Project II - Transformers

Deep Learning 2024

Kinga Frańczak, 313335
Grzegorz Zakrzewski, 313555

Contents

1. Introduction
2. Experiments
3. Results
4. Conclusions

Description of the research problem

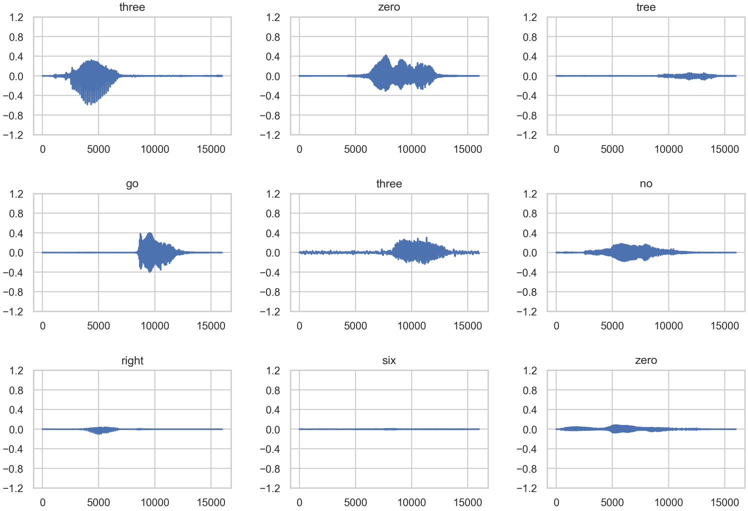


Figure: Sample audio clips from Speech Commands dataset.

Transformer - encoder

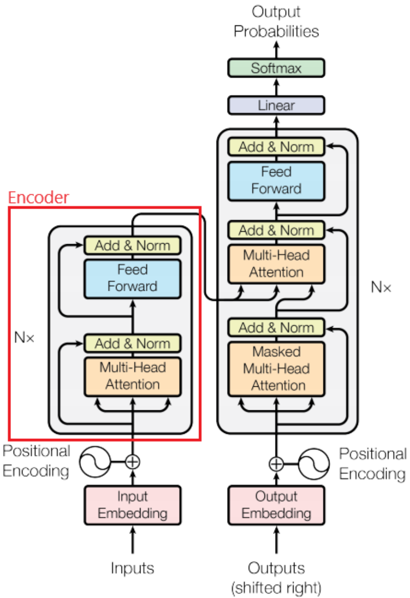
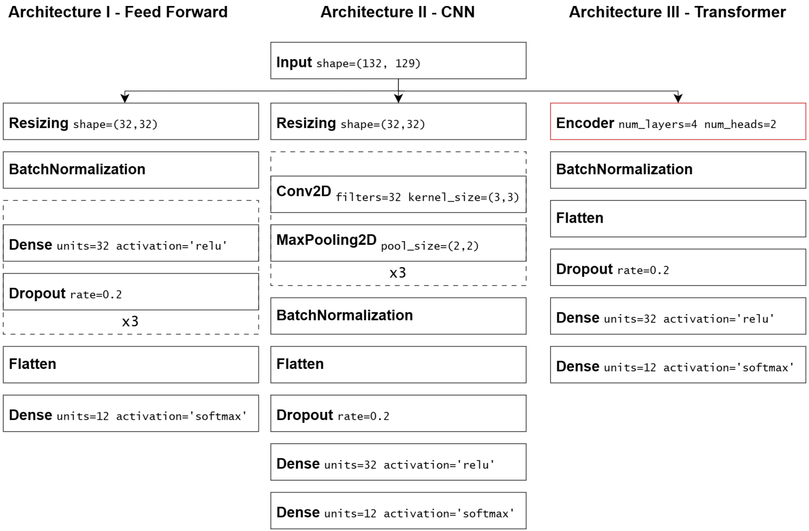


Figure: The original "Attention is All You Need" Transformer diagram.

Experiments - network architectures



Architecture I - Feed Forward

Architecture II - CNN

Architecture III - Transformer

Input `shape=(132, 129)`

Architecture I - Feed Forward

- **Resizing** `shape=(32,32)`
- **BatchNormalization**
- **Dense** `units=32 activation='relu'`
- **Dropout** `rate=0.2`
- x3
- **Flatten**
- **Dense** `units=12 activation='softmax'`

Architecture II - CNN

- **Resizing** `shape=(32,32)`
- **Conv2D** `filters=32 kernel_size=(3,3)`
- **MaxPooling2D** `pool_size=(2,2)`
- x3
- **BatchNormalization**
- **Flatten**
- **Dropout** `rate=0.2`
- **Dense** `units=32 activation='relu'`
- **Dense** `units=12 activation='softmax'`

Architecture III - Transformer

- **Encoder** `num_layers=4 num_heads=2`
- **BatchNormalization**
- **Flatten**
- **Dropout** `rate=0.2`
- **Dense** `units=32 activation='relu'`
- **Dense** `units=12 activation='softmax'`

Figure: Three neural network architectures used in the experiments.

Experiments - details

Experiment	Objective	Values
1	Architecture	Simple feed-forward CNN Transformer
2.1	Number of attention heads	2 4 6 8 10
2.2	Number of Encoder sub-layers	2 4 6 8
3	Handling *silence* and *unknown* classes	one network for all classes separate network for special cases

Table: Details of the experiments.

Experiment 1 - architecture

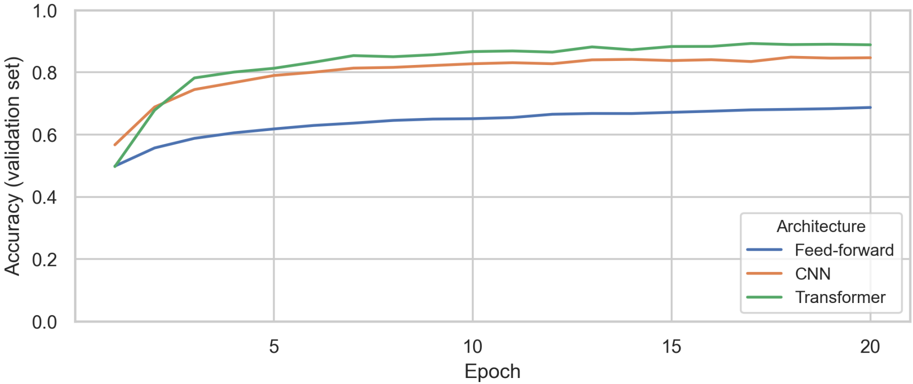


Figure: Accuracy computed on the validation subset for every epoch.

Experiments 2.1 & 2.2 - hyper-parameters

Number of		
attention heads	Accuracy	Validation accuracy
2	0.930 (0.004)	0.901 (0.003)
4	0.914 (0.008)	0.888 (0.012)
6	0.866 (0.036)	0.845 (0.025)
8	0.794 (0.021)	0.791 (0.016)
10	0.803 (0.023)	0.798 (0.012)

(a) Experiment 2.1 - number of attention heads.

Number of		
Encoder sub-layers	Accuracy	Validation accuracy
2	0.941 (0.003)	0.875 (0.004)
4	0.930 (0.004)	0.897 (0.006)
6	0.905 (0.015)	0.884 (0.009)
8	0.468 (0.132)	0.424 (0.174)

(b) Experiment 2.2 - number of Encoder sub-layers.

Table: The mean (and standard deviation) of the best values of the accuracy achieved by models.

Experiment 3 - handling *silence* and *unknown* classes.

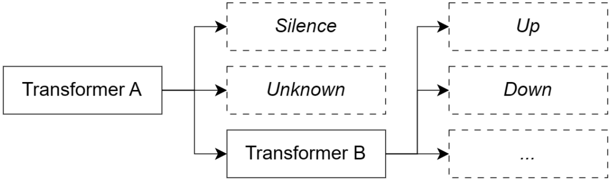
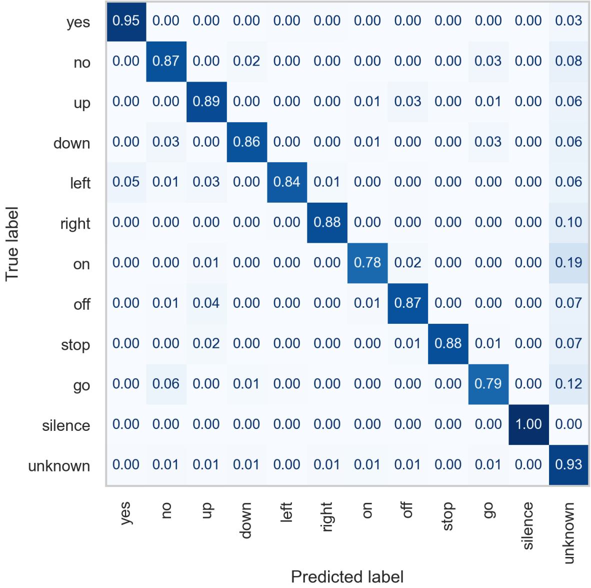


Figure: Experiment 3 - diagram

	Accuracy	Validation accuracy
Transformer A	0.947 (0.015)	0.934 (0.011)
Transformer B	0.940 (0.003)	0.913 (0.005)
Combination	0.944 (0.014)	0.917 (0.011)

Table: Experiment 3 - results.

Confusion matrix



	yes	no	up	down	left	right	on	off	stop	go	silence	unknown
yes	0.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03
no	0.00	0.87	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.08
up	0.00	0.00	0.89	0.00	0.00	0.00	0.01	0.03	0.00	0.01	0.00	0.06
down	0.00	0.03	0.00	0.86	0.00	0.00	0.01	0.00	0.00	0.03	0.00	0.06
left	0.05	0.01	0.03	0.00	0.84	0.01	0.00	0.00	0.00	0.00	0.00	0.06
right	0.00	0.00	0.00	0.00	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.10
on	0.00	0.00	0.01	0.00	0.00	0.00	0.78	0.02	0.00	0.00	0.00	0.19
off	0.00	0.01	0.04	0.00	0.00	0.00	0.01	0.87	0.00	0.00	0.00	0.07
stop	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.88	0.01	0.00	0.07
go	0.00	0.06	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.79	0.00	0.12
silence	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
unknown	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.93

Figure: Confusion matrix prepared on the validation dataset, normalized by true conditions (rows).

Conclusions

- The project was prepared according to the instructions.
- The best parameter settings:
  - number of attention heads $= 2$;
  - number of Encoder sub-layers $= 4$;
  - separate network for *silence* and *unknown* classes;
- The accuracy achieved on the validation subset (around 0.91) seems very high.
- The accuracy achieved on Kaggle is 0.67.