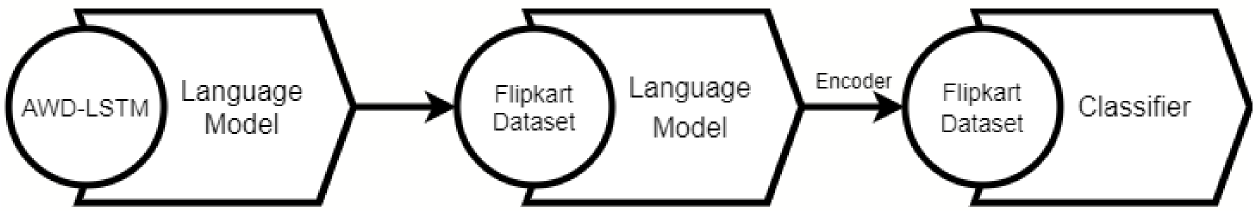
My Approach:

My approach for tackling this challenge is based on finding a simple yet still good solution, for getting good results. The following figure depicts the approach I undertake for this challenge.

Universal Language Model Fine-tuning for Text Classification



The approach shown above was introduced by Howard and Ruder et. al. in 2018. This paper provided a novel method for fine-tuning models for inductive transfer learning. This process includes using the model trained for a particular source task to be used to obtain good performance on other tasks (NLP tasks) as well, just like transfer learning in Computer Vision.

According to the paper, language modeling is the ideal source task and is considered analogous to ImageNet for NLP tasks. The pre-training is to be done on a large corpus of the same language which is to be used in the main task (ie category prediction for our case). This helps in effectively catching the main properties and aspects of language. The pre-training is done to help the model understand the general properties and semantics of the language at hand and then the models have to be tweaked a little to suit the specific task. This would be something like the Image-Net corpus, but, for language and NLP tasks.

We don't have to build the language model from scratch every time. It has to be performed only once. The resulting pre-trained model which we get can be reused for the next stages. I decided to use ULMFiT since it is found that pre-training was especially useful for small datasets and medium-sized datasets.

ULMFiT has revolutionized the field of NLP in the past few years, improving the scope of deep learning in NLP. It made training models in almost no time possible for various NLP tasks. This set the base for transfer learning for NLP and paved the way for ELMo, GPT, GPT-2, BERT, and XLNet.

I have used **AWD-LSTM architecture**. Why?

The AWD-LSTM stands for ASGD Weight-Dropped LSTM. It uses DropConnect and a variant of Average-SGD (NT-ASGD) along with several other well-known regularization strategies.

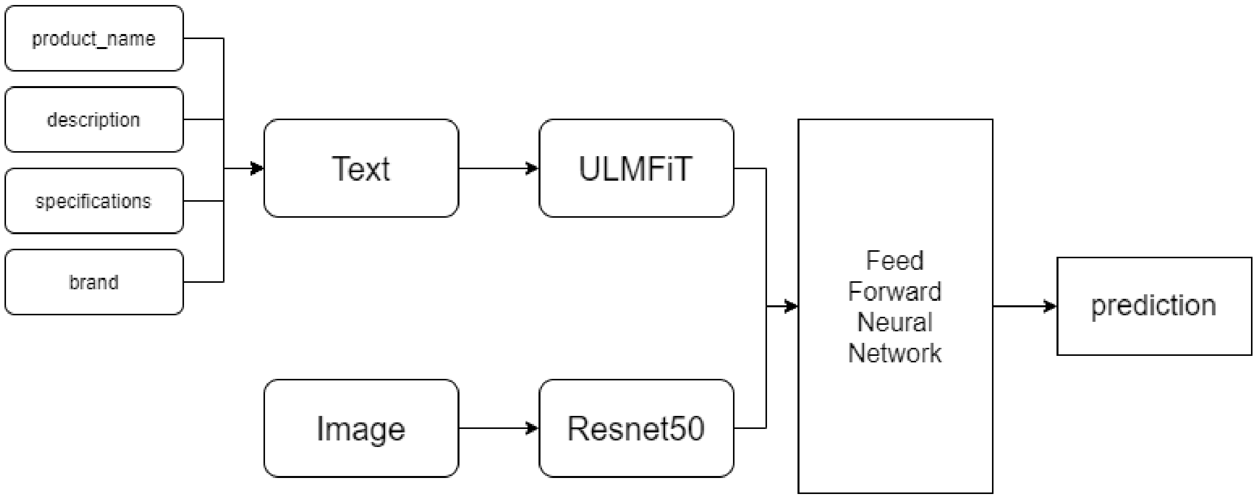
As discussed in the "My approach" section above, language modeling is found to be the ideal source task for learning the semantics of the language and using the encoder for other NLP tasks. Since, the AWD-LSTM has been dominating the state-of-the-art language modeling, it would be the right choice for our task at hand. It is also known that all the top research papers on word-level models incorporate AWD-LSTMs.

The research paper – [Regularizing and Optimizing LSTM Language Models](https://arxiv.org/abs/1708.02182) introduced the AWD-LSTM architecture for the first time. The paper investigates a set of regularization and optimization strategies for word-based language modeling tasks highly effective. Moreover, they can also be used with no modification to existing LSTM implementations.

Results:

Model Description		Accuracy
threshold_value=0 (all categories)	Only description	96.75
	Only description without top 2 cats	97.65
	description, brand, product_name	96.81
	description, brand, product_name without top 2 cats	96.89
threshold_value=10 (removing cats with less than 10 data points)	Only description	97.82
	Only description without top 2 cats	97.60
	description, brand, product_name	97.60
	description, brand, product_name without top 2 cats	97.68
threshold_value=30 (removing cats with less than 10 data points)	Only description	96.94
threshold_value=100 (removing cats with less than 10 data points)	Only description	97.83

Future work:



In the future, the architecture implemented in this project can be extended by taking in account the predictions from the images as well (refer to the above diagram). Due to time constraints, I was unable to implement it.

Another way to approach this challenge would be to use machine translation as shown in [this paper](https://arxiv.org/pdf/1812.05774.pdf). this method helps in predicting the whole category tree. It shows that state-of-the-art machine translation (MT) models surpass previous classification approaches in categorizing products in two large real-world e-commerce datasets. Moreover, novel roof-to-leaf category paths are found by using NMT models.

References:

Howard, J. and Sebastian Ruder. "Universal Language Model Fine-tuning for Text Classification." *ACL* (2018).

Merity, Stephen, N. Keskar and R. Socher. "Regularizing and Optimizing LSTM Language Models." *ArXiv* abs/1708.02182 (2018): n. pag.

Li, M., Stanley Kok and Liling Tan. "Don't Classify, Translate: Multi-Level E-Commerce Product Categorization Via Machine Translation." *ArXiv* abs/1812.05774 (2018): n. Pag.

Cevahir, A. and Murakami, K. (2016). Large-scale multi-class and hierarchical product categorization for an e-commerce giant. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 525–535.

Ha, J.-W., Pyo, H., and Kim, J. (2016). Large-scale item categorization in e-commerce using multiple recurrent neural networks. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 107–115. ACM.

Wirojwatanakul, Pasawee and A. Wangperawong. "Multi-Label Product Categorization Using Multi-Modal Fusion Models." *ArXiv* abs/1907.00420 (2019): n. pag.

Seth, Y. (2018, November 30). What makes the AWD-LSTM great? Retrieved April 10, 2021, from https://yashuseth.blog/2018/09/12/awd-lstm-explanation-understanding-language-model/

Ravi, A. (2020, May 31). Understanding ULMFiT - the shift Towards transfer learning in NLP. Retrieved April 10, 2021, from https://towardsdatascience.com/understanding-ulmfit-and-elmo-the-shift-towards-transfer-learning-in-nlp-b5d8e2e3f664

Kostas. (n.d.). Understanding Fastai's Fit_one_cycle method. Retrieved April 10, 2021, from https://iconof.com/1cycle-learning-rate-policy/

/author/sylvain-Gugger.html. (2018, April 07). The 1cycle policy. Retrieved April 10, 2021, from https://sgugger.github.io/the-1cycle-policy.html

Smith, Leslie N.. "Cyclical Learning Rates for Training Neural Networks." *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2017): 464-472.

Smith, Leslie N. and Nicholay Topin. "Super-convergence: very fast training of neural networks using large learning rates." *Defense + Commercial Sensing* (2019).

Smith, Leslie N.. "A disciplined approach to neural network hyper-parameters: Part 1 - learning rate, batch size, momentum, and weight decay." *ArXiv* abs/1803.09820 (2018): n. pag.