Clickbait Detection
Team: Pied Piper

Pratik Sanjay Patil, 11840850
Thummala Milind Kesar, 11841160

Abstract

Clickbait detection is a very important and daily occurring problem that most of us face. It refers to content whose main purpose is to attract attention and encourage visitors to click on a link to a particular web page. It is important to detect clickbait. In this report we present a model to detect clickbait in two scenarios - when only the headline of the link is present and when headline along with thumbnail is present. We have discussed feature extraction and modeling approaches and present their results and details in this report.

Previous Work

We had previously worked on the fake news detection task which is also a text classification task and very similar to the clickbait problem. As part of our improvement phase we were asked to work on the clickbait problem and to incorporate thumbnails as well.

Introduction

Clickbait Detection is a very popular problem in machine learning. In day-to-day lives, we encounter a lot of articles, which, with the help of clickbait, ensures that they get more views and exposure. This is very frustrating for the user and it also consumes a lot of time, so it is very important to capture such click-baits. The problem which is more popular just contains classification based on the title (text classification), In our project, we have taken the thumbnail associated with the text (title, in this case) into consideration as well. So first we will simply do a BERT classification on the first dataset and then on the second dataset, we will use BERT for text feature extraction and RESNET-18 for image feature extraction and then pass both of these into a pipeline and then predict the results.

Problem Definition

We worked on 2 problems.
1. Given the title of an article, we should be able to predict if the article is clickbait for that title or not (Using dataset-1).
2. Given a title and a thumbnail associated with it, we should be able to predict if the article/video is clickbait or not (Using dataset-2).

Objective

The objective is to identify whether any content (link) to an article/video is clickbait or not. We use the concept of transfer learning along with fine tuning for solving the problem and perform the feature extraction of text and image data using BERT and ResNet-18 respectively. The details of the architecture are specified below.

Technology Used
The models are defined in the "Procedure and Models Used" section.

Dataset Used
Dataset-1
This dataset contains three columns: title, text and label. This dataset has been taken from Kaggle (https://www.kaggle.com/c/clickbait-news-detection/data). Here we need to identify whether the content is clickbait or not based on the title/headline of the link.

Dataset-2
This dataset contains data of various youtube videos, it has various columns: label, ID (of the thumbnail), description, timestamp, view Count, like Count, then 10 columns of for comments, etc. But, we will only be needing the label, ID, and title as thumbnail and title are only needed to make a decision without actually opening the video. This dataset is not available publicly as it is part of a course (link) of Penn State University. So, we mailed the TA of this course and requested for the dataset.

Procedure and Models Used

Clickbait prediction using headline:
In this section we discuss our approach to solve the clickbait problem when we only have the headline of a particular article/video. We have used dataset 1 for this purpose. This problem is equivalent to a text classification task. We use an approach based on transfer learning for NLP with focus on feature extraction techniques for text data. We use pre-trained BERT for this. The penultimate layer of the BERT architecture represents the feature of the headline which we use as input to another hidden layer (size 512) and an output layer of size 2 for classification and then we finetune the architecture. We used softmax activation and fully connected layers.

Clickbait prediction using headline + thumbnail :
In this section we discuss our approach to solve the clickbait problem when given the thumbnail along with the headline of a particular article/video. We have used dataset 2 for this purpose which is data related to youtube videos. We use an approach based on transfer learning with focus on feature extraction techniques for text and image data.

Our architecture is as follows:
- First we extract the feature representation of the thumbnail using ResNet 18 and the feature extraction from headline using BERT. The penultimate layer in both these architectures, i.e., just before the output layer is the feature representation of our input.
- We concatenate these features from both the models as our combined feature and use this as input to a simple neural network architecture with a single hidden layer of size 512. We use fully connected layers and the output of two neurons with softmax activation corresponding to whether the input is clickbait or not.
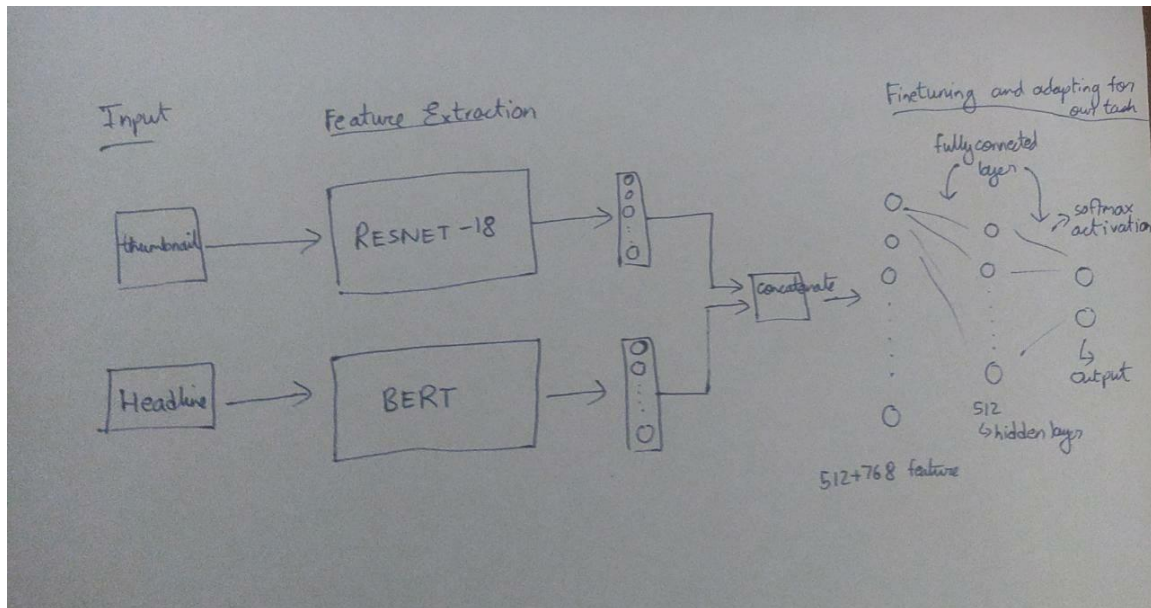- We have used appropriate dropout for regularization and relu activation for non linearity.

Figure 1

Implementation:

Clickbait prediction using headline + thumbnail :

We use 80% of the data for training, 10% for validation and 10% for testing. It is possible to finetune the weights of BERT and ResNet-18 for our downstream application but by experiments we found that freezing these weights and only training the weights of our hidden layer and output layer acts as a sort of regularization and generalizes well. Hence we have used this approach. We have used cross entropy loss and AdamW optimizer with a learning rate of 2×10^{-5}. We trained the model for 20 epochs. There was no bias in the class distribution of our training data. For tokenizing the data for Bert, we used a max length of 20 with padding. For preprocessing the images before passing into ResNet-18, we normalized the image and scaled it to 224x224 from 480x360.
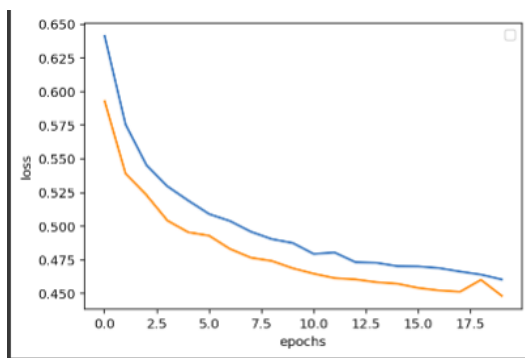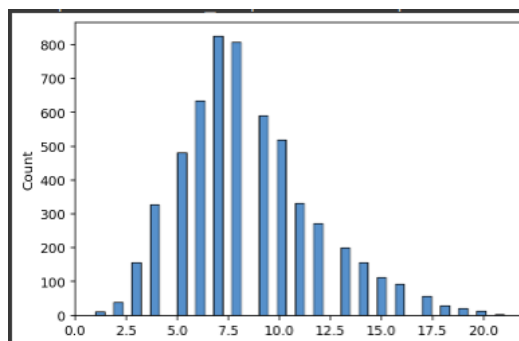

Figure 2

Figure 3

Figure 2: The training curve where orange line represents validation loss and blue line represents training loss. **Figure 3 :** The length of different headlines which used to determine the length of input to Bert (20 with padding).

Clickbait prediction using headline:

We use 70% of the data for training, 15% for validation and 15% for testing. It is possible to finetune the weights of BERT for our downstream application but by experiments we found that freezing these weights and only training the weights of our hidden layer and output layer acts as a sort of regularization and generalizes well. Hence we have used this approach. We have used cross entropy loss and AdamW optimizer with a learning rate of 2×10^{-5}. For this dataset there was a bias in the class distribution for the training process and hence we accounted for this by giving the class weights in the loss function.

Results and Performance:

Clickbait prediction using headline + thumbnail :

We present the accuracy, F1-metrics of the model. We also show the accuracy of two more models which correspond to using only the thumbnail and using only the headline.
We see that our approach outperformed using only a thumbnail or only the headline.
For the purpose of training using only a thumbnail or only headline we trained the model under similar conditions except that there was no concatenation operation with the missing feature representation.
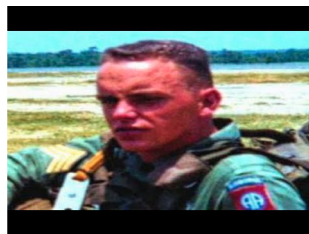
	Accuracy	Weighted Avg f1-score
Only thumbnail	0.8	0.79-0.8
Only headline	0.78	0.78
Our approach	0.85	0.85

Hence we see that our approach outperforms using only thumbnail or headline and generalizes well as well as takes the best of both worlds.

Sample clickbait predictions of our model:

```
Title: "Man Found In Vietnam Jungle After 40 Years Stuns Police"
Image:
```


The example is a clickbait video which is correctly predicted by our model.

Title: "You Laugh = You're a Gamer"
Image:

This is another example of clickbait correctly predicted by our model.

Clickbait prediction using headline:
In this case, we got an accuracy of 0.73 and weighted average f1 score 0.74.

Conclusion:
We have made a model for clickbait detection in two different scenarios - (1) we are given only the headline (2) we are given headline as well as thumbnail. We discuss the implementation details of our approach and show some advantages.

References
- Code - link
- https://www.kaggle.com/c/clickbait-news-detection/data
- https://www.kaggle.com/c/clickbait-thumbnail-detection
- https://www.analyticsvidhya.com/blog/2020/07/transfer-learning-for-nlp-fine-tuning-bert-for-text-classification/
- https://becominghuman.ai/extract-a-feature-vector-for-any-image-with-pytorch-9717561d1d4c
- https://arxiv.org/pdf/1810.04805.pdf
- https://machinelearningmastery.com/transfer-learning-for-deep-learning/
- https://pytorch.org/tutorials/beginner/finetuning_torchvision_models_tutorial.html
- https://towardsdatascience.com/an-overview-of-resnet-and-its-variants-5281e2f56035