Deep Learning Project:

Image Captioning on the COCO Dataset

University of Milano-Bicocca

Our team

Alex Calabrese

Data Science Student

Antonio Sabbatella

Data Science Student

Table of contents

COCO (Common Objects in Context)

- Large-scale object detection, segmentation, and **captioning** dataset 2014.
- Contains over **330,000 images** with annotated objects.
- Widely used in computer vision research and development.
- **Size: 17 GB.**

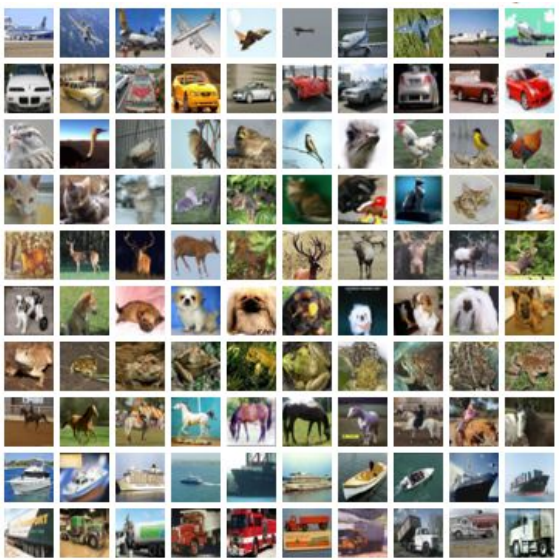


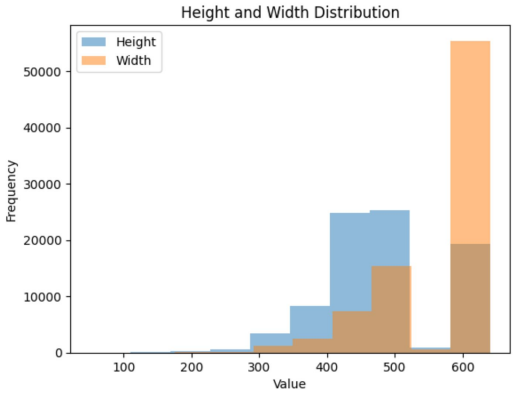
Image Captioning

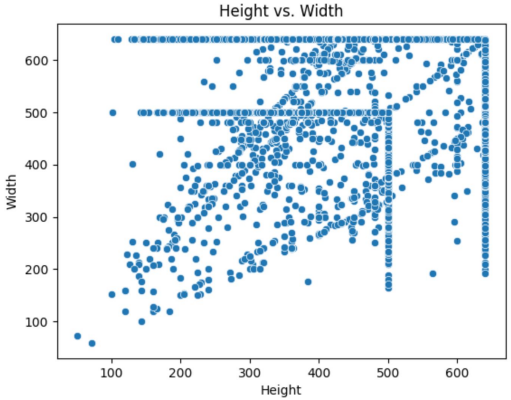
Image captioning is the process of generating a textual description for given images.

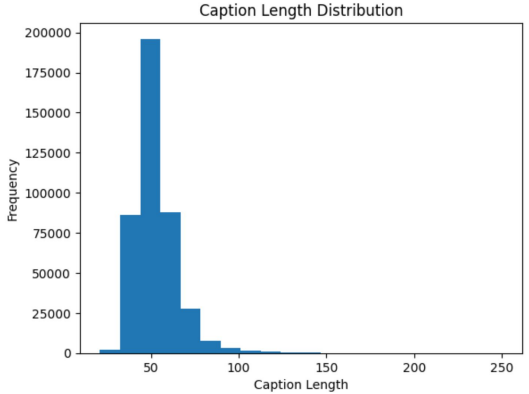
Input:

Output: *"A cat lying on a couch with a remote lying next to it."*

Exploration

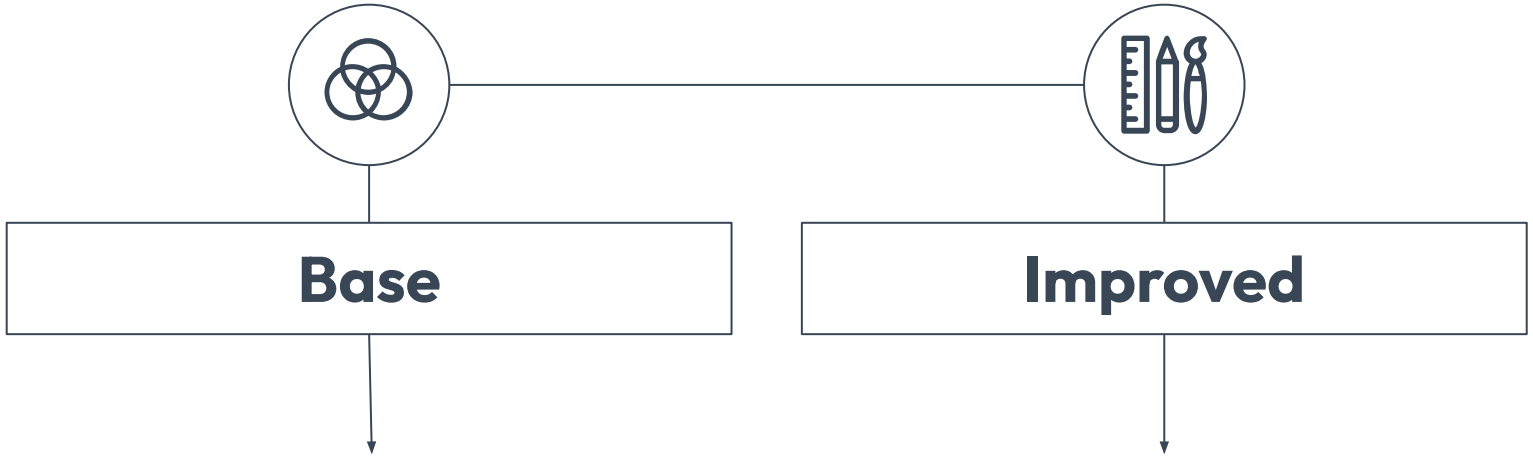






Heights are clustered **~500 pixels** while **widths** near **~600 pixels**.
Captions contain between **25 to 75 words** with a peak around 50 words.

Data Augmentation



Used in all the baselines

This are the most used techniques and are displayed in various img. Captioning. projects

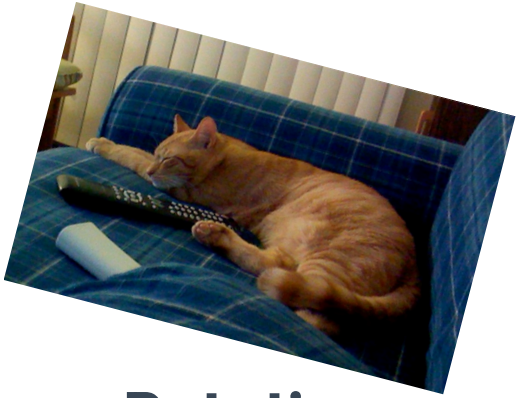
Best Model

Expanded “base” data augmentation for improving the best model performance

Base data augmentation

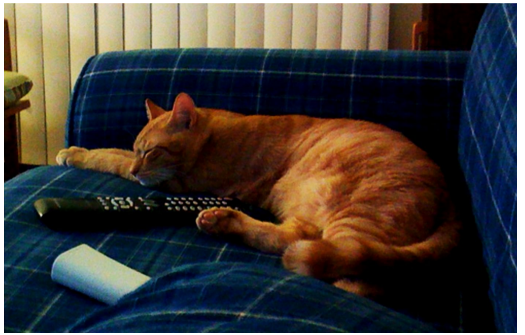
Flip

Randomly flips image horizontally
Factor: 50% chance



Rotation

Rotates image by r. angle
Factor: ±0.2 radians (±11.50)



Contrast

Alters contrast levels by a factor ±0.3

Experiments with Custom Models

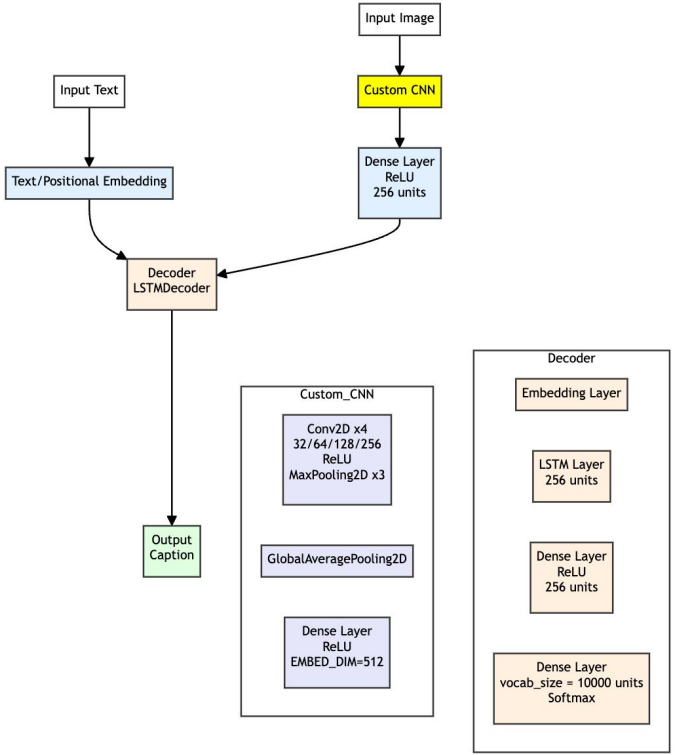
LSTM

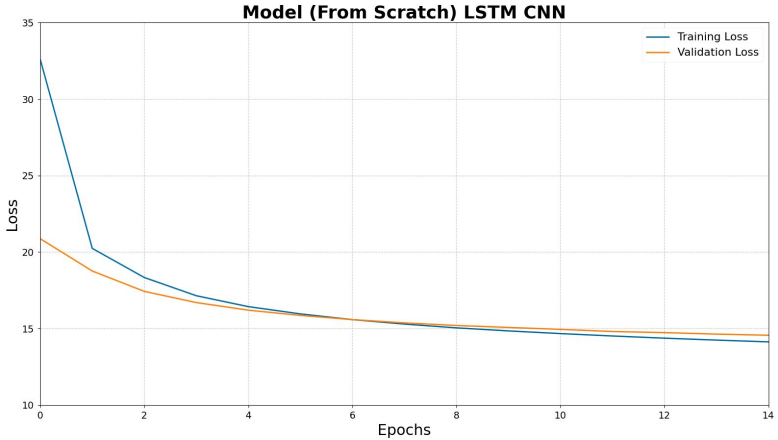
LSTMs are recurrent neural networks that handle long-term dependencies in sequential data using memory cells and gates.

Transformer

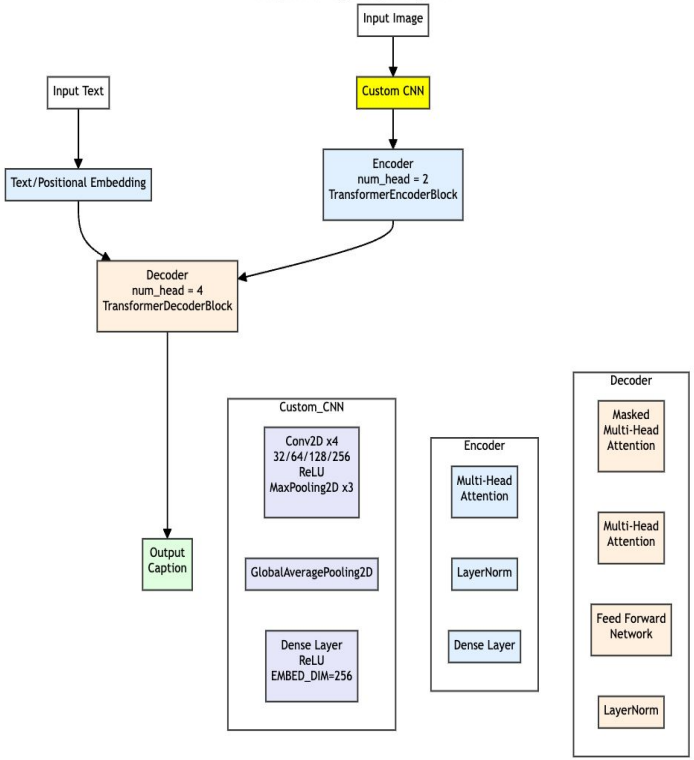
Transformers, use self-attention mechanisms to process entire sequences in parallel, capturing global dependencies enabling better performance on many language tasks.

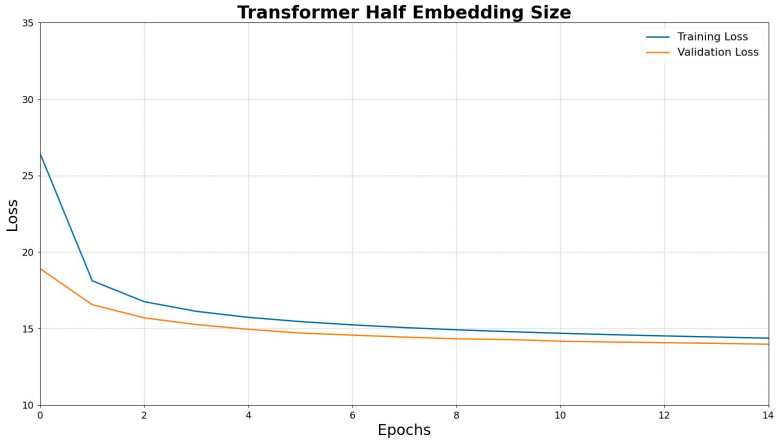
LSTM with Custom CNN



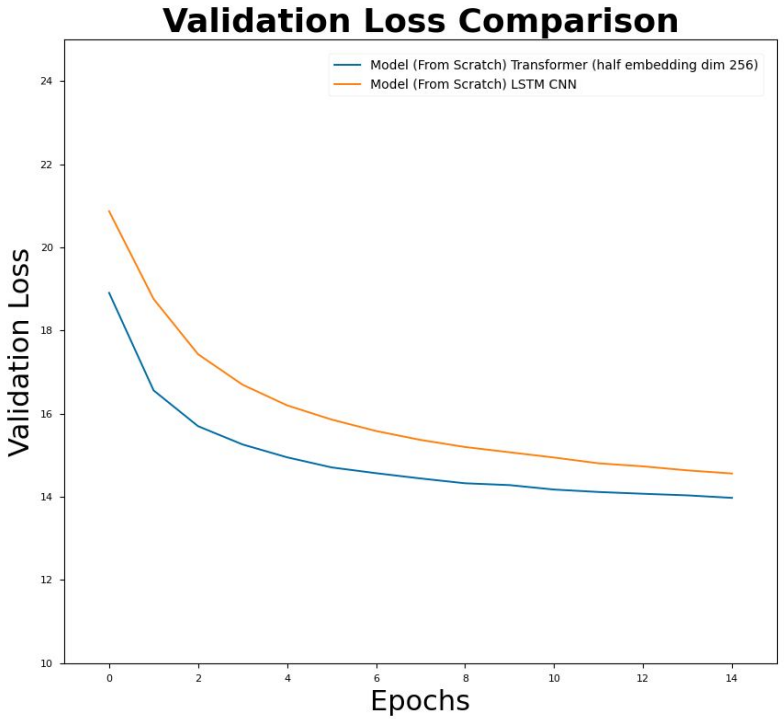


Transformer with Custom CNN



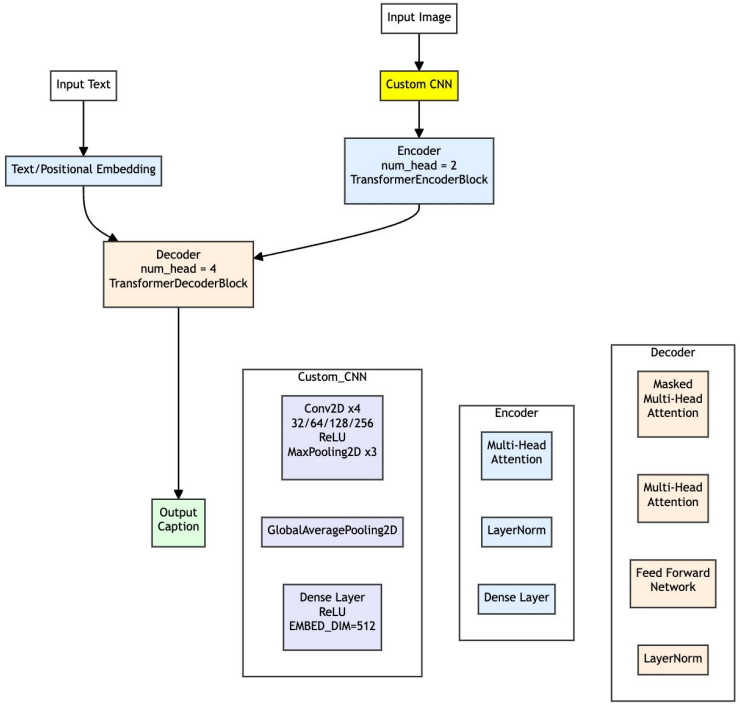


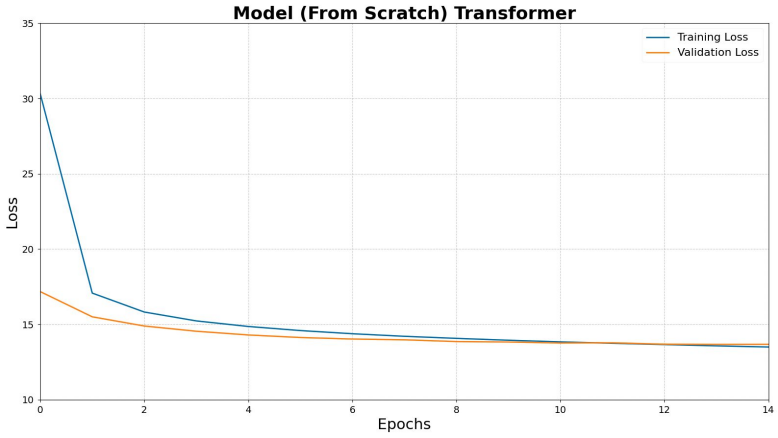
LSTM vs Transformer



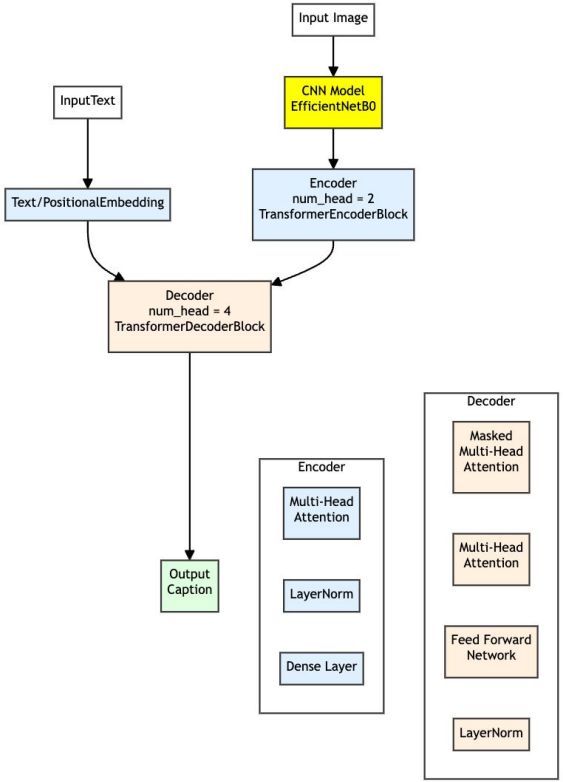
Model based on **transformers outperforms** the model based on LSTM.

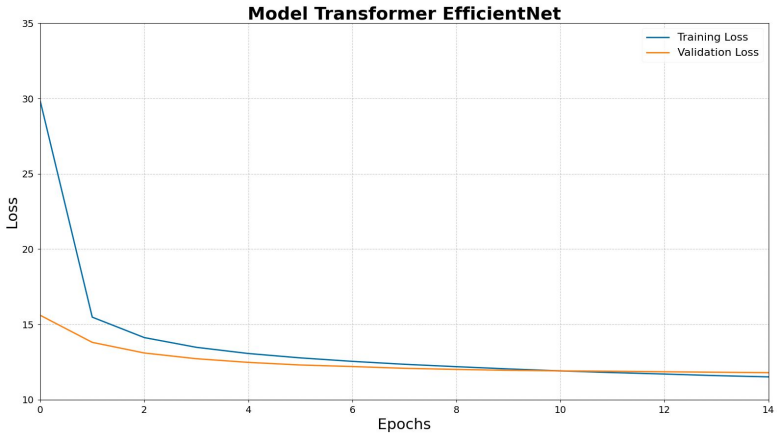
Transformer with Custom CNN



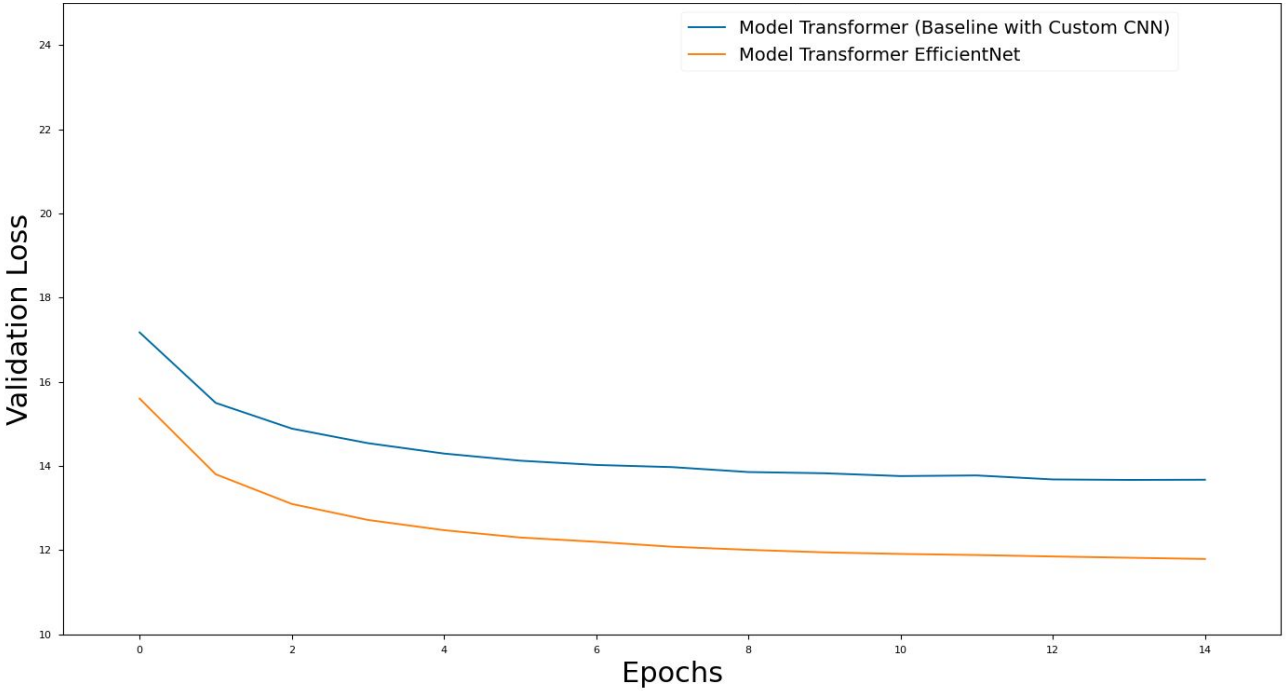


Transformer with EfficientNet CNN





Custom CNN vs EfficientNet

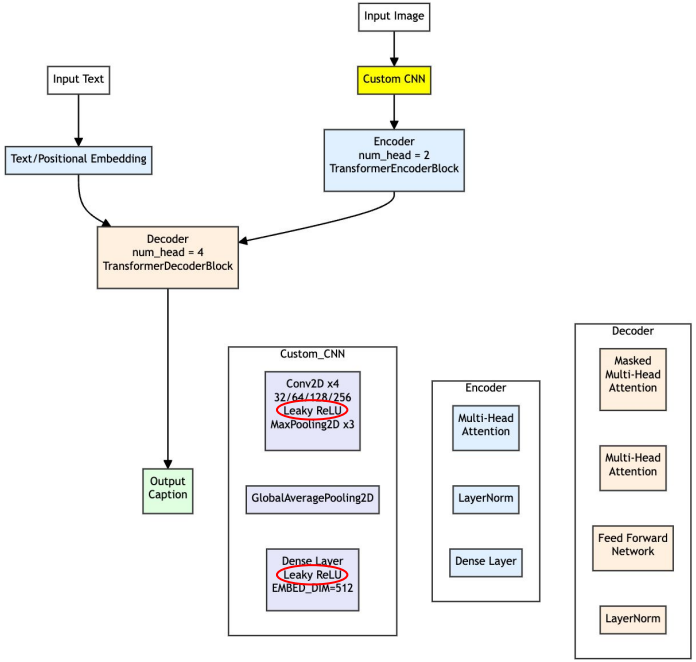


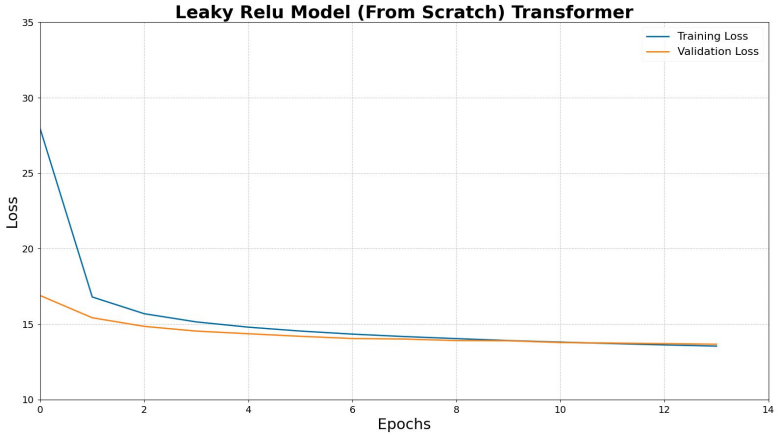
Pre-trained model based on **EfficientNet outperforms** the Model with a Custom CNN.

Pretrained (EfficientNetB0) > Custom CNN

Let's see if we can match EfficientNet!

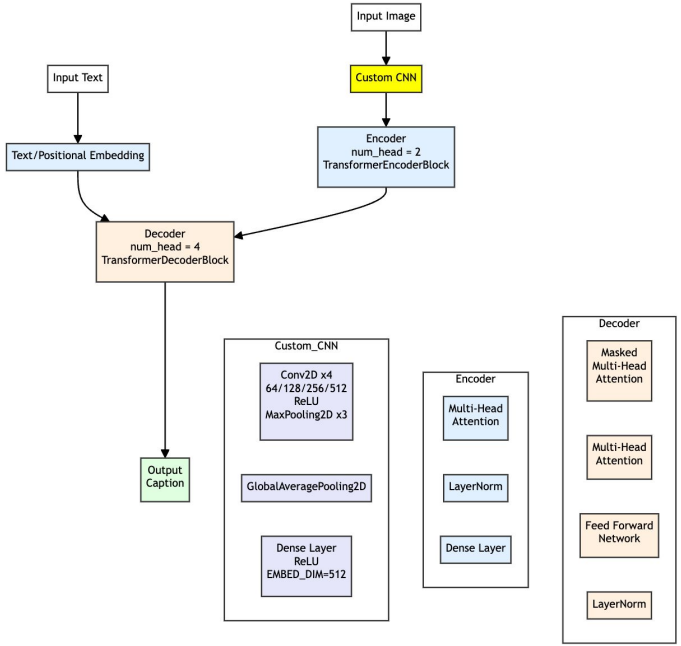
Transformer with Leaky ReLU

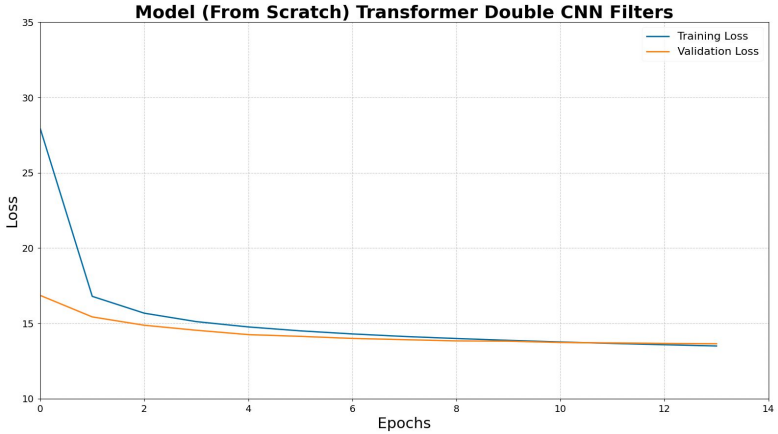




+0.94% Improvement compared to Base Transformer with Custom CNN model

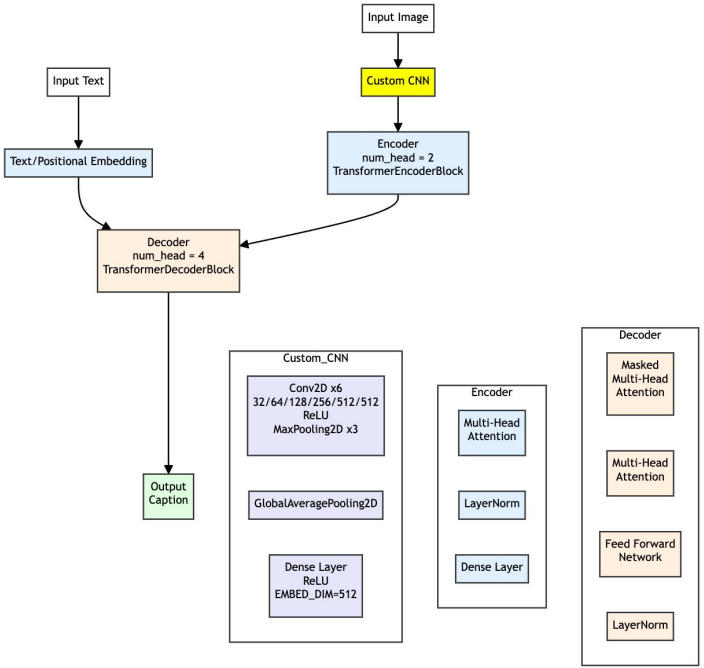
Transformer with 2x CNN Size

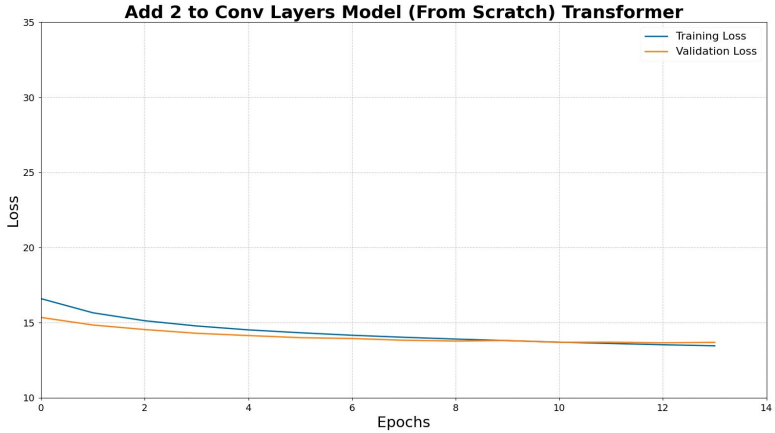




+0.17% Improvement compared to Base Transformer with Custom CNN model

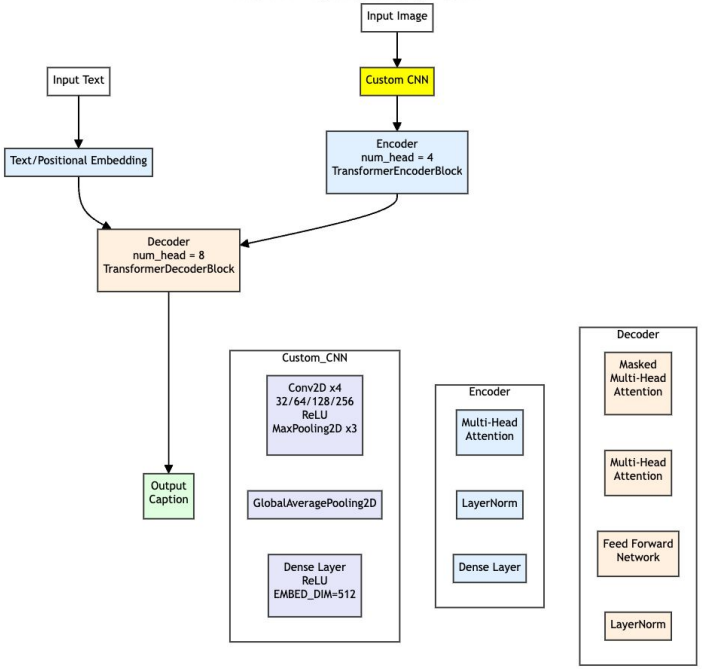
Transformer with +2 Conv layers

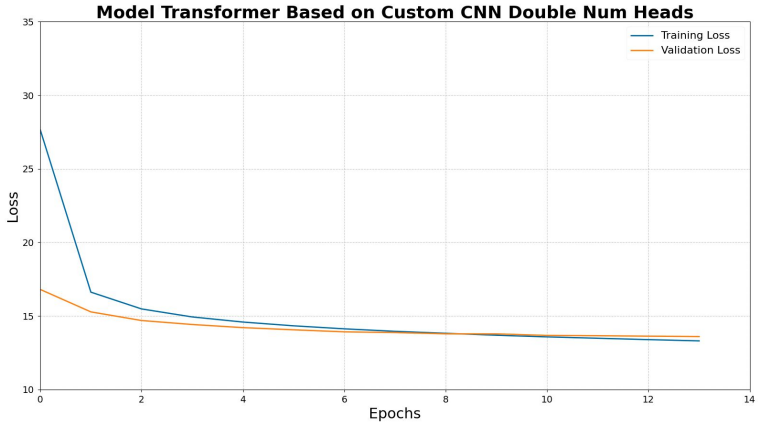




-0.11% Improvement compared to Base Transformer with Custom CNN model

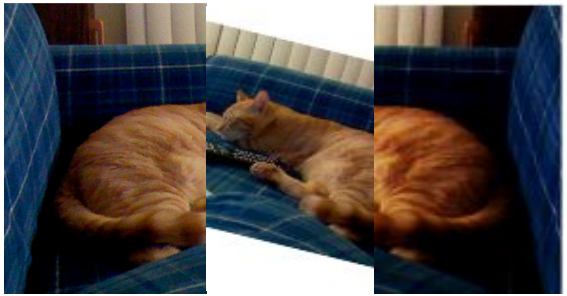
Transformer with 2x num_heads





+0.52% Improvement compared to Base Transformer with Custom CNN model

Improved data augmentation



Base

Random Horizontal flip,
Random Rotation,
Random Contrast

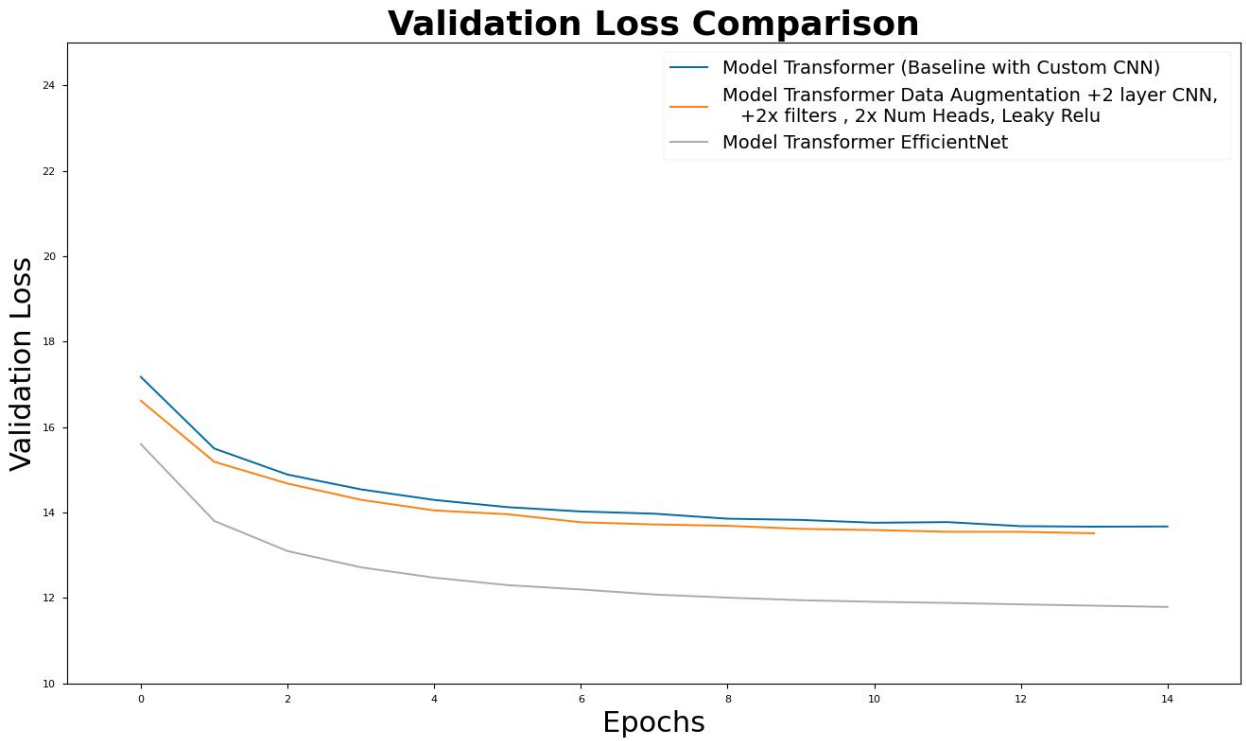
Zoom

Randomly zooms in/out
Factor: ±20% of original size

Brightness

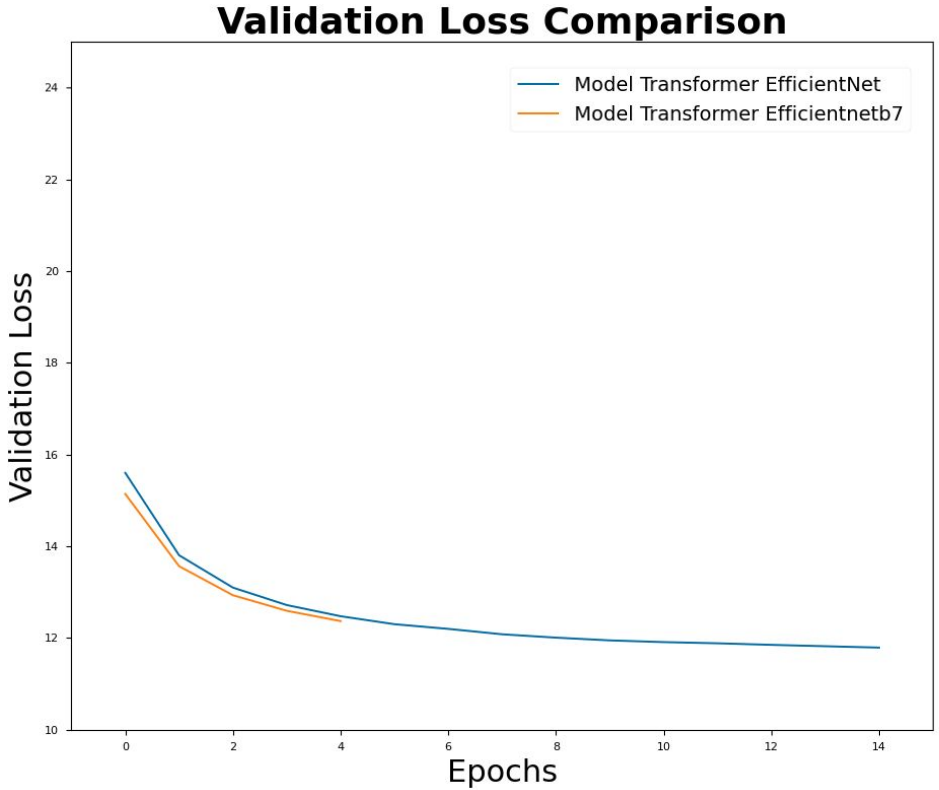
Randomly adjusts brightness
Factor: ±20% intensity

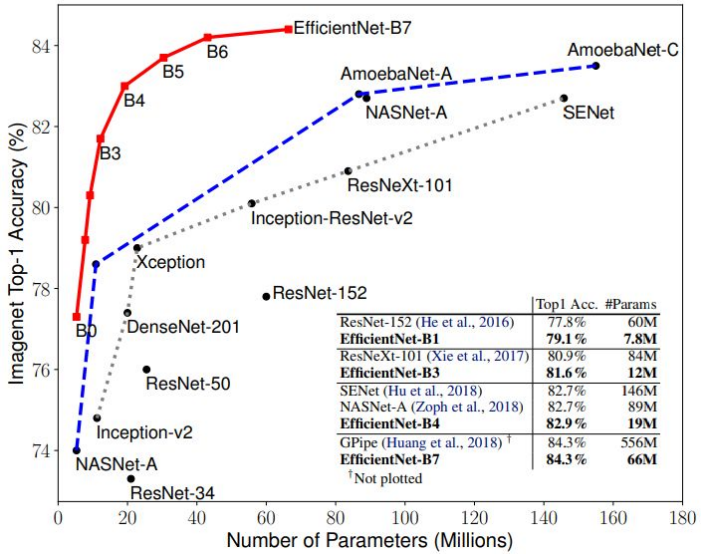
Architecture Optimization



After numerous improvements to the original transformer architecture Efficient Net remains vastly superior.

Architecture Optimization

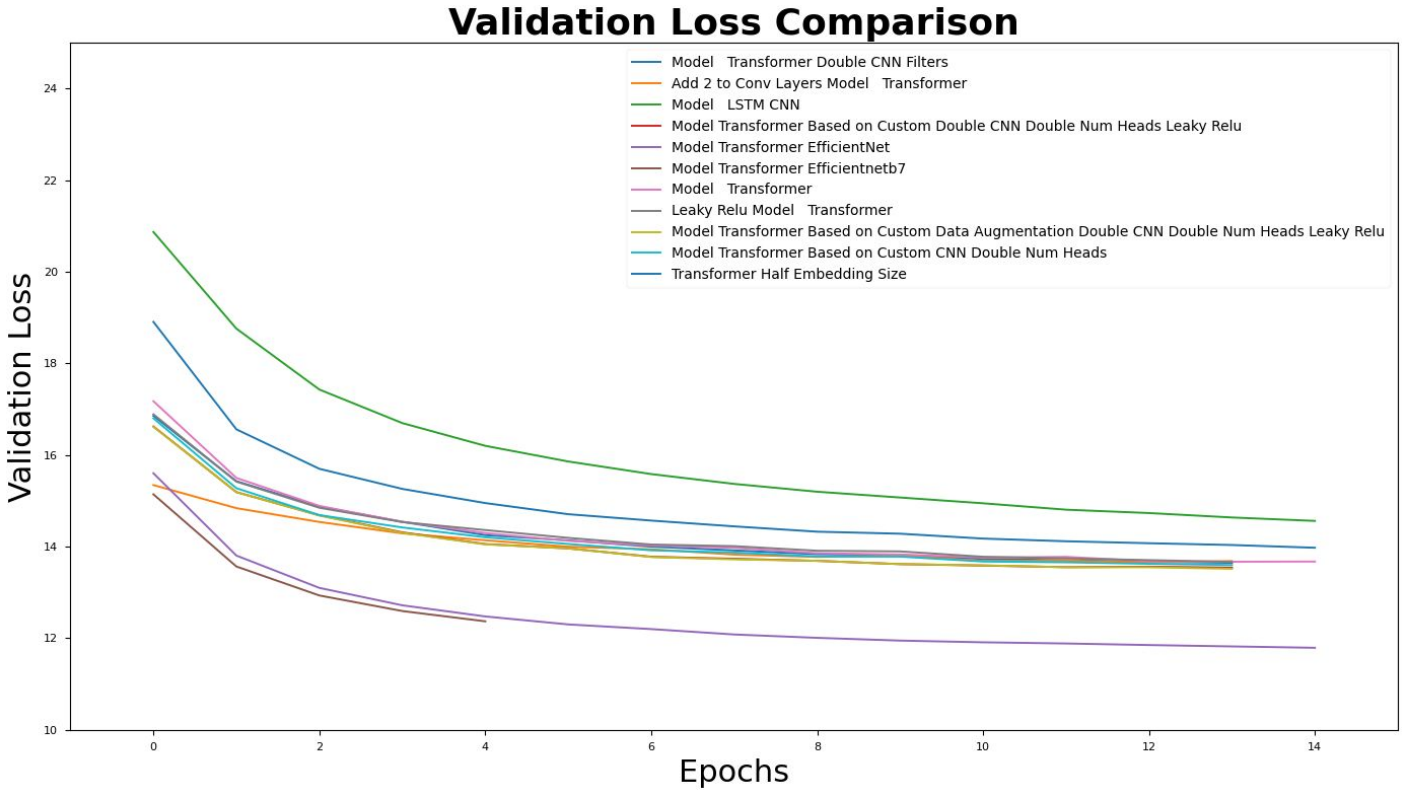




EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks" Mingxing Tan et. al. 2019

Efficient (B0) due to it's small size could be trained for longer and achieve better performance compared to the B7 using 3.2h on Nvidia T4 GPU.

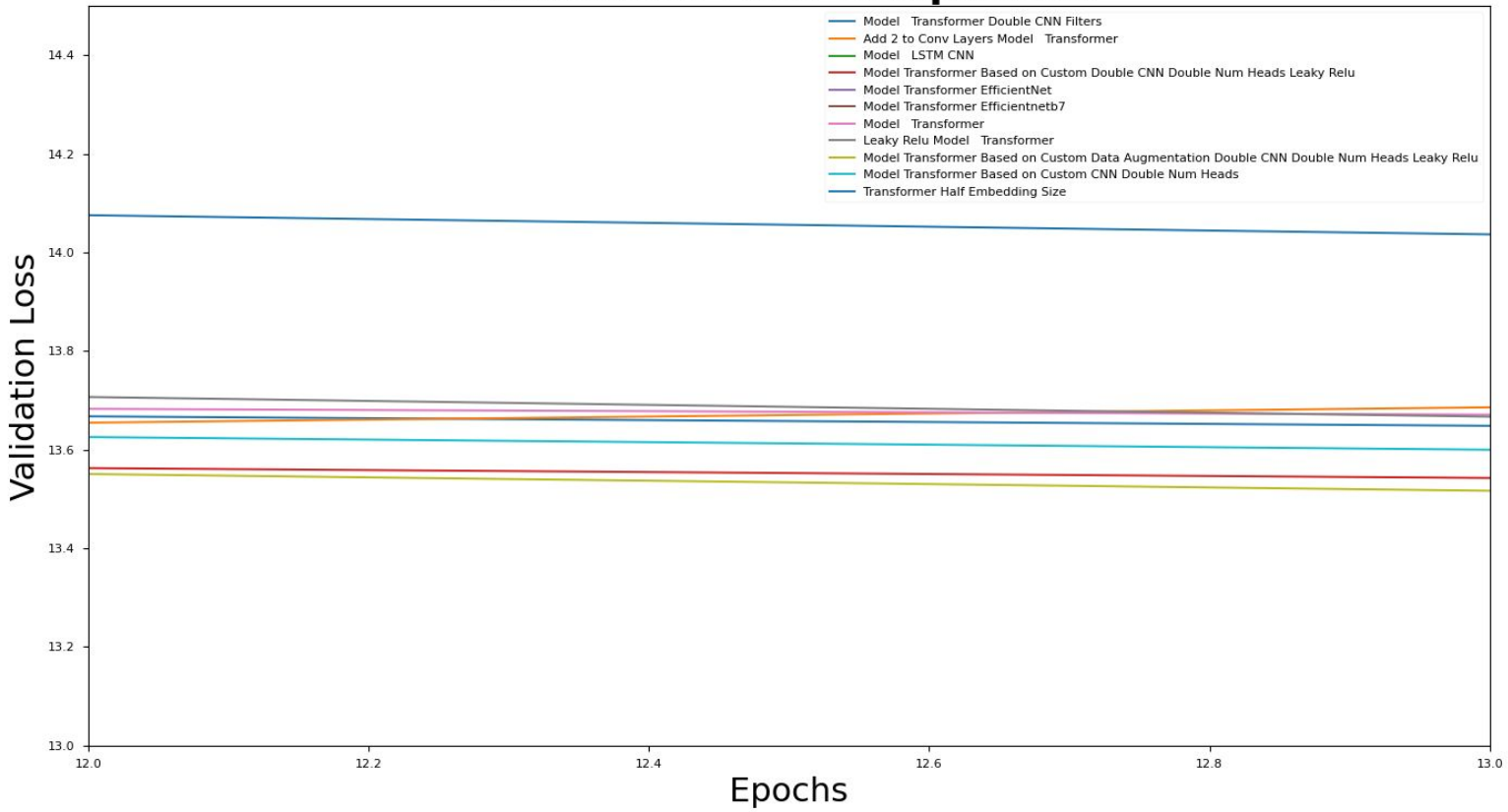
Comparison



EfficientNetB0 & EfficientNetB7 outperform the other models.

Comparison

Validation Loss Comparison



Regularization

Default for all the Transformer model	**Ridge and Lasso regression?**
Dropout included in feed forward decoder layers: *Dropout(0.3)*.	There are **no particular signs of overfitting** therefore we decided not* to include Ridge or Lasso regularization.

* After trying to reduce the dataset and increase the number of training epochs to overfit the network, adding regularization in all layers of the CNN (L1 with different hyperparameters did not lead to substantial differences).

Metrics, Learning Rates Loss Function

Sparse Categorical Cross Entropy: measures the difference between predicted and actual captions.

Learning Rate Schedule:
Custom LRSchedule class implementing warm-up strategy that gradually increases learning rate (1/15 total training steps) and maintains constant learning rate: 1e-4.

Optimizer: Adam.

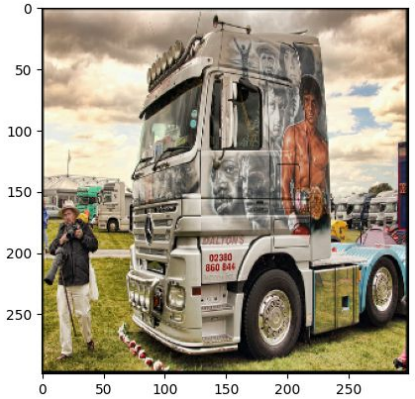
Training Monitoring: Early Stopping , custom callbacks for Model Checkpoint during training.

Model Comparison

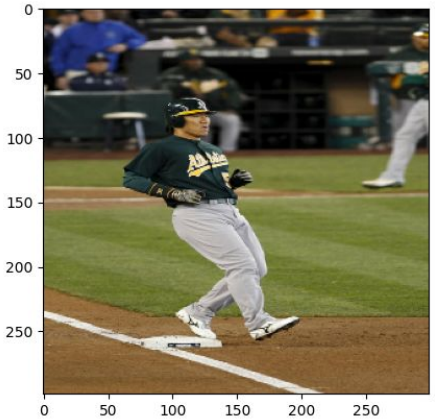
Model	Epoch	Validation Loss	% Improvement
Model Transformer EfficientNet	14	11.8223	↑ 13.52%
Model Transformer Efficientnetb7	5	12.3706	↑ 9.51%
Model Transformer Based on Custom Data Augmentation Double CNN Double Num Heads Leaky Relu	14	13.5168	↑ 1.13%
Model Transformer Based on Custom Double CNN Double Num Heads Leaky Relu	14	13.5429	↑ 0.94%
Model Transformer Based on Custom CNN Double Num Heads	14	13.6002	↑ 0.52%
Model Transformer Double CNN Filters	14	13.6487	↑ 0.17%
Leaky Relu Model Transformer	14	13.6672	↑ 0.03%
Model Transformer	14	13.6712	0.00%
Add 2 to Conv Layers Model Transformer	14	13.6862	↓ -0.11%
Transformer Half Embedding Size	14	14.0370	↓ -2.68%
Model LSTM CNN	14	14.6381	↓ -7.07%

Prediction of best Custom Model

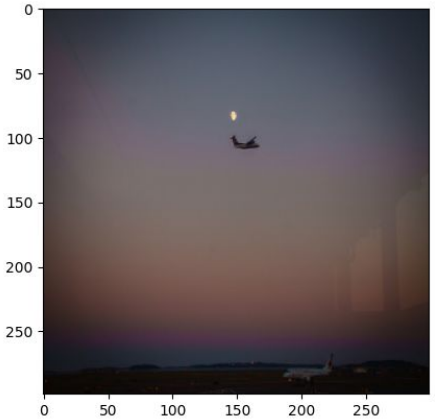
4 M parameter CNN + 16M transformer



Two men standing on the side of the train.



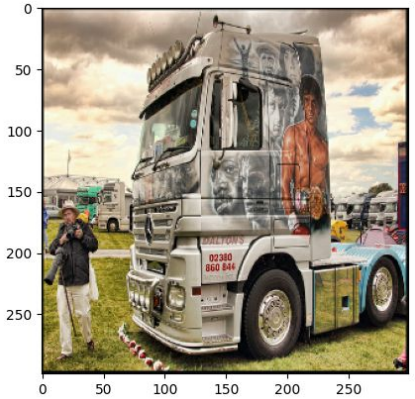
Two baseball teams playing the baseball game of baseball.



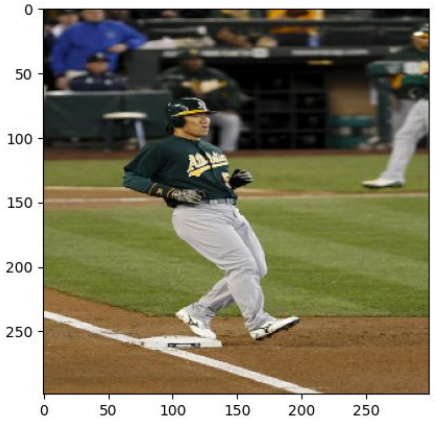
The airplane has landed in the sky.

Prediction of Efficient Net Model

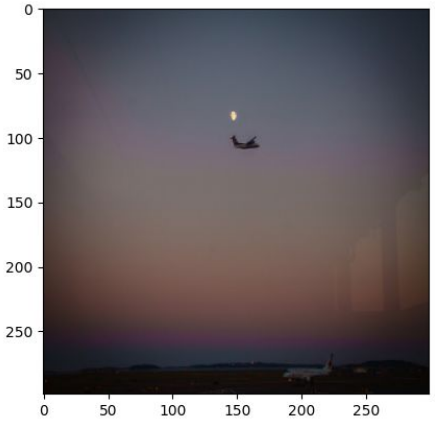
4 M parameter CNN + 16M transformer



The truck has been loaded in the middle of a field.

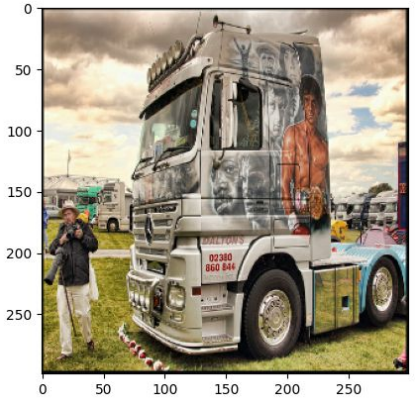


Baseball player is getting ready to throw his pitch during baseball game.

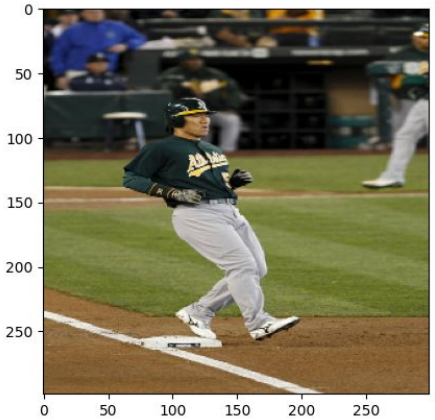


The kite flying over a body with the ocean on a sunny sunset.

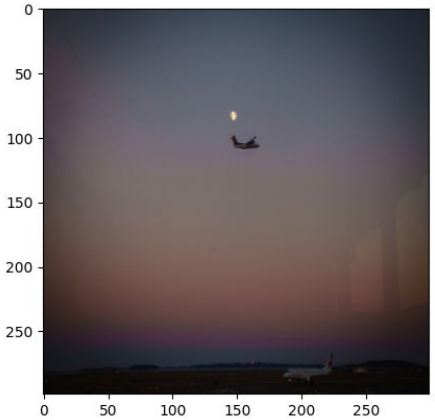
Florence 2 base v0.2b



A large truck with a picture of a man on the side of it.

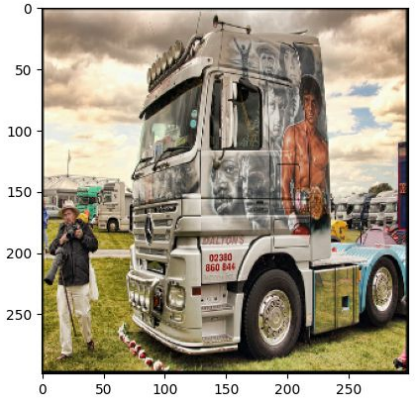


A baseball player running to first base during a game.

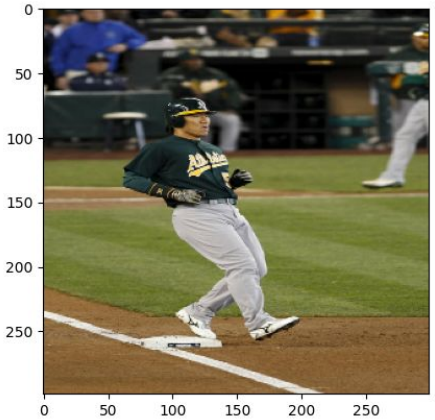


A plane flying in the sky with a full moon in the background.

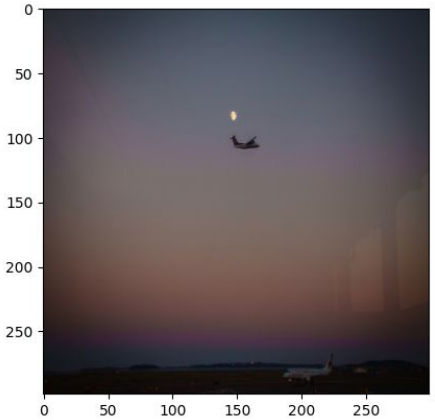
Prediction of PaliGemma 3b



A large truck with a painting of a man on the side, showcasing a variety of details. The truck has a large windshield...



A baseball player, wearing a green and yellow jersey and a black helmet, sprints to first base after hitting a ball...



A plane flies high in the sky at night, its tail shining brightly against the clear sky. The plane is on the ground...

Failed & Alternative Experiments

Type	Details
Florence-2 Finetuning	The predicted captions were not semantically correct.
Add more transformer layers	Using more resources, model complexity can be significantly enhanced.
Evaluate other architectures (e.g. GRU)	Explore alternative architectures, such as GRU, to potentially improve performance.

Thanks!

Do you have any questions?