



Apprentissage Profond

Génération de proverbes

Groupe L34

Élèves :

THEVENET Louis

LEBOBE Timothé

TENE Zacharie

SABLAYROLLES Guillaume

31 Mars 2025

Table des matières

1. Génération de proverbes en anglais	3
2. Consitution de la base de données	3
2.1. Acquisition des données	3
2.2. Partionnement des données	3
2.3. Script de chargement des données	3
3. Création et entraînement du modèle	4
3.1. Création du modèle	4
3.2. Entraînement	4
3.3. Génération	5
4. Analyse des résultats	5
4.1. Entraînement sur les proverbes anglais	5
4.2. Entraînement sur les proverbes traduits	7
5. Conclusion	7

1. Génération de proverbes en anglais

Nous avons choisi de créer un modèle de génération de proverbes. La base de donnée est trouvable dans le dossier `raw_data/` à la racine de ce [dépôt GitHub](#)

Voici un exemple de proverbes de notre base d'entraînement :

- He that brings good news, knocks hard.
- Anger and haste hinder good counsel.
- Big thunder, little rain.
- Romeo must die in order to save the love.
- The point is plain as a pike staff.

2. Consitution de la base de données

2.1. Acquisition des données

Puisqu'il est plus simple de trouver des données en langue anglaise, nous avons choisi de nous limiter à cette langue et avons utilisé des scripts Python de scrapping pour récolter des données sur différents sites internet.

Nous avons 3200 proverbes originaux anglais et 35000 proverbes en incluant des proverbes traduits d'autres langues et prévoyons de tester le modèle sur ces deux bases.

2.2. Partionnement des données

Nous avons choisi de partionner les données de la manière suivante qui est un standart dans l'apprentissage profond :

- 80% des données pour l'apprentissage
- 10% des données pour le test
- 10% des données pour la validation

Nous pourrons augmenter la part d'apprentissage si le manque de donnée a un impact trop important.

2.3. Script de chargement des données

Nous avons réalisé un script de téléchargement et traitement des données. Un exemple d'utilisation est donné dans le fichier `main.ipynb`, il suffit d'appeler la fonction `make_dataset.load_data()` qui renvoie les proverbes classés par sources.

```
1  proverbs_db.txt: 34142 proverbs
2  proverbs_db_only_english.txt: 2208 proverbs
3  proverbs_digest.txt: 1000 proverbs
4  Total: 37350 proverbs
5  Total length: 1853527 characters
6
7
8  Examples of proverbs:
9  Money's for buying and a horse is for riding.
10 A set of white teeth does not indicate a pure heart.
11 Time discloses the truth.
12 The earth has ears, the wind has a voice.
13 The fox will catch you with cunning, and the wolf with courage.
```

3. Création et entraînement du modèle

3.1. Création du modèle

Nous allons faire du fine-tuning à partir du modèle [facebook/opt-125](#) qui est un modèle type GPT-3 basé sur l'architecture Transformers.

Nous testerons également de partir du modèle [TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T](#) qui est plus récent et contient près de dix fois plus de paramètres.

On charge le modèle et son Tokenizer à l'aide des fonctions `AutoModelForCausalLM.from_pretrained()` et `AutoTokenizer.from_pretrained()` de la librairie `transformers`.

On décide des sources de proverbes que l'on va utiliser, puis on les fusionne en une seule liste:

```
1 selected_proverbs_groups = [  
2     "proverbs_db.txt",  
3     "proverbs_digest.txt"  
4 ]  
5  
6 proverbs = []  
7 for group in selected_proverbs_groups:  
8     proverbs.extend(all_proverbs[group])
```

On utilise ensuite la librairie `datasets` pour préparer ces données à l'entraînement. La tokenisation du dataset consiste à calculer la taille du proverbe le plus long et ajouter du padding aux autres pour uniformiser les tailles.

On utilise ensuite la librairie `peft` afin d'utiliser la technique LoRA (Low-Rank Adaptation). Ainsi, on ajoute un petit nombre de nouveaux paramètres entraînaables au modèle afin de l'adapter à la nouvelle tâche.

Paramètres de la configuration LoRA:

```
1 LoraConfig(  
2     r=8,  
3     lora_alpha=16,  
4     target_modules=["q_proj", "v_proj"],  
5     lora_dropout=0.05,  
6     bias="none",  
7     task_type="CAUSAL_LM"  
8 )
```

La méthode `get_peft_model` permet d'obtenir un nouveau modèle à partir de cette configuration de notre modèle initial.

3.2. Entraînement

A l'aide de la librairie `transformers`, on définit les paramètres d'entraînement:

```
1 TrainingArguments(  
2     output_dir="./results",  
3     per_device_train_batch_size=8,  
4     per_device_eval_batch_size=8,  
5     num_train_epochs=1,  
6     logging_dir='./logs',  
7     logging_steps=10,
```

```

8     eval_strategy="no",
9     save_strategy="epoch",
10    report_to="none"
11 )

```

Finalement, on met en commun notre modèle, nos paramètres d'entraînement, notre dataset tokenisé et notre tokenizer via la classe `Trainer` et on peut lancer l'entraînement avec la méthode `train()`.

3.3. Génération

Après entraînement du modèle, on crée un pipeline de génération:

```

1 generator = pipeline("text-generation", model=model, tokenizer=tokenizer)

```

On peut maintenant utiliser le modèle pour terminer un début de proverbes:

```

1 prompt = "A"
2 results = generator(prompt, max_length=max_length, num_return_sequences=3,
3 do_sample=True, temperature=0.7)
4 for i, result in enumerate(results):
5     print(result['generated_text'])

```

Quelques proverbes obtenus:

- A nice dog is a dog.
- A man who keeps his family safe can never be found.
- A child that has no teeth is a coward.

4. Analyse des résultats

Les données de sortie étant des proverbes, elles sont difficilement évaluable et comparable et sont sujettes à l'appréciation humaine.

Ainsi nous donnerons quelques résultats non triés à titre d'exemple. Puis dans une seconde partie, nous détaillerons une expérience que nous avons réalisé.

4.1. Entraînement sur les proverbes anglais

On utilise ici les deux sources :

- `proverbs_db_only_english.txt`
- `proverbs_digest.txt`

Pour rappel, cette base de données représente 3208 proverbes originaux en anglais.

4.1.1. Modèle de départ facebook/opt-125m

Proverbes obtenus à partir de « A » :

Proverbe	Commentaire
A man's only dead when he eats his own.	Intéressant
A man can make a woman forget her brother's death.	Etrange
A man's heart is full of gold.	Intéressant
A good thing is a good thing.	Effectivement

A woman's dream is to be a widow.	Etrange
A man dies in the wind, a horse dies in the wind	Etrange
A good example of a good example of a good example of a bad example of a bad example of a bad example.	Incohérent
A few days before you start a new one, you will be remembered for a long time.	Incohérent
A little of light can be a good thing	Intéressant

Proverbes obtenus à partir de « Some » :

Proverbe	Commentaire
Some people have a hard time keeping a family man.	Intéressant
Some people are good, some people are good.	Incohérent
Some people are lucky to live in a tree.	Intéressant
Some of them are to be proud of their first friend.	Intéressant
Some times the best is when you are there to be with.	Incohérent
Some people have a disease.	Etrange
Some people are better than others.	Intéressant
Some of us are better than others	Intéressant
Some people are too silly to forget.	Intéressant
Some of those is nothing compared to the sum of them	Etrange/Intéressant

Intéressants 9

Etranges 4

Incohérents 4

On constate que certains résultats sont incohérents ou étrange. Certains proverbes ressemblent plus à des vérités générales et on trouve quelques proverbes intéressants.

4.1.2. Modèle de départ TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T

Les données sélectionnées ne sont pas assez importante pour obtenir des résultats satisfaisants avec ce modèle.

Exemple de sorties pour « This » :

Proverbe	Commentaire
This is the story of an old woman.	
This is the way of the world.	
This is the best.	
This is the way to get your money's worth.	
This is the day.	
This is what you call a new dress.	
This is a story about a little girl.	
This is the day the Lord hath spoken. It is the day that the Lord hath spoken.	
This is what happens when we start out on the wrong path.	
This is my favourite drink.	

4.2. Entraînement sur les proverbes traduits

4.2.1. Modèle de départ `facebook/opt-125m`

On entraîne d'abord sur **20000** proverbes.

4.2.2. Modèle de départ `TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T`

La RAM disponible ne nous a permis que de sélectionner un maximum de 5000 proverbes.

5. Conclusion