ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΜΑΤΙΚΗΣ

Αναφορά 3ης Εργασίας: Μηχανική Μάθηση και Εφαρμογές

Μανούσος Λιναρδάκης, it22064

Μέρος 1: Χρήση του GPT-2:

1. Τρέχοντας το κώδικα τις εκφώνησης (και αλλάζοντάς το λίγο για να φαίνεται το prompt στο notebook με print, και να τρέχει μόνο για 3 prompts), μπορούμε να δούμε τα αποτελέσματα για τα εξής 3 prompts:

Το πρώτο είναι η πρώτη πρόταση από το βιβλίο Wizard of Oz:

```
Dorothy lived in the midst of the great Kansas prairies, with Uncle
Henry, who was a farmer, and Aunt Em, who was the farmer's wife.
```

Η απάντηση του GPT-2 είναι:

```
Dorothy lived in the midst of the great Kansas prairies, with Uncle
Henry, who was a farmer, and Aunt Em, who was the farmer's wife. After
he had made some changes, the Indians began to grow and multiply. The
great land was given to us as "the country that gave birth to the man's
body," and the Indians had to be removed from the lands to be placed in
other lands. At last the Indians were able to take possession of the
land and the land was taken, and we, as a whole, have a right to a
presentation of this same soil.

We do not want to have our presentations to be a part of it, because, as
we know, we have been in that territory many and many years, and the
Indians have a lot more land, and they must give us a presentation
before we can do that again. We have had no right, in any part of this
country ever to have a part whatever in our presentations. No right to
make this land our presentation; we have always had a right to the
presentation, for it is what we have been through
```

παρατηρούμε ότι για αυτό το prompt, το GPT πάραξε ένα κείμενο που (εκτός από το γεγονός ότι δεν ακολουθεί την ιστορία του Wizard of Oz) δεν έχει πολύ “συνοχή”. Για παράδειγμα, το πρόσωπο του κειμένου αλλάζει από “he” (στις πρώτες προτάσεις) σε “we” στο υπόλοιπο κείμενο, αλλάζοντας την μορφή της αφήγησης. Επίσης, χρησιμοποιεί αρκετές φορές την λέξη “presentation” χωρίς να βγάζει ιδιαίτερο νόημα. Παρόλα αυτά το κείμενο είναι αρκετά ρεαλιστικό και μοιάζει με τον τρόπο που θα έγραφε ένας άνθρωπος (αν εξαιρέσουμε την αλλαγή προσώπου). Γενικότερα, το GPT-2 σύνδεσε το αρχικό prompt με το θέμα της “ιδιοκτησίας γης” και συγκεκριμένα με την μετεγκατάσταση των Ινδιάνων, κάτι που είναι ιστορικά σωστό! Το κείμενο συνεχίζει να αναφέρεται σε αυτό το θέμα, και στην επόμενη παράγραφο. Συνεπώς το συγκεκριμένο κείμενο είναι μία αρκετά πετυχημένη παραγωγή κειμένου.

Το δεύτερο prompt που δοκιμάστηκε είναι η γνωστή φράση του Shakespeare:

```
To be or not to be: that is the question.
```

Το μοντέλο έδωσε το παρακάτω αποτέλεσμα:

```
To be or not to be: that is the question.

But to me, this was not about this. This was about the fact -- it was
something that was part of that whole debate. And now here's the thing,
```

```
if you're a feminist or a humanist and you've been doing this long
enough, you'll realize that, I think, the question that you have to ask
when you say "feminism" or "humanism" is whether or not that's what you
really think of as feminism -- and to me, feminism as a concept, as a
concept as it relates to things that, I think, are more or less the same
concept. And I have to say, this is one of those things that's a pretty
simple and easy answer and that's the sort of question that makes me
think in that way. But at the same time, that's just something that I've
been doing and that's the sort of thing that I think is going to make
feminists, and I will say this, if it's
```

Εδώ (εκτός από τι πάλι δεν συνεχίζει τις λέξεις του Shakespeare), μιλάει για το θέμα “feminism” or “humanism”. Αυτή τη φορά η ποιότητα χειροτέρεψε, καθώς κάποιες φορές επαναλαμβάνει τις ίδιες φράσεις (πχ “`and to me, feminism as a concept, as a concept`”) ή γράφει προτάσεις που δεν βγάζουν πολύ νόημα και δεν έχουν τόση σχέση με το υπόλοιπο κείμενο – θα μπορούσαν δηλαδή να είχαν αφαιρεθεί (πχ “`This was about the fact -- it was something that was part of that whole debate.`”). Συνεπώς, το τελικό αποτέλεσμα δεν φαίνεται πολύ σαν ανθρώπινη “γραφή” και δεν μπορεί να θεωρηθεί τόσο πετυχημένη παραγωγή κειμένου.

Τέλος το τρίτο prompt που δοκιμάστηκε είναι η πρώτη πρόταση σε ταινίες Star Wars:

```
In a galaxy far, far away,
```

Το αποτέλεσμα είναι:

```
In a galaxy far, far away, in a galaxy far, away, in a galaxy far, the
world is a place where you can go to the right place in the right time
and place of the wrong time. And in all of that time, if I'm on the
wrong side, what can I do to change it?

That's where we're going. You're going from a time where you're sitting
in a room with a television and you're talking and you're thinking about
things and you're looking for the right thing that has to do with the
season and the people you want to find. The season is already done.
We're at a point where we're able to get things done. So there's no
question about that.

But let's look at this season. What kind of season is there in the
universe where there is this idea that there's an entire universe that's
in a perfect place in which you can have a family without having your
spouse be on the wrong side of something?
```

Εδώ πάλι φαίνεται σαν να μην το έγραψε άνθρωπος καθώς ξανά επαναλαμβάνει φράσεις (όπως φαίνεται και στην αρχή του κειμένου, που επαναλαμβάνει το “in a galaxy far away” σε διάφορα σημεία). Επίσης, κάποιες προτάσεις δεν βγάζουν νόημα (πχ “`to the right place in the right time and place of the wrong time`”) και είναι λανθασμένες συντακτικά (το “in the right time” πρέπει να γίνει “**at** the right time”). Γενικότερα, το κείμενο αυτό δεν έχει πολύ συνοχή και έχει συντακτικά λάθη κάνοντάς το ένα παράδειγμα μη επιτυχημένης παραγωγής κειμένου.

Συνεπώς το μοντέλο τις περισσότερες φορές δεν βγάζει συνεκτικά και “ρεαλιστικά” κείμενα.

2. Αναζητώντας στο διαδίκτυο, βρήκα την ακόλουθη [πηγή] του huggingface που δοκιμάζει διάφορες στρατηγικές αποκωδικοποίησης.

Τα prompts που χρησιμοποιήθηκαν για τα παρακάτω test είναι η πρώτη γραμμή του wizard of oz και η φράση από το star wars “In a galaxy far, far away,”:

```
prompts = ["Dorothy lived in the midst of the great Kansas prairies, with Uncle Henry, who was a farmer, and Aunt Em, who was the farmer's wife.", "In a galaxy far, far away,"]
```

Ακολουθούν σχολιασμοί και αποτελέσματα των στρατηγικών αποκωδικοποίησης που χρησιμοποιήθηκαν:

<u>Αποτελέσματα Greedy Αναζήτηση:</u>

1ο prompt:

```
Dorothy lived in the midst of the great Kansas prairies, with Uncle Henry, who was a farmer, and Aunt Em, who was the farmer's wife. The two of them were very close, and they were very good friends.

The next day, the two of them went to the house of the farmer, and there they met the two of them, and they were very happy. They were very happy, and they were very happy.

The next day, the two of them went to the house of the farmer, and there they met the two of them, and they were very happy. They were very happy, and they were very happy.

The next day, the two of them went to the house of the farmer, and there they met the two of them, and they were very happy. They were very happy, and they were very happy.

The next day, the two of them went to the house of the farmer, and there they met the two of them, and they were very happy. They were very happy, and they were very happy.

The next day, the two of
```

Το greedy search επιλέγει τη λέξη με τη μεγαλύτερη πιθανότητα ως επόμενη λέξη. Στο παράγωμενο κείμενο παρατηρούμε ότι το μοντέλο επαναλαμβάνεται πολλές φορές, κάτι που είναι λογικό καθώς κάθε φορά επιλέγουμε τη λέξη με την μέγιστη πιθανότητα (κάτι που μπορεί να οδηγήσει σε επανάληψη sequence).

Αυτή η επανάληψη γίνεται ακόμα χειρότερη στο 2ο prompt “In a galaxy far, far away,”:

```
In a galaxy far, far away, the galaxy is a vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast, vast
```

εδώ παρατηρούμε ότι έχει επαναλάβει πολλές φορές τη λέξη “vast” και συνεχίζει να το κάνει αυτό μέχρι το max_tokens.
Συνεπώς η άπληστη αναζήτηση επαναλαμβάνει πολλές φορές τις ίδιες λέξεις!

Αποτελέσματα Beam Αναζήτησης:
1ο prompt:

```
Dorothy lived in the midst of the great Kansas prairies, with Uncle
Henry, who was a farmer, and Aunt Em, who was the farmer's wife.

In the spring of 1848, the family moved to a small town on the western
side of the Missouri River, near the Missouri River.

In the spring of 1848, the family moved to a small town on the western
side of the Missouri River, near the Missouri River.

In the spring of 1848, the family moved to a small town on the western
side of the Missouri River, near the Missouri River.

In the spring of 1848, the family moved to a small town on the western
side of the Missouri River, near the Missouri River.

In the spring of 1848, the family moved to a small town on the western
side of the Missouri River, near the Missouri River.

In the spring of 1848, the family moved to a small town on the western
side of the Missouri River, near the Missouri River.

In the spring of 1848, the family moved to a small town on the western
side
```

Εδώ πάλι βλέπουμε ότι επαναλαμβάνει φράσεις!

Ας δούμε τα αποτελέσματα του 2ου prompt:

```
In a galaxy far, far away, there is a galaxy far, far, far away.

In a galaxy far, far away, there is a galaxy far, far, far away.

In a galaxy far, far away, there is a galaxy far, far, far away.

In a galaxy far, far away, there is a galaxy far, far, far away.

In a galaxy far, far away, there is a galaxy far, far, far away.

In a galaxy far, far away, there is a galaxy far, far, far away.
```

αυτή τη φορά υπάρχει πάλι επανάληψη της ίδιας φράσης, αλλά όχι μόνο μίας λέξης (όπως έγινε πριν στο grid search).

Συνεπώς το beam search είναι λίγο καλύτερο από το grid search καθώς (στο grid search) στο 2ο prompt είχαμε επανάληψη μόνο μίας λέξης (κάτι που δεν μοίαζει σαν να το έγραψε άνθρωπος). Τώρα (με το beam search) επαναλαμβάνονται προτάσεις οι οποίες και πάλι δεν προσφέρουν κάτι καινούργιο (στο κείμενο) αλλά τουλάχιστον δεν είναι μόνο μία λέξη που επαναλαμβάνεται συνεχώς, κάνοντας έτσι το κείμενο λίγο πιο ρεαλιστικό.

Χρήση Beam Αναζήτησης + no_repeat_ngram_size:
Χρησιμοποιώντας το no_repeat_ngram_size, λέμε στο μοντέλο να μην επαναλαμβάνει τις ίδιες ακολουθίες στην παραγόμενη έξοδο. Τα αποτελέσματα είναι σαφώς καλύτερα από το απλό Beam Search, καθώς "λύνει" το πρόβλημα που παρατηρήθηκε και πριν με την επανάληψη προτάσεων. Για το 1ο prompt το αποτέλεσμα είναι:

```
Dorothy lived in the midst of the great Kansas prairies, with Uncle
Henry, who was a farmer, and Aunt Em, who was the farmer's wife.

In the spring of 1848, the family moved to a small town on the outskirts
of K
```

Για το 2ο prompt:

```
In a galaxy far, far away, there is no such thing as a safe place to be.

It's a place where you can be anything you want. You can go anywhere,
anywhere. There's no place like home. It's the only place you'll ever be
able to go, and that's why I'm here. I want you to know that I love you.
And I know you're going to love me, too.
```

τα παραπάνω κείμενα που παράχθηκαν μοιάζουν περισσότερο σαν να τα έγραψε άνθρωπος, αλλά δεν είναι πολύ μεγάλα σε μέγεθος, λόγω της "απαγόρευσης" όρων από το no_repeat_ngram_size.

Χρήση top-k:
Τα prompts, αποτελέσματα και ο σχολιασμός αυτής της στρατηγικής βρίσκονται στο πρώτο ερώτημα (καθώς αυτή η μέθοδος χρησιμοποιήθηκε για την απάντηση του 1ου). Το συμπέρασμα της στρατηγικής αυτής είναι ότι το μοντέλο τις περισσότερες φορές δεν βγάζει συνεκτικά και "ρεαλιστικά" κείμενα. Παρόλα αυτά βγάζει καλύτερα αποτελέσματα από το beam & greedy search καθώς τα παραγόμενα κείμενα είναι μεγάλα σε μέγεθος και δεν επαναλαμβάνουν συχνά τις ίδιες φράσεις/λέξεις.

Χρήση top-p:
1ο prompt:

```
Dorothy lived in the midst of the great Kansas prairies, with Uncle
Henry, who was a farmer, and Aunt Em, who was the farmer's wife. They
married in 1479, to whom they have continued to live ever since. The two
children were: George a son of Samuel, who died of heart disease in
1477; and Alice a daughter of Samuel, who died of disease in 1484. In
consequence, they live well, have a common life, have a good home, and
live in peace.
```

```
The eldest daughter, a daughter of Edward, died in 1492; the son is
buried in the old Church of England in Leicester, which she lived in
until the mid-1630s; his widow, Esther, died in 1650.

In 1744, Charles Dandridge became a member of the Continental
Parliament. The Continental Parliament was elected in 1745. It was
formed with the support of a majority in Congress; the president was
John Adams, who was appointed by George Washington, a member of the
Continental Parliament in 1776, and John Jay, a member of Congress until
1801. When Jay
```

παρατηρούμε ότι το κείμενο που παράχθηκε χρησιμοποιεί ποικίλες λέξεις κάνοντας το κείμενο να μοιάζει σαν να το έγραψε άνθρωπος. Επίσης, το κείμενο μένει κοντά σε ένα από τα “θέματα” του prompt, το οποίο είναι η οικογένεια (καθώς στο prompt περιγράφουμε πως ζει η Dorothy μαζί με τον θείο και θεία της). Βέβαια, το κείμενο αν το διαβάσουμε βλέπουμε ότι δεν βγάζει κάποιες φορές νόημα (πχ γράφει ότι τα δύο παιδιά George and Alice πέθαναν από πρόβλημα καρδιάς και στην συνέχεια λέει ότι “κατά συνέπεια” ζούνε ωραία).

2ο prompt:

```
In a galaxy far, far away, one of the brightest galaxies that ever
exists, lies hidden beneath the sky.

In a galaxy far, far away, one of the brightest galaxies that ever
exists, lies hidden beneath the sky. The universe looks really, really
dark. When NASA's Keck Observatory and the SETI program found the
galaxy, it became a mystery and was left out of the search. The Keck
observatory, which includes Kota and Chico, in California, also turned
up a giant, glowing gas giant called a dwarf galaxy known as M1. But a
team of scientists from UCLA, New Mexico and California State University
has found something even stranger: They can observe it from inside a
star.

The Keck Telescope, now under construction at the European Southern
Observatory in Chile, is the largest telescope in the world to study the
Keck Cluster. The telescope has a broad field of view of at least three
million light-years, in contrast to the thousands or even millions that
are visible over
```

σε αυτό το παραγόμενο κείμενο βλέπουμε και πάλι ότι το top-p είναι πιο κοντά στο θέμα του αρχικού prompt, και συγκεκριμένα μιλάει για το διάστημα. Επιπρόσθετα, το κείμενο αυτή τη φορά είναι πιο συνεκτικό και “σωστό" από τι στο προηγούμενο prompt και για αυτό θυμίζει περισσότερο την ανθρώπινη γραφή.

Συνεπώς το top-p είναι λίγο καλύτερο από το top-k καθώς μένει στο ίδιο “θέμα” με το prompt που του έχουμε δώσει ενώ ταυτόχρονα φτιάχνει αρκετά ρεαλιστικά κείμενα.

Χρήση top-k & top-p:

1ο prompt:

```
Dorothy lived in the midst of the great Kansas prairies, with Uncle
Henry, who was a farmer, and Aunt Em, who was the farmer's wife. On the
same day, she and her father took a boat to the town, where they were
found by a man from the same company. The young woman and the young man
who accompanied them were killed in the collision with a herd of cattle,
and they were buried at the cemetery. They were buried in a large grave
at Woburn, a little north of Wichita, and the remains of Mrs. Ortega are
said to be there. They were buried in a large mound on a hillside near
the town. There was an old brick church built in 1837. The church of
Mrs. Ortega was found in the vicinity of the burial place of her mother.
Another church in which the family is said to have lived is the
Methodist Church, at Kuehlman, Kansas City. The oldest surviving church
is located in the churchyard of the church where they lived. It is said
that they were baptized at the stake and that they were buried near the
spot where they died
```

παρατηρούμε ότι το παραγόμενο κείμενο είναι πιο ρεαλιστικό από πριν (που χρησιμοποιήσαμε μόνο το top-p). Και πάλι μένει κοντά στο αρχικό θέμα του prompt, χωρίς όμως να υπάρχει κάποιο σημείο που να μην βγάζει ιδιαίτερο νόημα ή να φεύγει πολύ από το θέμα.

2ο prompt:

```
In a galaxy far, far away, and from a galaxy far, far away, is a
strange, strange world.

The galaxy is a vast, unspoiled, multiverse, filled with many galaxies,
and many planets, many universes, many universes, many universes, and
many universes. Every universe is inhabited by several intelligent
beings, and even many of them have their own unique characteristics, but
the main star in the universe is the largest galaxy in the entire known
universe and is the most powerful. And this is not all.

This planet is the "Gigantic Moon" and all the stars that inhabit it are
of this world. Every star is one massive, massive star. Every star is
made up of billions of galaxies. Each galaxy is filled with trillions of
stars that are all the same size. Each galaxy is home to millions of
galaxies. Every galaxy is a huge star in which there are many galaxies
and billions of galaxies and millions of planets, billions of planets,
and hundreds of planets. So
```

Ομοίως και εδώ παρατηρούμε ότι πάλι μένει στο θέμα του prompt και μιλάει για πλανήτες. Συνεπώς ο συνδυασμός top-k & top-p έδωσε τα καλύτερα αποτελέσματα μέχρι στιγμής (όπου καλύτερα σημαίνει πιο κοντά στην ανθρώπινη γραφή). Στο Μέρος 2 χρησιμοποιώ το συνδυασμό αυτό για να δοκιμάσω το ζητούμενο μοντέλο.

Μέρος 2: Προσαρμογή μοντέλου

Παρατήρηση: Το μοντέλο το εκπαίδευσα για συνολικά 5 εποχές όπως γράφει στον κώδικα της εκφώνησης. Είναι αξιοσημείωτο ότι άλλαξα λίγο τον αρχικό κώδικα έτσι ώστε αν υπάρχει ήδη αποθηκευμένο το μοντέλο (στο κατάλογο αποθήκευσης) να συνεχίζει να το εκπαιδεύει (κάνοντάς το load από τον κατάλογο πρώτα). Οπότε, αρχικά έτρεξα το colab για 1 εποχή, ύστερα έκανα save το μοντέλο και σε άλλο χρόνο το έκανα load από τον κατάλογο και το έτρεξα άλλες 2 + 2 εποχές. Το τελικό μοντέλο το έκανα save στον κατάλογο model_save.

1. Το prompt που δοκιμάστηκε είναι η φράση του Shakespeare:

```
To be or not to be: that is the question.
```

Το μοντέλο έδωσε το παρακάτω αποτέλεσμα:

```
To be or not to be: that is the question. What good is a question of
precedure if you start down the south passage and speak of the Rebellion?
I know the precedure of the law. I feel confident we can overcome it
I must speak with the Jei Council immediately, Your Honor. The situation has
become more complicated.
Ani, come on.
Da queen's a bein grossly nice, mesa tinks. Pitty hot.
the Republic is not what it once was. The Senate is full of greedy,
squabbling delegates who are only looking out for themselves and their home
sytems. There is no interest in the common good no civility, only politics
its disgusting. I must be frank, Your Majesty, there is little chance the
Senate will act on the invasion.
Chancellor Valorum seems to think there is hope.
If I may say so, Your Majesty, the Chancellor has little real power he is
mired down by baseless accusations of corruption. A manufactured scandal
surrounds him
```

Παρατηρούμε ότι χρησιμοποιώντας το prompt αυτό, το μοντέλο συνέχισε το κείμενο με βάση τους υπότιτλους star wars. Επίσης, η λέξη “question” του prompt οδήγησε στην επόμενη πρόταση του παραγόμενου κειμένου να περιέχει τη λέξη “question”. Παρατηρούμε λοιπόν ότι το μοντέλο προσπαθεί να παράξει κείμενο που να είναι όσο πιο κοντά γίνεται στο prompt (με βάση μοτίβα διαλόγων των ταινιών star wars, καθώς αυτά μαθαίνει).

2. Prompts που χρησιμοποιήθηκαν:

```
"the dark side", "the force"
```

Χρησιμοποιώντας το prompt “the dark side” 2 φορές παρατηρούμε ότι η έξοδος – τα κείμενα που παράχθηκαν είναι τα ίδια! Συγκεκριμένα και τις 2 φορές που χρησιμοποιήθηκε το συγκεκριμένο prompt παράχθηκε το παρακάτω κείμενο:

```
the dark side of the Force is a pathway to many abilities some consider
to be unnatural.
What happened to him?
He became so powerful... the only thing he was afraid of was losing his
power, which eventually, of course, he did. Unfortunately, he taught his
apprentice everything he knew, then his apprentice killed him in his
sleep. (smiles) Plagueis never saw it coming. It's ironic he could save
```

```
others from death, but not himself.
Is it possible to learn this power?
Not from a Jedi.
(holo) Palpatine thinks General Grievous is on Utapau. We have had no reports of this from our agents.
(holo) How could the Chancellor have come by this information and we know nothing about it? We have had contact with Baron Papanoida and he said no one was there.
```

Το κείμενο που ακολουθεί το “the dark side” είναι ακριβώς το ίδιο με αυτό στους υπότιτλους star_wars.txt.

Χρησιμοποιώντας το prompt “the force” 2 φορές πήραμε διαφορετικές εξόδους. Το πρώτο κείμενο που παράχθηκε είναι:

```
the force is with us, my Master.
Welcome home, Lord Tyranus. You have done well.
I bring you good news, my Lord. The war has begun.
Excellent. (smiling) Everything is going as planned.
Where is your apprentice?
On his way back to Naboo. He is escorting Senator Amidala home.
(continuing) I must admit without the clones, it would not have been a victory.
Victory? Victory, you say?
```

Η “παράγραφος” αυτή είναι επίσης ίδια από τους υπότιτλους!

Το επόμενο κείμενο που παράχθηκε είναι:

```
the force grows dark, Anakin, and we are all affected by it. Be wary of your feelings.
Anakin, this afternoon the Senate is going to call on me to take direct control of the Jedi Council.
The Jedi will no longer report to the Senate?
They will report to me... personally. The Senate is too unfocused to conduct a war. This will bring a quick end to things.
I agree, but the Jedi Council may not see it that way.
There are times when we must all endure adjustments to the constitution in the name of security.
With all due respect, sir, the Council is in no mood for more constitutional amendments.
Thank you, my friend, but in this case I have no choice... this war must be won.
```

που επίσης είναι το ίδιο κείμενο με τους υπότιτλους.

Συμπερασματικά, το μοντέλο παράγει κείμενο που είναι πολύ “πιστό” στους υπότιτλους των ταινιών star wars. Επίσης, όπως φαίνεται και από το prompt “the dark side” μπορεί να υπάρξει περίπτωση που δημιουργεί το ίδιο κείμενο (για το ίδιο prompt).

3. Τρέχοντας τον κώδικα, βγάζει:

```
The GPT-2 model has 148 different named parameters.

==== Embedding Layer ====

transformer.wte.weight                                (50259, 768)
transformer.wpe.weight                                 (1024, 768)

==== First Transformer ====

transformer.h.0.ln_1.weight                                (768,)
transformer.h.0.ln_1.bias                                  (768,)
transformer.h.0.attn.c_attn.weight                     (768, 2304)
transformer.h.0.attn.c_attn.bias                          (2304,)
transformer.h.0.attn.c_proj.weight                     (768, 768)
transformer.h.0.attn.c_proj.bias                           (768,)
transformer.h.0.ln_2.weight                                (768,)
transformer.h.0.ln_2.bias                                  (768,)
transformer.h.0.mlp.c_fc.weight                        (768, 3072)
transformer.h.0.mlp.c_fc.bias                             (3072,)
transformer.h.0.mlp.c_proj.weight                      (3072, 768)
transformer.h.0.mlp.c_proj.bias                            (768,)

==== Output Layer ====

transformer.ln_f.weight                                    (768,)
transformer.ln_f.bias                                      (768,)
```

Αρχικά, βλέπουμε ότι το GPT-2 έχει 148 διαφορετικές “named” παραμέτρους. Βλέποντας τις παραμέτρους αυτές, μπορούμε να συμπεράνουμε τα εξής για το πλήθος τους:

Για το Embedding Layer (μετασχηματίζει το κείμενο σε αριθμητικά διανύσματα για να μπορέσει να τα επεξεργαστεί το μοντέλο):

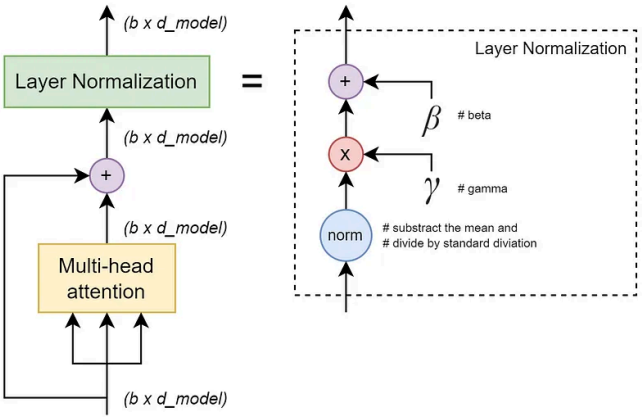
- transformer.wte.weight = weight matrix για τα token embeddings. Δείχνει ότι υπάρχουν 50.259 ξεχωριστά embeddings και το κάθε (embedding) αναπαρίσταται από ένα διάνυσμα με **768** διαστάσεις.
- transformer.wpe.weight = weight matrix για τα positional embeddings. Δείχνει ότι υπάρχουν 1024 θέσεις, η καθεμία αναπαρίσταται από διάνυσμα 768 διαστάσεων.

Το dmodel είναι μια παράμετρος που αναφέρεται στον αριθμό των διαστάσεων των embeddings του μοντέλου. Επομένως, dmodel = 768.

Για το First Transformer:

- ln_1: (γενικά ln = layer normalization) υπάρχουν 768 (=dmodel) παράμετροι τόσο για το weight όσο και για το bias.

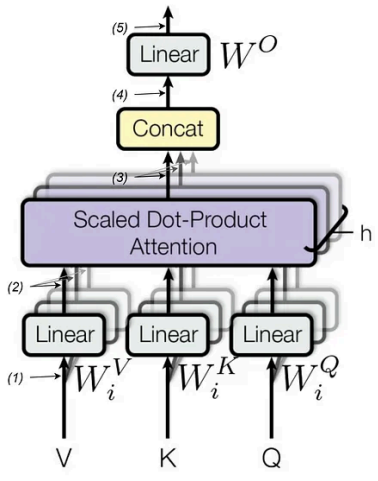
Εξήγηση:
Το layer-normalization περιγράφεται από το ακόλουθο σχήμα:



Στο μοντέλο GPT-2, το layer-normalization περιλαμβάνει τον υπολογισμό του μέσου όρου και της τυπικής απόκλισης στις διαστάσεις της εισόδου (dmodel=768). Στη συνέχεια, τα αποτελέσματα αυτά χρησιμοποιούνται για την κανονικοποίηση των εισόδων πριν από την εφαρμογή των παραμέτρων scaling (γ) και shifting (β). Αυτές οι παράμετροι επιτρέπουν στο μοντέλο να προσαρμόζει τις κανονικοποιημένες εξόδους, παρέχοντας ευελιξία στην αναπαράσταση. Επομένως, τα διανύσματα bias (β) και weight (γ) έχουν την ίδια διάσταση με τις ενσωματώσεις εισόδου (768) για να διασφαλιστεί η συνέπεια στη διαδικασία κανονικοποίησης.

- attn.c_attn: Οι παράμετροι του μηχανισμού προσοχής. Το weight matrix έχει διαστάσεις (768, 2304) και το bias έχει μέγεθος 2304.
- attn.c_proj: Η “έξοδος” του μηχανισμού προσοχής. Το weight matrix απέκτησε διαστάσεις (768, 768) και το bias έχει μέγεθος 768.

Εξήγηση:
Η έξοδος έχει αυτές τις παραμέτρους – διαστάσεις καθώς το multiple head attention (MHA) μοιάζει σχηματικά ως εξής:



$$W^O.shape = (d_{model}, d_{model})$$
$$W_i^Q.shape = (d_{model}, d_{qkv})$$
$$W_i^K.shape = (d_{model}, d_{qkv})$$
$$W_i^V.shape = (d_{model}, d_{qkv})$$
$$d_{qkv} = d_k = d_v = d_{model}/n_{heads} = d_{model}/h$$

$1)(b, d_{model})$
$2)(b, d_{qkv})$
$3)(b, d_{qkv})$
$4)(b, d_{model})$
$5)(b, d_{model})$

Η έξοδος έχει τις διαστάσεις του W^O που είναι (dmodel, dmodel)=(768, 768).

Επίσης, οι παράμετροι του MHA – που περιγράφονται στο attn.c_attn – έχουν διαστάσεις (dmodel, 3*dmodel) = (768, 3*768) = (768, 2304). Ο πολλαπλασιασμός με το 3 γίνεται γιατί βλέπουμε ότι υπάρχουν 3 “είσοδοι” στο κάτω μέρος (του σχήματος).

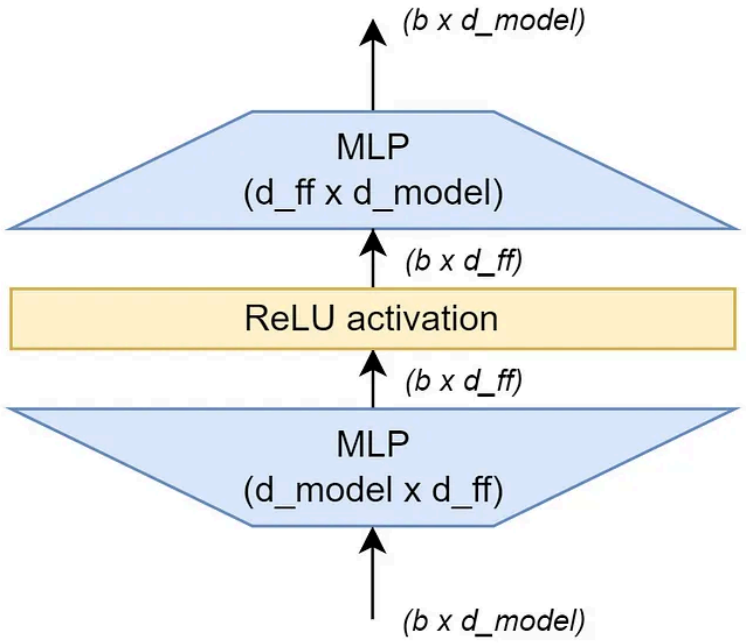
- ln_2: παράμετροι του επόμενου layer normalization, όπως το ln_1, με 768 παραμέτρους το weight και bias (παρόμοια εξήγηση με ln_1).
- mlp.c_fc: Οι παράμετροι του feed-forward. Το weight matrix έχει μέγεθος (768, 3072) και το bias έχει 3072, όπου το 3072 είναι το dff = 4*d_model.
- mlp.c_proj: Η έξοδος του feed-forward. Το weight matrix έχει μέγεθος (3072, 768) και το bias έχει 768.

Εξήγηση feed-forward:
Ο τύπος που χρησιμοποιείται στο paper (“Attention is all you need”) για το feed-forward είναι:

$$\mathrm{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (2)$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer. Another way of describing this is as two convolutions with kernel size 1. The dimensionality of input and output is $d_{\text{model}} = 512$, and the inner-layer has dimensionality $d_{ff} = 2048$.

Σχηματικά αυτό γράφεται:



όπου d_ff = 4*d_model. Μπορούμε να δούμε ότι στην αρχή η είσοδος μπαίνει – όπως περιγράφεται στο c_fc – με διαστάσεις (d_model, d_ff) = (768, 3072) και μετά περνάει από ReLU. Έπειτα, η έξοδος βγαίνει – όπως επισημαίνεται στο c_proj – με διαστάσεις (d_ff, d_model) = (3072, 768).

Για το output layer: δείχνει τις παραμέτρους του τελικού layer normalization (ln_f), που είναι 768 παράμετροι για το weight και bias (όπως και στα προηγούμενα layer normalization – η εξήγηση για το πως προέκυψαν αυτές οι παράμετροι είναι όμοια με αυτή της ln_1).

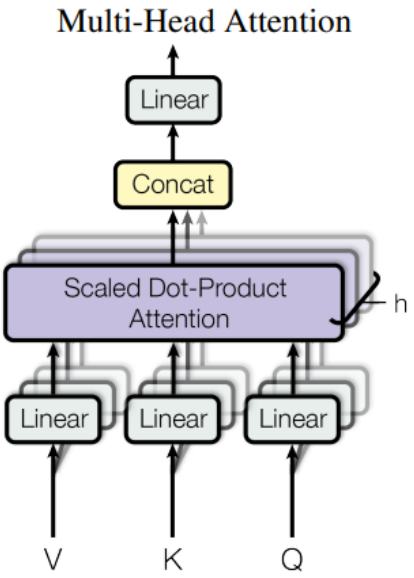
Τα σχήματα που χρησιμοποιήθηκαν για την απάντηση αυτού του ερωτήματος μπορούν να βρεθούν στις ακόλουθες πηγές [1], [2], [3].

Bonus ερώτημα

Μελετώντας προσεκτικά την κλάση MultiHeadAttention, συμπεραίνουμε ότι δεν υπάρχει λάθος στο κώδικα.

Για να αιτιολήσουμε αυτή τη πρόταση, χρειάζεται αρχικά να μελετήσουμε το MultiHead Attention (MHA) του paper "Attention is all you need".

Σύμφωνα με το paper, το multi-head attention περιγράφεται με το ακόλουθο σχήμα:



Συνήθης shape των V, K, Q είναι το (seq_length, d_model), όπου το seq_length αντιπροσωπεύει το μήκος κάθε ακολουθίας και το d_model τη διάσταση του χώρου χαρακτηριστικών. Στο paper έχουμε επίσης και το d_k για το οποίο ισχύει d_k = d_model / num_heads. Ακολουθεί το αντίστοιχο απόσπασμα:

> In this work we employ $h = 8$ parallel attention layers, or heads. For each of these we use $d_k = d_v = d_{\text{model}}/h = 64$. Due to the reduced dimension of each head, the total computational cost is similar to that of single-head attention with full dimensionality.

Για την υλοποίηση του MHA, βρήκαν ότι είναι χρήσιμο να προβάλλουν h φορές (όπου h = αριθμός heads) με διαφορετικό γραμμικό τρόπο τα Q, K, V σε dimensions d_k, d_k, d_v (=d_k) αντίστοιχα. Αντίστοιχη αναφορά από το paper:

> **3.2.2 Multi-Head Attention**
>
> Instead of performing a single attention function with d_{model}-dimensional keys, values and queries, we found it beneficial to linearly project the queries, keys and values h times with different, learned linear projections to d_k, d_k and d_v dimensions, respectively. On each of these projected versions of queries, keys and values we then perform the attention function in parallel, yielding d_v-dimensional

Αυτό φαίνεται και από το κάτω μέρος του σχήματος του MHA (πριν το scaled dot-product attention), όπου τα Q, K, V εισέρχονται (και κατά συνέπεια προβάλλονται) σε h διαφορετικά γραμμικά επίπεδα. Έτσι, το i-οστό scaled dot-product attention (σύμφωνα με το paper) είναι:

$$\text{head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Όπως βλέπουμε και από το απόσπασμα αυτό, αυτά που εισέρχονται στο Attention είναι τα QW_i^Q, KW_i^K, VW_i^V, όπου το καθένα από αυτά έχει shape (seq_length, d_k), καθώς το κάθε Wi^X έχει shape (d_model, d_k), οπότε αν το πολλαπλασιάσουμε με το Q (ή το K, ή το V), προκύπτει ότι το γινόμενο έχει shape (seq_length, d_model) x (d_model, d_k) ⇒ (seq_length, d_k). Ακολουθεί και η αντίστοιχη παραπομπή για τα shapes των Wi^X.

> Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

Είναι αξιοσημείωτο ότι ο μηχανισμός προσοχής εφαρμόζεται παράλληλα σε καθένα από αυτά τα projected versions των Q, K, V, όπως αναφέρεται και εδώ:

3.2.2 Multi-Head Attention

> Instead of performing a single attention function with d_{model}-dimensional keys, values and queries, we found it beneficial to linearly project the queries, keys and values h times with different, learned linear projections to d_k, d_k and d_v dimensions, respectively. On each of these projected versions of queries, keys and values we then perform the attention function in parallel, yielding d_v-dimensional output values. These are concatenated and once again projected, resulting in the final values, as depicted in Figure 2.

Αφού υπολογιστούν όλα τα $head_i$, τα κάνουμε concat και έπειτα τα πολλαπλασιάζουμε με το τελικό linear W^O που έχει shape (h*d_v, d_model) = (d_model, d_model). Ακολουθεί και η αντίστοιχη παραπομπή από το paper:

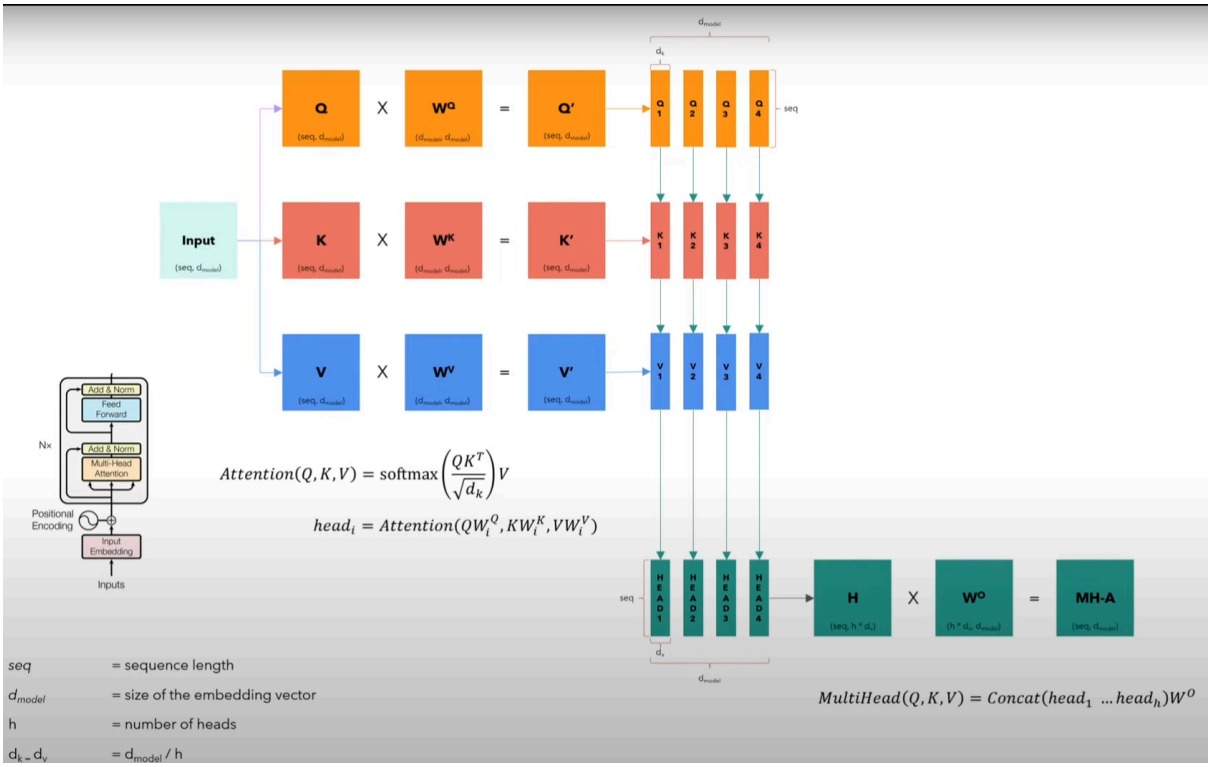
$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$

οπότε όλη η διαδικασία του MHA, με βάση το paper είναι:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_\text{h})W^O$$
$$\text{where head}_\text{i} = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

> Where the projections are parameter matrices $W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$, $W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$ and $W^O \in \mathbb{R}^{hd_v \times d_{\text{model}}}$.

Ας συγκρίνουμε λοιπόν την διαδικασία αυτή με το κώδικα του datacamp. Αρχικά, ολόκληρος ο κώδικας μπορεί να περιγραφεί από το ακόλουθο σχήμα:



Σχήμα [1] (image source – youtube)

Βλέπουμε ότι έχουμε τις εισόδους Q, K, V με shapes (seq_length, d_model) – όπως επισημάναμε και όταν αναλύσαμε το paper. Ύστερα πολλαπλασιάζουμε τα Q, K, V με τα γραμμικά matrices W^Q, W^K, W^V που έχουν διαστάσεις (d_model, d_model). Αυτό έχει ως αποτέλεσμα την παραγωγή matrices Q', K', V' που έχουν διαστάσεις (seq_length, d_model), αφού (seq_length, d_model) x (d_model, d_model) ⇒ (seq_length, d_model). Τα Q', K', V' τα χωρίζουμε (με την split_heads στο κώδικα) σε διαφορετικά γραμμικά σύνολα με διαστάσεις (seq_length, d_k) τόσες φορές όσος και ο αριθμός των heads (όπως δηλαδή αναφέραμε και στην ανάλυση του paper). Έτσι, δημιουργούμε τις εισόδους του attention – QW_i^Q, KW_i^K, VW_i^V του paper (οι οποίες όπως είχαμε υπολογίσει και προηγουμένως έχουν shape (seq_length, d_k) μετά από τη split_heads). 'Επειτα χρησιμοποιώντας ως εισόδους στο Attention τα QW_i^Q, KW_i^K, VW_i^V, υπολογίζουμε το attention στο κάθε $head_i$ παράλληλα όπως γίνεται στο paper. Τα αποτελέσματα του attention που προκύπτουν (τα i heads που είναι με πράσινο χρώμα στο σχήμα) τα κάνουμε concat και ύστερα πολλαπλασιάζουμε με W^O για να βρουμε το MHA.

Συνεπώς, το σχήμα (και κατά συνέπεια και ο κώδικας) ακολουθεί σωστά τη "γενική" διαδικασία του MHA του paper όπως το αναλύσαμε.

Πιο συγκεκριμένα, στο constructor (του κώδικα του datacamp):

```
class MultiHeadAttention(nn.Module):
    def __init__(self, d_model, num_heads):
        super(MultiHeadAttention, self).__init__()
        # Initialize dimensions
        self.d_model = d_model # Model's dimension
        self.num_heads = num_heads # Number of attention heads
        self.d_k = d_model // num_heads # Dimension of each head's key,
```

```
query, and value
        # Linear layers for transforming inputs
        self.W_q = nn.Linear(d_model, d_model) # Query transformation
        self.W_k = nn.Linear(d_model, d_model) # Key transformation
        self.W_v = nn.Linear(d_model, d_model) # Value transformation
        self.W_o = nn.Linear(d_model, d_model) # Output transformation
```

παρατηρούμε ότι γίνονται οι υπολογισμοί των διαστάσεων d_model και d_k, όπου d_k = d_model / num_heads, όπως γίνεται στο paper (η πράξη // εξασφαλίζει ότι το αποτέλεσμα θα είναι πάντα ακέραιος, κάτι που το θέλουμε για τον υπολογισμό της διάστασης d_k). Επίσης, όπως είδαμε και από το σχήμα, αρχικοποιούμε τα W_q, W_k, W_v και W_o με διαστάσεις (d_model, d_model).

Ας δούμε τώρα τη forward του κώδικα:

```
    def forward(self, Q, K, V, mask=None):
        # Apply linear transformations and split heads
        Q = self.split_heads(self.W_q(Q))
        K = self.split_heads(self.W_k(K))
        V = self.split_heads(self.W_v(V))
        # Perform scaled dot-product attention
        attn_output = self.scaled_dot_product_attention(Q, K, V, mask)
        # Combine heads and apply output transformation
        output = self.W_o(self.combine_heads(attn_output))
        return output
```

από εδώ βλέπουμε ότι αρχικά “πολλαπλασιάζει” το Q, K, V με τα γραμμικά matrices W_q, W_k και W_v αντίστοιχα (όπως στο σχήμα). Έτσι, προκύπτουν τα Q’, K’, V’ του σχήματος [1]. Έπειτα χρησιμοποιεί τη split_heads έτσι ώστε να παράξει τις εισόδους του attention QW_i^Q, KW_i^K, VW_i^V.

Ακολουθεί ο κώδικας του split_heads:

```
def split_heads(self, x):
     # Reshape the input to have num_heads for multi-head attention
     batch_size, seq_length, d_model = x.size()
     return x.view(batch_size, seq_length, self.num_heads,
self.d_k).transpose(1, 2)
```

Η split_heads επιστρέφει tensor με shape (batch_size, num_heads, seq_length, d_k), κατί που εν τέλει αντιπροσωπεύει σωστά τις εισόδους QW_i^Q, KW_i^K, VW_i^V που θέλουμε να βάλουμε ύστερα στο Attention, καθώς η κάθε είσοδος έχει shape (seq_length, d_k) όπως υπολογίσαμε και προηγούμενως, και επίσης είναι “χωρισμένη” σε num_heads. Δηλαδή η split_heads τελικά δημιουργεί τα κατάλληλα QW_i^Q, KW_i^K, VW_i^V που μπαίνουν ύστερα ως είσοδος στο attention για την παραγωγή attention του κάθε $head_i$.

Ύστερα τα head attentions γίνονται combine (concat) στη combine_heads:

```
    def combine_heads(self, x):
        # Combine the multiple heads back to original shape
```

```
        batch_size, _, seq_length, d_k = x.size()
        return x.transpose(1, 2).contiguous().view(batch_size, seq_length,
self.d_model)
```

και γυρίζουν ύστερα σε διάσταση d_model έτσι ώστε να τα περάσουμε από το τελικό Linear W_o.

Όσο για το scale_dot_product attention του κώδικα, είναι σωστά υλοποιημένο και ακολουθεί τις οδηγίες του paper. Επίσης, οι υπολογισμοί των attention γίνεται παράλληλα (όπως αναφέρεται στο paper) καθώς το input shape των “Q, K, V” στο scale_dot_product function είναι της μορφής [batch_size, num_heads, seq_length, d_k] δηλαδή έχουμε αρχικοποιήσει το num_heads που εξασφαλίζει την “παραλληλία” των υπολογισμών.

Συνεπώς, ο κώδικας είναι σωστά υλοποιημένος. Ακολουθούν και άλλες πηγές που έχουν πολύ παρόμοια υλοποίηση:

- https://github.com/hkproj/pytorch-transformer/blob/main/model.py#L83
- https://nlp.seas.harvard.edu/annotated-transformer/
- https://d2l.ai/chapter_attention-mechanisms-and-transformers/multihead-attention.html

Ένα “λάθος” που παρατηρήθηκε “εκτός” του MHA είναι στη forward του Transformer class χρησιμοποιείται masking για την παραγωγή της εξόδου στον encoder και ξανά ύστερα όταν η έξοδος του encoder χρησιμοποιείται στο decoder. Αυτό δεν συμβαδίζει με το “τυπικό” μοντέλο του Transformer καθώς ο encoder (στο τυπικό μοντέλο) μπορεί να παρακολουθεί όλες τις θέσεις της εισόδου, για την κατανόηση του context. Για αυτό, αφαίρεσα από τον κώδικα το masking της εισόδου/εξόδου του encoder. Η καινούργια forward μοιάζει ως εξής:

```
def forward(self, src, tgt):
    src_mask, tgt_mask = self.generate_mask(src, tgt)
    src_embedded =
self.dropout(self.positional_encoding(self.encoder_embedding(src)))
    tgt_embedded =
self.dropout(self.positional_encoding(self.decoder_embedding(tgt)))

    enc_output = src_embedded
    for enc_layer in self.encoder_layers:
        # Does not apply mask
        enc_output = enc_layer(enc_output, None)

    dec_output = tgt_embedded
    for dec_layer in self.decoder_layers:
        dec_output = dec_layer(dec_output, enc_output, None, tgt_mask)

    output = self.fc(dec_output)
    return output
```

Τυπική Αρχιτεκτονική Transformer: