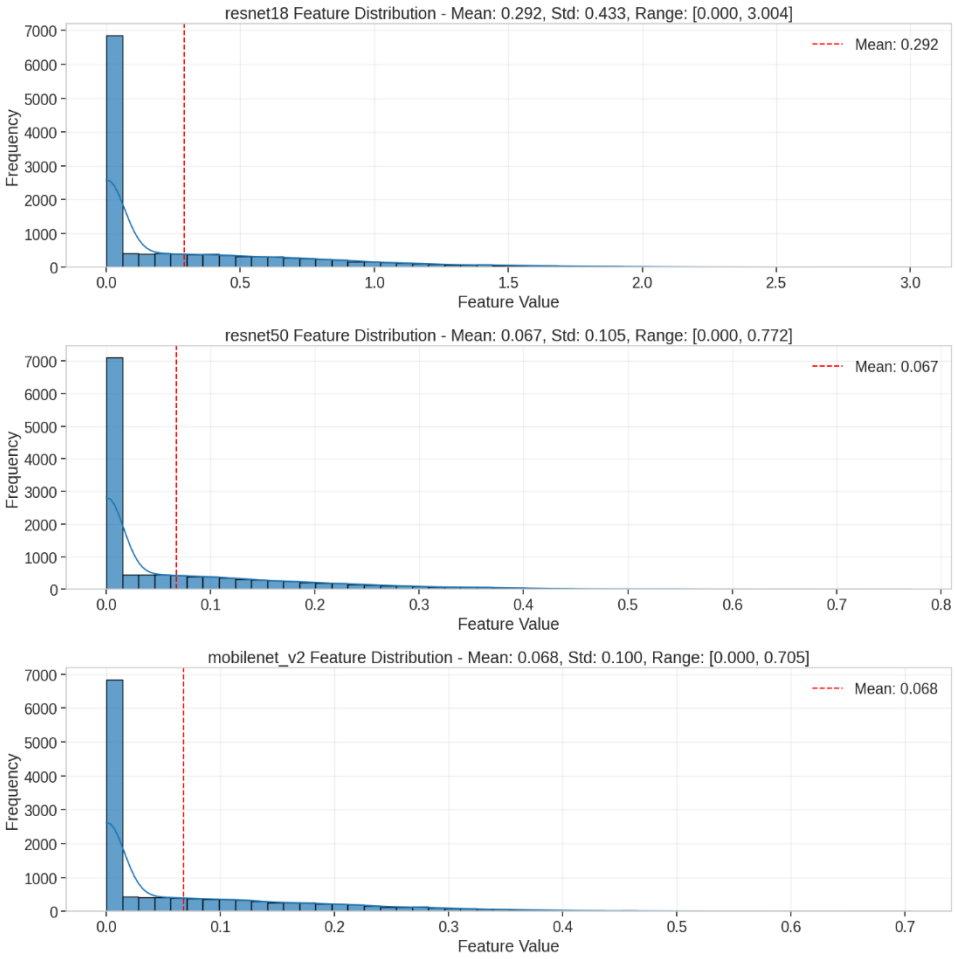


شکل ۷

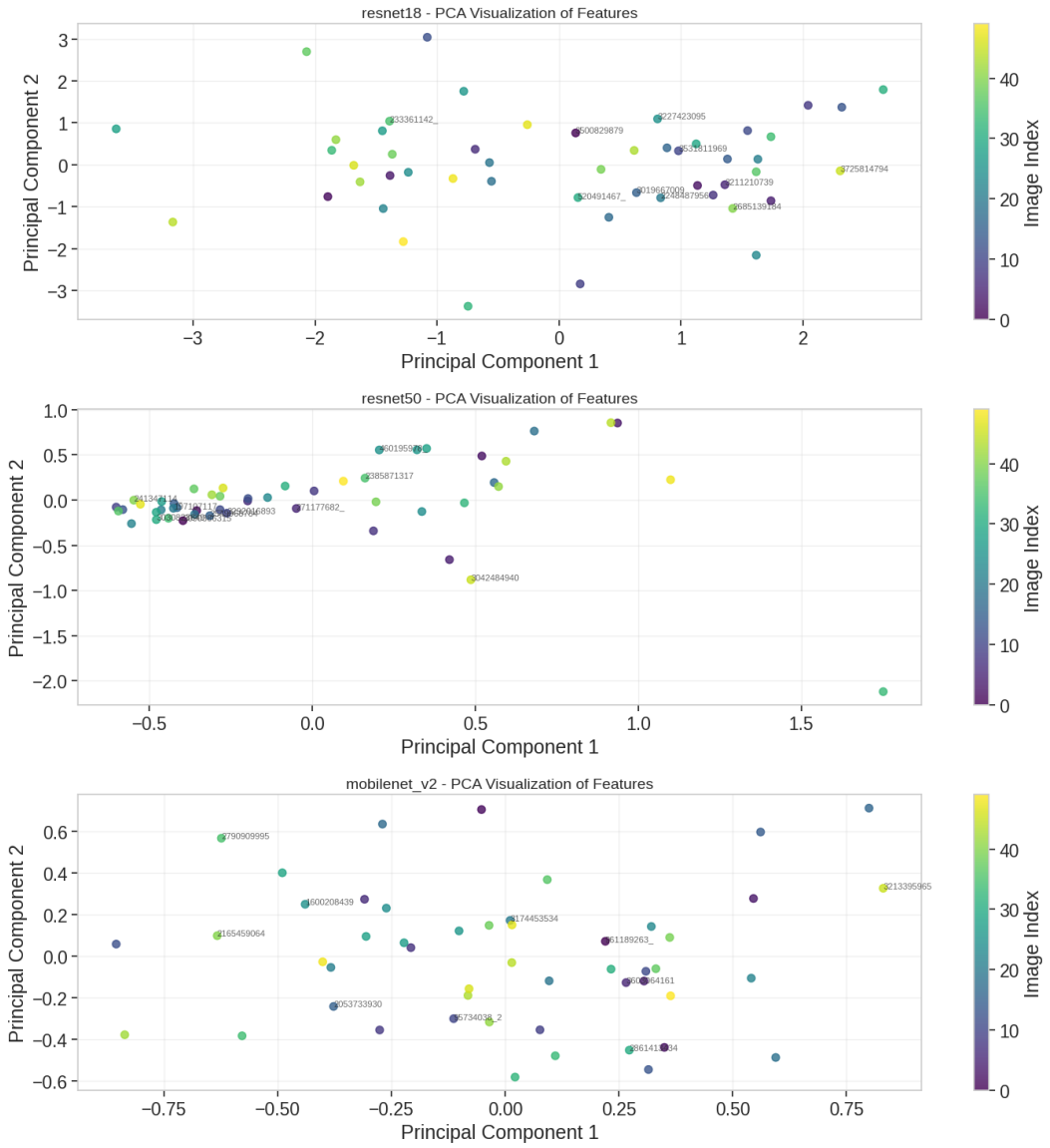


شکل ۸

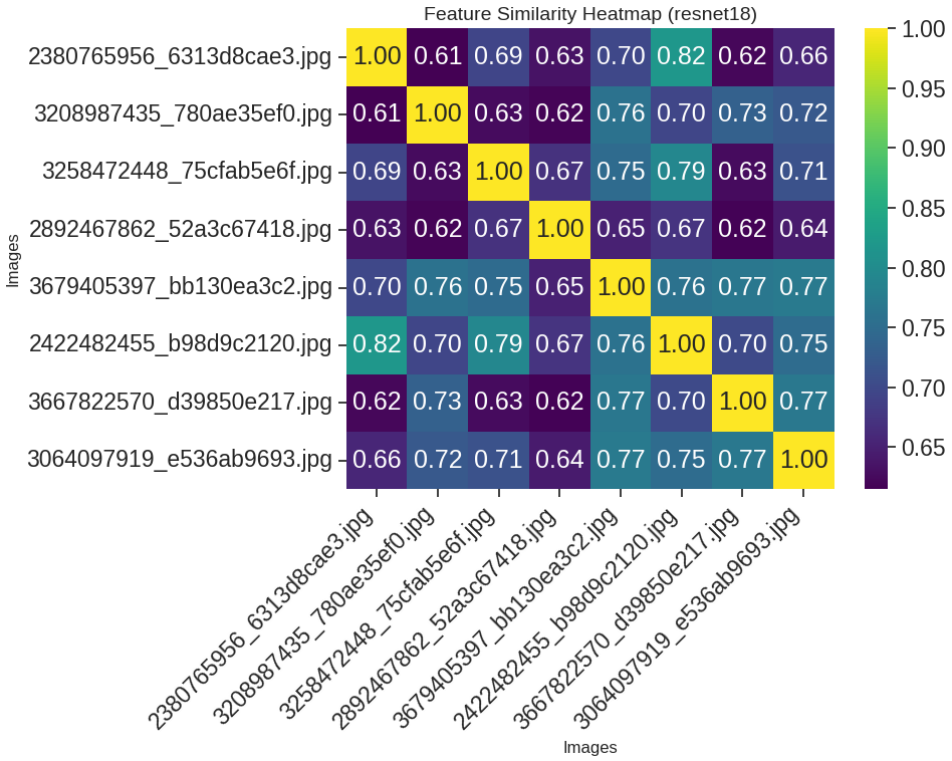
برای استخراج ویژگی ها از سه مدل از پیش آموزش دیده ResNet18، ResNet50 و MobilenetV2 استفاده می کنیم. شکل ۹ نشان دهنده نمودار توزیع ویژگی های برای این سه مدل است، مقادیر ویژگی های در ResNet18 محدوده گسترده تری قرار دارند و انحراف معیار بالایی دارند اما در ResNet50 و MobilenetV2 مقادیر ویژگی های در محدوده کوچک تری قرار دارند و انحراف معیار کمتری دارند. شکل ۱۰ نشان دهنده نمودار PCA برای تحلیل ویژگی های استخراج شده توسط این مدل ها را نشان می دهد، هدف اصلی این نمودار ها مقایسه نحوه استخراج ویژگی ها توسط این مدل ها و نمایش توزیع ویژگی ها در فضای Principal Component است. شکل ۱۱ نشان دهنده نمودار Heatmap شباهت ویژگی های استخراج از تصاویر با استفاده از مدل ها است، هدف اصلی این نمودار مقایسه شباهت بین تصاویر مختلف بر اساس ویژگی های استخراج شده توسط مدل است. جدول ۱ نشان دهنده یک مقایسه کامل بین این سه مدل است.

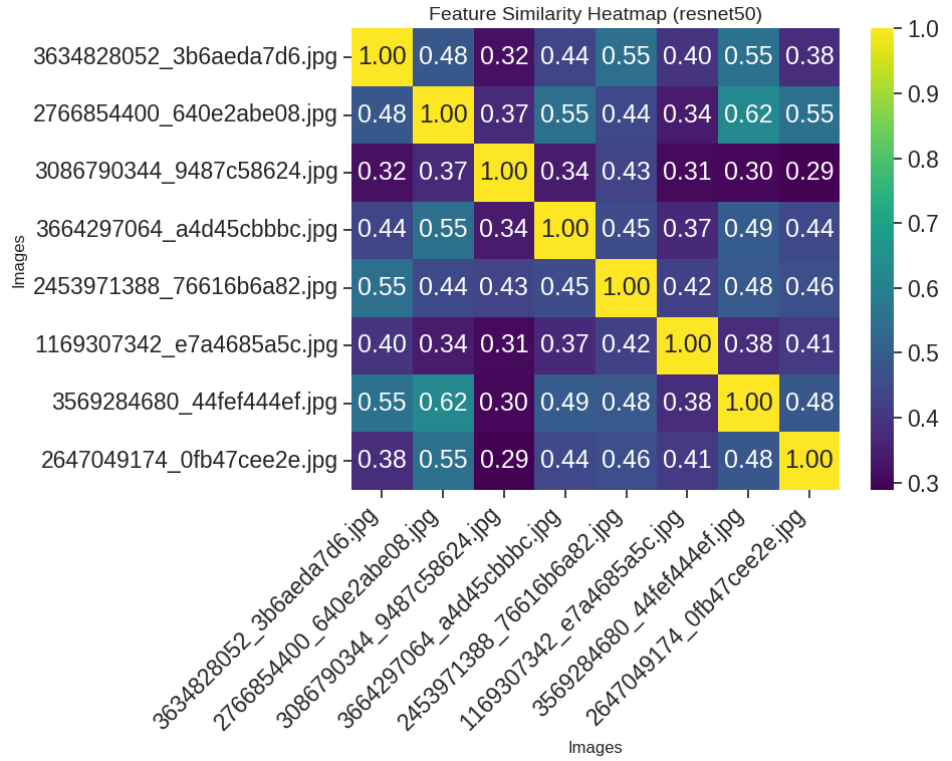

resnet18 Feature Distribution - Mean: 0.292, Std: 0.433, Range: [0.000, 3.004]
Mean: 0.292
Frequency
Feature Value
resnet50 Feature Distribution - Mean: 0.067, Std: 0.105, Range: [0.000, 0.772]
Mean: 0.067
Frequency
Feature Value
mobilenet_v2 Feature Distribution - Mean: 0.068, Std: 0.100, Range: [0.000, 0.705]
Mean: 0.068
Frequency
Feature Value

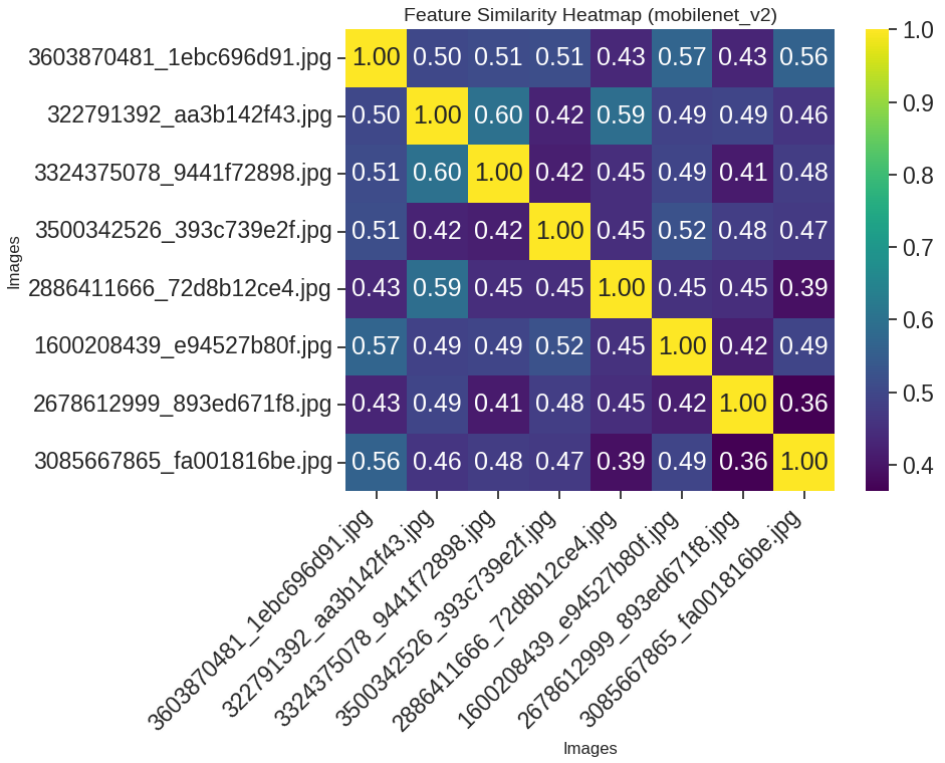
شکل ۹


resnet18 - PCA Visualization of Features
Principal Component 2
Principal Component 1
Image Index
resnet50 - PCA Visualization of Features
Principal Component 2
Principal Component 1
Image Index
mobilenet_v2 - PCA Visualization of Features
Principal Component 2
Principal Component 1
Image Index

شکل ۱۰


Feature Similarity Heatmap (resnet18)
2380765956_6313d8cae3.jpg 1.00 0.61 0.69 0.63 0.70 0.82 0.62 0.66
3208987435_780ae35ef0.jpg 0.61 1.00 0.63 0.62 0.76 0.70 0.73 0.72
3258472448_75cfab5e6f.jpg 0.69 0.63 1.00 0.67 0.75 0.79 0.63 0.71
2892467862_52a3c67418.jpg 0.63 0.62 0.67 1.00 0.65 0.67 0.62 0.64
3679405397_bb130ea3c2.jpg 0.70 0.76 0.75 0.65 1.00 0.76 0.77 0.77
2422482455_b98d9c2120.jpg 0.82 0.70 0.79 0.67 0.76 1.00 0.70 0.75
3667822570_d39850e217.jpg 0.62 0.73 0.63 0.62 0.77 0.70 1.00 0.77
3064097919_e536ab9693.jpg 0.66 0.72 0.71 0.64 0.77 0.75 0.77 1.00
Images
2380765956_6313d8cae3.jpg
3208987435_780ae35ef0.jpg
3258472448_75cfab5e6f.jpg
2892467862_52a3c67418.jpg
3679405397_bb130ea3c2.jpg
2422482455_b98d9c2120.jpg
3667822570_d39850e217.jpg
3064097919_e536ab9693.jpg
Images
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65


Feature Similarity Heatmap (resnet50)
3634828052_3b6aeda7d6.jpg 1.00 0.48 0.32 0.44 0.55 0.40 0.55 0.38
2766854400_640e2abe08.jpg 0.48 1.00 0.37 0.55 0.44 0.34 0.62 0.55
3086790344_9487c58624.jpg 0.32 0.37 1.00 0.34 0.43 0.31 0.30 0.29
3664297064_a4d45cbbbc.jpg 0.44 0.55 0.34 1.00 0.45 0.37 0.49 0.44
2453971388_76616b6a82.jpg 0.55 0.44 0.43 0.45 1.00 0.42 0.48 0.46
1169307342_e7a4685a5c.jpg 0.40 0.34 0.31 0.37 0.42 1.00 0.38 0.41
3569284680_44fef444ef.jpg 0.55 0.62 0.30 0.49 0.48 0.38 1.00 0.48
2647049174_0fb47cee2e.jpg 0.38 0.55 0.29 0.44 0.46 0.41 0.48 1.00
Images
3634828052_3b6aeda7d6.jpg
2766854400_640e2abe08.jpg
3086790344_9487c58624.jpg
3664297064_a4d45cbbbc.jpg
2453971388_76616b6a82.jpg
1169307342_e7a4685a5c.jpg
3569284680_44fef444ef.jpg
2647049174_0fb47cee2e.jpg
Images
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3



شکل ۱۱

Model	Parameters(M)	Feature Dimension	Feature Mean	Feature Std
ResNet18	11.7	512	0.2917	0.4335
ResNet50	25.6	2048	0.0669	0.1055
Mobilenet v2	3.5	1280	0.0676	0.0999s

جدول۱

فرآیند آموزش به این صورت است که تصویر ورود

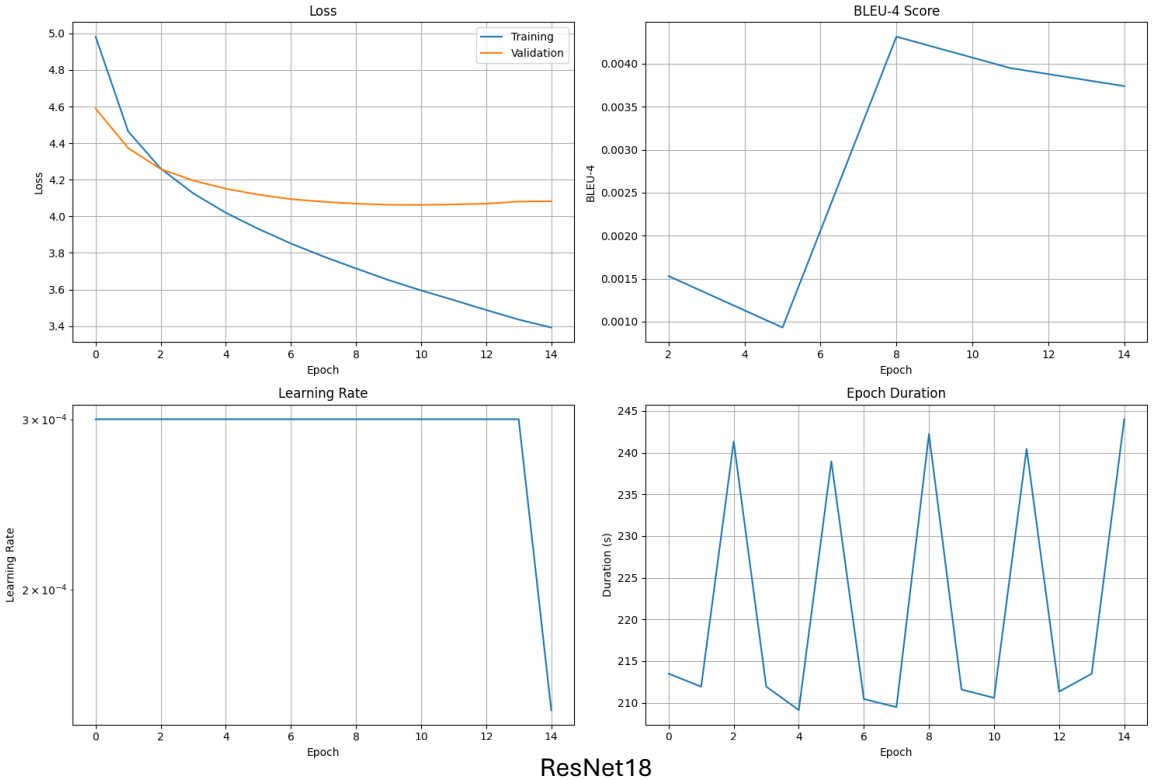
۴. نتایج کمی و کیفی

برای آموزش مدل از Batch Size=32، Dropout Rate=0.5، Decoder Type=LSTM ، Learning Rate=3e-4 و Number of Epoch=60 استفاده کرده ایم. اولین مدلی که به عنوان Encoder استفاده کردیم ResNet18 است، در مدل ترکیبی که با استفاده از این مدل ساخته می شود تعداد کل پارامتر ها قابل آموزش 5425322 است، تعداد پارامتر های Encoder برابر 131840 است و تعداد پارامتر های Decoder برابر 5161898 است. آموزش مدل پس از 16 Epoch متوقف شد، خطایی آموزش نهایی 3.3019 است، خطای Validation نهایی 4.0808 است، بهترین خطای Validation ای که به دست آمد 4.0627 است و بهترین BLUE-4 Socre ای که به دست آمد 0.0043 است. شکل ۱۲ نشان دهنده نمودار های Loss، BLUE-4 Score، Learning Rate و Epoch Duration این مدل است.بهترین Epoch آموزش 11 Epoch بود که در آن خطای Validation برابر

4.0627 بود و بهترین BLUE Epoch آموزش 9 Epoch بود که در آن BLUE Score برابر 0.0043 بود. جدول ۲ نشان دهنده یک مقایسه کامل از نتایج مدل ترکیبی با استفاده از Encoder های متفاوت است. شکل ۱۳ نشان دهنده نمودار های Loss، BLUE-4 Score، Learning Rate و Epoch Duration مدل با استفاده از ResNet50 Encoder و MobileNet Encoder است.

Endoer	Total Parameters	Encoder Parameters	Decoder Parameters	Final Train Loss	Final Validation Loss	Best Validation Loss	Best BLUE-4 Score	Best Epoch	Best BLUE Score Epoch
ResNet18	5425322	131840	5161898	3.3019	4.0808	4.0627	0.0043	11	9
ResNet50	5818538	525056	5161898	3.2292	4.0233	3.9971	0.0109	11	15
MobileNet V2	5621930	328448	5161898	3.2774	4.1010	4.0658	0.0030	11	10

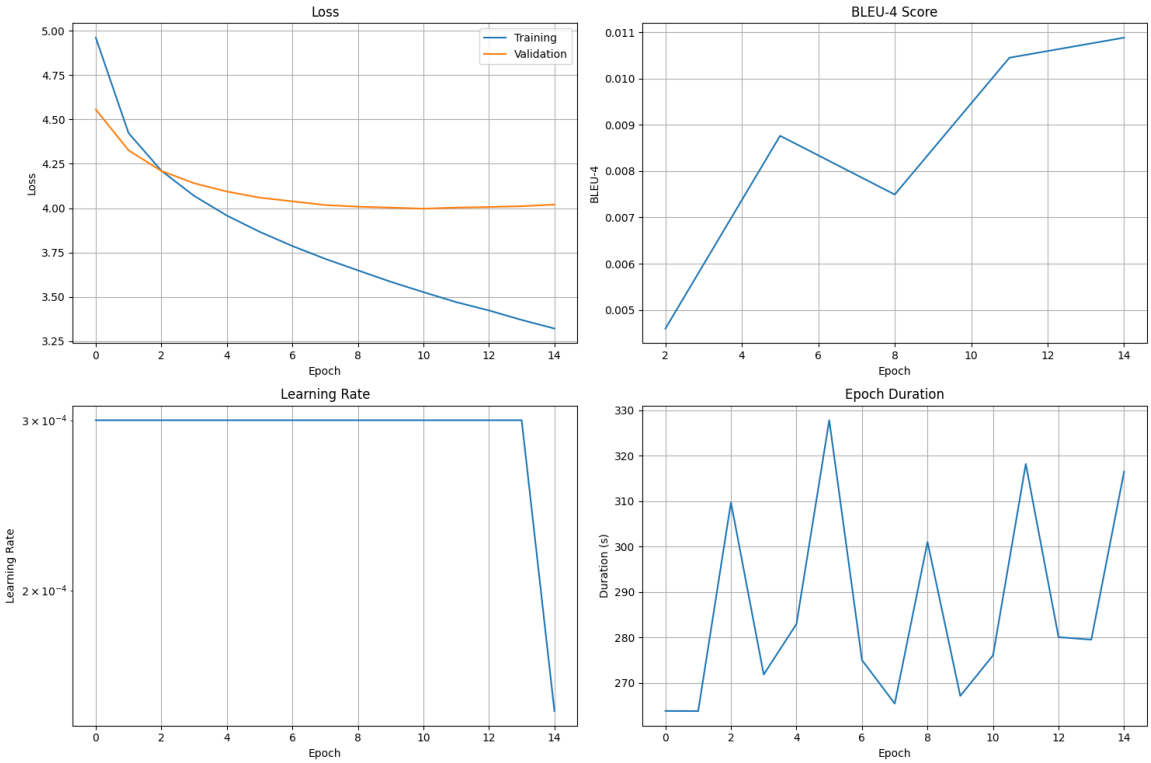
جدول ۲



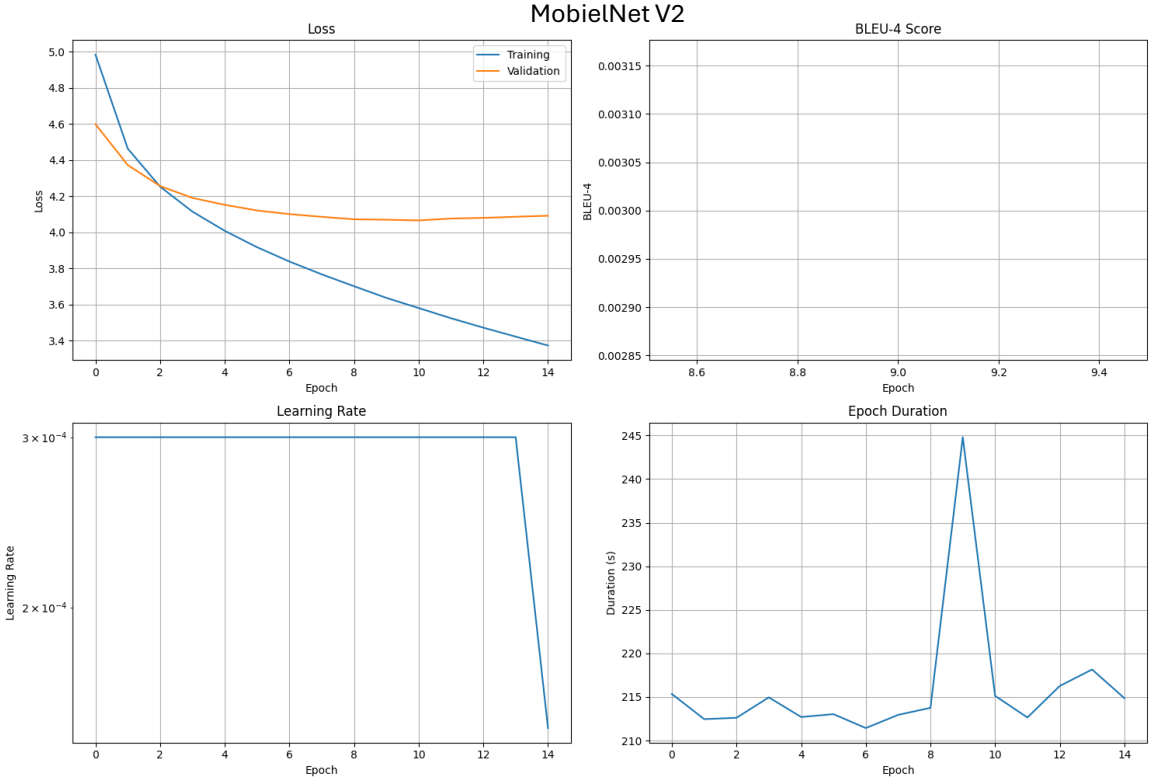
ResNet18

شکل ۱۲

ResNet50



MobielNet V2

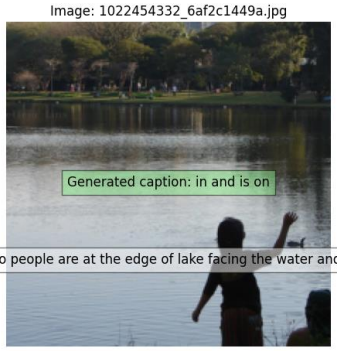


شکل ۱۳

با بررسی نتایج می توان نتیجه گرفت که ResNet50 بهترین Encoder برای استفاده در مدل ترکیبی است. شکل ۱۴ نشان دهنده تصویر، Caption های واقعی آن تصویر و Caption های تولید شده توسط مدل است.

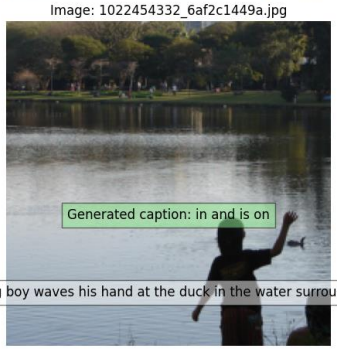
ResNet18

Image: 1022454332_6af2c1449a.jpg



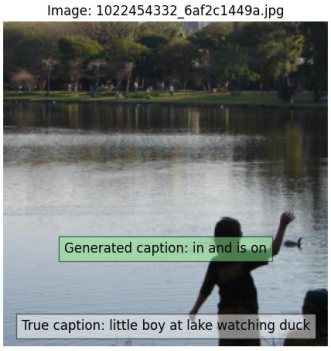
True caption: two people are at the edge of lake facing the water and the city skyline

Image: 1022454332_6af2c1449a.jpg



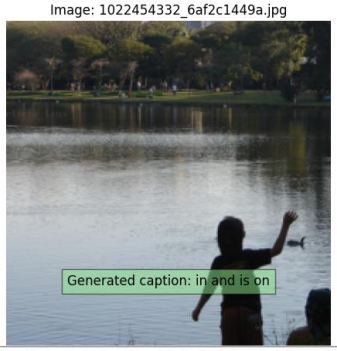
True caption: young boy waves his hand at the duck in the water surrounded by green park

Image: 1022454332_6af2c1449a.jpg



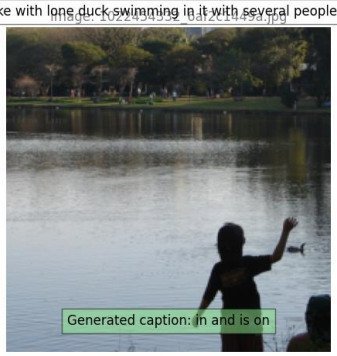
True caption: little boy at lake watching duck

Image: 1022454332_6af2c1449a.jpg



True caption: large lake with lone duck swimming in it with several people around the edge of it

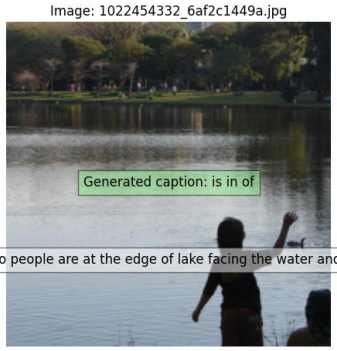
Image: 1022454332_6af2c1449a.jpg



True caption: child and woman are at waters edge in big city

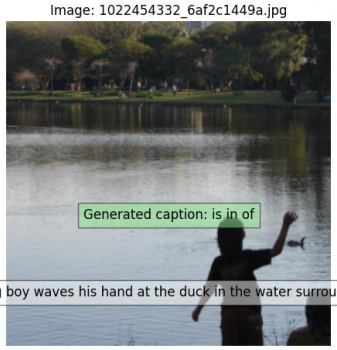
ResNet50

Image: 1022454332_6af2c1449a.jpg



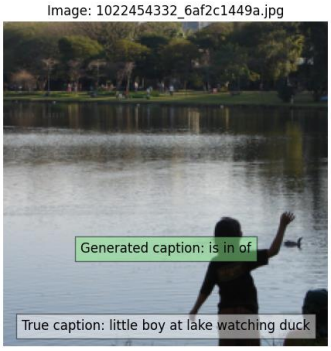
True caption: two people are at the edge of lake facing the water and the city skyline

Image: 1022454332_6af2c1449a.jpg



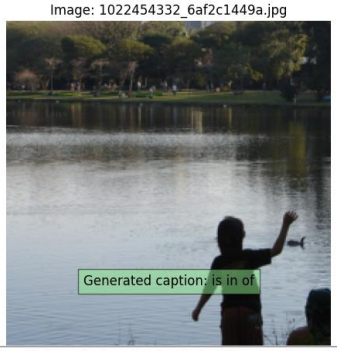
True caption: young boy waves his hand at the duck in the water surrounded by green park

Image: 1022454332_6af2c1449a.jpg



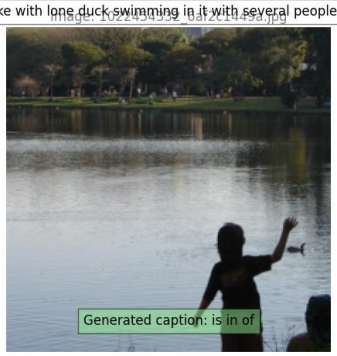
True caption: little boy at lake watching duck

Image: 1022454332_6af2c1449a.jpg



True caption: large lake with lone duck swimming in it with several people around the edge of it

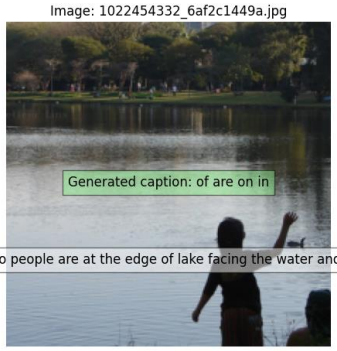
Image: 1022454332_6af2c1449a.jpg



True caption: child and woman are at waters edge in big city

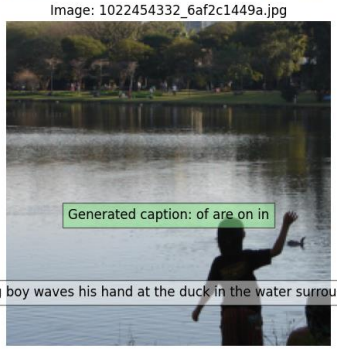
MobielNet V2

Image: 1022454332_6af2c1449a.jpg



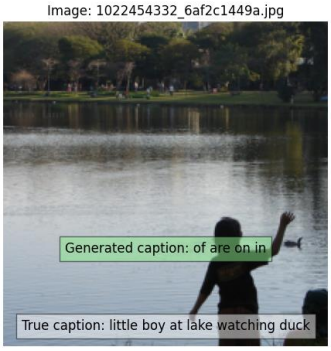
True caption: two people are at the edge of lake facing the water and the city skyline

Image: 1022454332_6af2c1449a.jpg



True caption: young boy waves his hand at the duck in the water surrounded by green park

Image: 1022454332_6af2c1449a.jpg

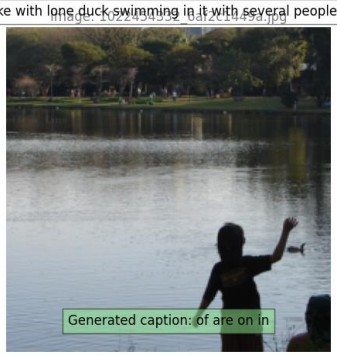


True caption: little boy at lake watching duck

Image: 1022454332_6af2c1449a.jpg

True caption: large lake with lone duck swimming in it with several people around the edge of it

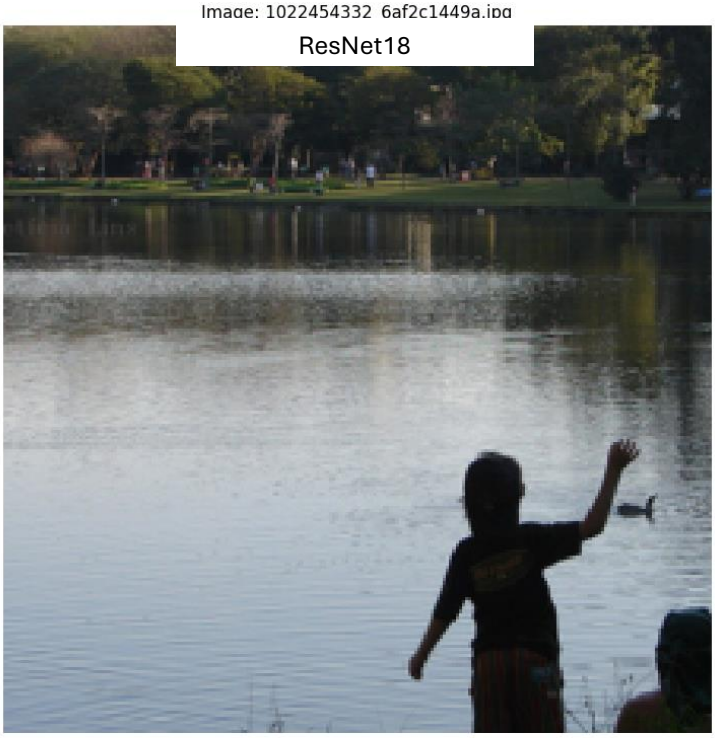
Image: 1022454332_6af2c1449a.jpg



True caption: child and woman are at waters edge in big city

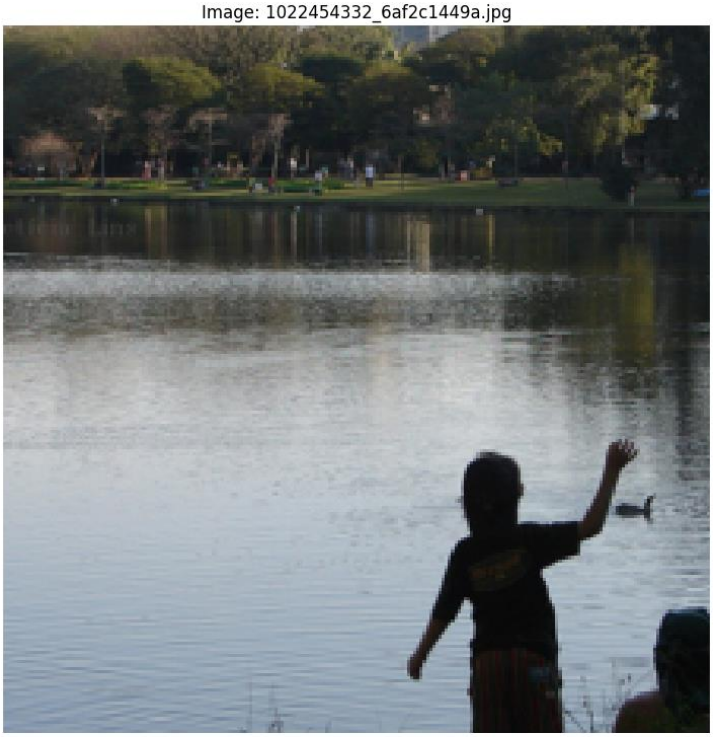
شکل ۱۴

شکل ۱۵ نشان دهنده مقایسه بین Greedy Decoding و Beam Search در مدل ترکیبی با سه Encoder مختلف است.



ResNet50

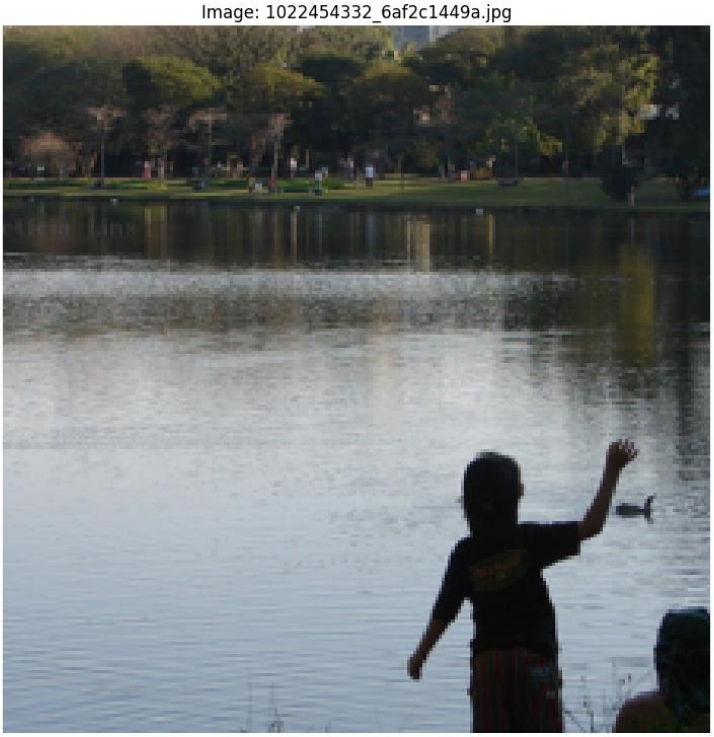
Image: 1022454332_6af2c1449a.jpg



True caption: little boy at lake watching duck
Greedy search: is in of
Beam search (k=3): is on
Beam search (k=5): is on

MobielNet V2

Image: 1022454332_6af2c1449a.jpg



True caption: large lake with lone duck swimming in it with several people around the edge of it
Greedy search: of are on in
Beam search (k=3): is in of and is on
Beam search (k=5): is in of

شکل ۱۵

۵. خلاصه مقاله Show and Tell: A Neural Image Caption Generator

- چکیده: این مقاله یک مدل End-to-End برای تولید خودکار Caption تصاویر معرفی می‌کند. این مدل از دو بخش اصلی تشکیل شده است، یک شبکه عصبی کانولوشنی برای استخراج ویژگی های تصویر و یک شبکه عصبی بازگشتی برای تولید Caption . این مدل با عنوان NIC(Neural Image Caption) قادر است با مشاهده یک تصویر بدون متن یک جمله به زبان طبیعی تولید کند که موضوع یا عملکرد اصلی تصویر را توضیح دهد.
- هدف مقاله: هدف این مقاله توسعه یک سیستم است که بتواند به صورت خودکار و بدون دخالت انسان یک Caption قابل فهم از تصاویر تولید کند. این کار شامل درک تصویر با استفاده از شبکه عصبی کانولوشنی و سپس تبدیل آن به یک توالی مناسب از کلمات با استفاده از شبکه عصبی بازگشتی است.
- معماری مدل: اولین بخش معماری این مدل Encoder است، در این بخش از یک شبکه عصبی کانولوشنی از پیش آموزش داده شده مانند Inception-V3 برای استخراج ویژگی های تصویر استفاده شده است و خروجی این شبکه به عنوان ورودی

اولیه برای شبکه عصبی بازگشتی استفاده می شود. دومین بخش معماری مدل Decoder است، در این بخش از یک شبکه عصبی بازگشتی مثل LSTM برای تولید کلمات استفاده شده است و این شبکه عصبی با استفاده از تابع هزینه -log likelihood آموزش داده شده است. ورودی مدل یک تصویر است و خروجی آن یک دنباله از کلمات که کپشن را تشکیل می دهند.

- آموزش مدل: مدل با بیشینه سازی احتمال جمله صحیح S با توجه به تصویر I آموزش می بیند، برای آموزش از روش Maximum Likelihood Estimation و الگوریتم SGD استفاده می شود و هدف تولید یک توالی از کلمات است که احتمال شرطی $P(S|I)$ را بیشینه کند. برای تولید جملات در مرحله تست از روش هایی مانند Beam Search و Greedy Decoding استفاده شده است. این مدل از Teacher Forcing در زمان آموزش استفاده می کند و در زمان Inferencing از Sampling یا Beam Search برای تولید کپشن استفاده می کند.
- معیار های ارزیابی: BLUE Score یکی از معیار های ارزیابی در تولید متن است. در مقایسه با مدل های قبلی NIC نتایج بسیار بهتری داشته است. BLUE-1 در Pascal VOC از 25 به 59 افزایش یافته است، BLUE-1 در Flickr30k از 56 به 66 افزایش یافته است، BLUE-1 در SBU از 19 به 28 افزایش یافته است و BLUE-4=27.7 در MSCOCO به دست آمده است. معیار های METEOR و CIDEr همبستگی بالاتری با ارزیابی انسانی دارند، نتایج NIC با این معیار ها هم بهتر از مدل های قبلی است و در بسیاری موارد توصیف های تولید شده قابل تشخیص از توصیف های انسانی نبوده اند.
- دیتاست ها: در این مقاله از دیتاست های Pascal VOC ، Flickr8k ، Flickr30k ، SBU و MSCOCO استفاده شده است.
- ایده کلی: ایده اصلی این مقاله از Statistical Machine Translation الهام گرفته شده است، در SMT مدل یک جمله را به جمله دیگری ترجمه می کند اما در اینجا مدل یک تصویر را به یک جمله تبدیل می کند.
- نتایج: NIC یک رویکرد End-to-End است که نیازی به مراحل جداگانه مثل Detection + Template-Based-Captioning ندارد. در تمامی دیتاست ها عملکرد State-of-the-Art را دارد. کپشن های تولید شده اغلب طبیعی، قابل خواندن و دقیق هستند. مدل به خوبی از Beam Search برای تولید کپشن های بهتر استفاده می کند. نتایج نشان می دهند که افزایش اندازه دیتاست به بهبود چشم‌گیر عملکرد کمک می کند. در دیتاست SBU که لیبل ها خودکار و غیر دقیق هستند عملکرد مدل ضعیف تر است. در ارزیابی انسانی مدل هنوز دور از نتیجه انسان ها است.

۶. خلاصه مقاله Show, Attend and Tell

- هدف مقاله: این مقاله یک مدل جدید در حوزه تولید خودکار Caption تصویر با استفاده از Mechanism Attention معرفی می کند. این مدل قادر است بدون نیاز به تشخیص شیء جداگانه یک توالی از کلمات را تولید کند که Caption

معنایی تصویر را فراهم کند. هدف اصلی پژوهش افزایش دقت Caption ها با استفاده از Attention ، ایجاد قابلیت Interpretability در مدل CNN-RNN و مقایسه Soft Attention و Hard Attention در عملکرد و کارایی است.

- معماری مدل: قسمت Encoder مدل یک شبکه عصبی کانولوشنی است که برای استخراج Feature Map های فضایی استفاده شده است. قسمت Encdoer یک LSTM برای تولید کلمات بر اساس ویژگی های تصویر و یک ماژول Attention است که در هر مرحله از RNN وزن هایی را روی Feature Map محاسبه می کند و فقط به بخش های مهم تصویر توجه می کند.
- انواع توجه: در Soft Attention مدل با وزن دهی به تمام Feature ها توجه می کند، Soft Attention قابل آموزش با Backpropagation است و دقیق تر و قابل تفسیر تر است. در Hard Attention مدل فقط به یک قسمت تصویر توجه می کند، آموزش تصادفی و با Policy Gradient است و سریع تر ولی غیر قابل پیش بینی تر است.
- روش تحقیق: از دیتاست های معروف Flickr8k, Flickr30k, MSCOCO برای آموزش استفاده شده است. معیار های ارزیابی BLUE-1, BLUE-4, ROUGE-L, CIDEr-D, METEOR هستند. آموزش با استفاده از Teacher Forcing, Negative Log-Likelihood Loss, Adam Optimizer صورت می گیرد.
- نتایج: جدول زیر نشان دهنده نتایج به دست آمده است. نتایج نشان دهنده بهبود چشم گیر مدل با استفاده از Attention نسبت به مدل های بدون Attention است.

		BLEU				
Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Flickr8k	Google NIC(Vinyals et al., 2014)†Σ	63	41	27	—	—
	Log Bilinear (Kiros et al., 2014a)◦	65.6	42.4	27.7	17.7	17.31
	Soft-Attention	**67**	44.8	29.9	19.5	18.93
	Hard-Attention	**67**	**45.7**	**31.4**	**21.3**	**20.30**
Flickr30k	Google NIC†◦Σ	66.3	42.3	27.7	18.3	—
	Log Bilinear	60.0	38	25.4	17.1	16.88
	Soft-Attention	66.7	43.4	28.8	19.1	**18.49**
	Hard-Attention	**66.9**	**43.9**	**29.6**	**19.9**	18.46
COCO	CMU/MS Research (Chen & Zitnick, 2014)a	—	—	—	—	20.41
	MS Research (Fang et al., 2014)†a	—	—	—	—	20.71
	BRNN (Karpathy & Li, 2014)◦	64.2	45.1	30.4	20.3	—
	Google NIC†◦Σ	66.6	46.1	32.9	24.6	—
	Log Bilinear◦	70.8	48.9	34.4	24.3	20.03
	Soft-Attention	70.7	49.2	34.4	24.3	**23.90**
	Hard-Attention	**71.8**	**50.4**	**35.7**	**25.0**	23.04

- نحوه کارکرد توجه: Encoder یک Feature Map فضایی از تصویر می سازد. Decoder در هر مرحله یک Context Vector را با استفاده Attention Weights از Feature Map ها می سازد. Attention Weights با استفاده از یک MLP محاسبه می شوند. مدل به صورت End-to-End آموزش داده می شود.
- مزایای استفاده از توجه: افزایش دقت Caption، قابلیت Interpretability، عدم نیاز به object detection، کاهش overfitting و نمایش visual-semantic alignment مزیت های توجه هستند.
- تحلیل Caption ها: مدل می تواند Caption های دقیق تولید کند مثل A woman is holding a remote control و می تواند Caption های غلط تولید کند ولی با استفاده از Attention Map می توان دلیل خطا را متوجه شد.

- نتیجه گیری: مقاله Neural Image Caption Generation with Visual Attention نشان می دهدکه با استفاده از Attention می توان مدلی ساخت که قادر به تولید Caption های دقیق تری باشد. قابلیت تفسیر مدل با استفاده از Attention Map فراهم شده است