Traductor Espanyol-Anglès basat en Xarxes Neuronals Recurrents

Jordi Planagumà Sánchez
Màster en Enginyeria Informàtica
Intel·ligència artificial

Samir Kanaan Izquierdo
Carles Ventura Royo

28 de Desembre de 2016

FITXA DEL TREBALL FINAL

Títol del treball:	*Traductor Espanyol-Anglès basat en xarxes neuronals recurrents*
Nom de l'autor:	*Jordi Planagumà Sánchez*
Nom del consultor/a:	*Samir Kanaan Izquierdo*
Nom del PRA:	*Carles Ventura Royo*
Data de lliurament (mm/aaaa):	*12/2016*
Titulació o programa:	*Màster en Enginyeria Informàtica*
Àrea del Treball Final:	*Intel·ligència artificial*
Idioma del treball:	*Català*
Paraules clau	*Traductor, DeepLearning, XNR*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

Donades les noves tendències i els avanços que s'han fet en el camp de la intel·ligència artificial, es vol adquirir experiència realitzant les tasques de investigació pertinents per adquirir un major coneixement tant a nivell teòric com pràctic d'aquestes tècniques. Així el treball no pretén ser una recerca o un desenvolupament de una idea innovadora, sinó de aprendre de les que existeixen en la actualitat.

El treball consistirà en el desenvolupament i test de xarxes neuronal recurrents basades en dos tipus de cel·les de memòria: LSTM (Long-Short Term Memory) i GRU (Gated Recurrent Unit). Aquestes xarxes s'entrenaràn inicialment per a la realització de traduccions de texts en un idioma en un altre. No són importants els idiomes escollits, sinó la tècnica; ja que el mateix sistema podrà ser entrenat en diferents idiomes.

La realització requereix disposar de una ingent quantitat de traduccions ja existents. Aquesta informació s'obtindrà de conjunts de dades existents i que són en format obert. Com a cas excepcional s'optaria per la extracció de la informació de la xarxa. La aplicació de l'aprenentatge consisteix en la cerca de una configuració que maximitzi les traduccions correctes. Donat que aquest és un concepte subjectiu a la persona que llegeix la traducció, es definirà un criteri que podrà ser des de "res a veure" a "molt encertat".

Com a conclusions del treball es podrà obtenir una idea dels volums de dades necessaris per a obtenir resultats acceptables i obtenir un millor coneixement del funcionament de les xarxes neuronals recurrents i les cel·les de memòria que la composen.

Índex

1. Introducció

1.1 Context i justificació del Treball

El treball consisteix en la realització de un traductor Espanyol-Anglès fent servir tècniques basades en xarxes neuronals. Avui dia estan guanyant pes les tècniques d'auto aprenentatge i aquest treball és un punt de partida el qual es considera una base entre el què s'ha observat i el com una computadora pot descriure-ho. Com tot problema basat en xarxes neuronals artificials, la seva finalitat es la de imitar aquelles que no ho són. Concretament, les del cervell humà. Els avenços d'avui dia son significatius però encara disten de totes les funcions que som capaços. Així que, encara es requereix de la intervenció humana per tal de dur a terme, en aquest cas, tasques de traducció, sobretot en els casos amb paraules poc freqüents o en contexts especialitzats.

1.2 Objectius del Treball

Com a objectius del treball es troben els següents punts:

- Obtenció de un gran volum de dades per a l'entrenament de la xarxa neuronal.
- Aplicació amb èxit de una xarxa neuronal recurrent.
- Mesurar i avaluar els temps requerits d'entrenament per a diferents volums de dades.
- Avaluar i/o extrapolar quantes dades són necessàries per a obtenir traduccions acceptables.
- Experimentar amb diferents formes de entrenament. Textos llargs en primer lloc i textos curts en segon lloc. Diferents grups de dades.
- Entrenar la xarxa neuronal per a la traducció de textos en Espanyol a l'idioma Anglès amb resultats acceptables.
- Comparar els resultats amb altres traductors.

1.3 Enfocament i mètode seguit

En primer lloc són necessàries dades. És possible trobar traduccions de frases simples en repositoris de dades obertes tals com http://tatoeba.org/. Tanmateix son frases simples i no estan enfocades a cap context. I, com a repositori de frases simples, els resultats serien acceptables si el que es pretén traduir es una frase simple. Com el que es pretén és fer-ho lliure de context, al criteri de la xarxa neuronal, hi ha la possibilitat de obtenir les dades de pàgines web que faciliten la cerca per paraula i en els resultats s'obtenen les traduccions en que aquella paraula surt en un context. Com es el cas de http://www.linguee.es/. Com a opció en ment, a valorar, hi ha altres pàgines que ofereixen bones traduccions senzilles com http://wordreference.com i que podrien

ser un reforç posterior al que ja s'ha aprés. Això pot ser un motiu de des-virtualització de la xarxa i fer que el rendiment baixi.
També hi ha disponible un repositori amb 2 milions de traduccions disponibles a http://www.statmt.org/europarl/. Són de caire "parlamentari" així que més trauccions seràn necessàries per a fer-lo més genèric.
En segon lloc és necessari aplicar l'algorisme de forma efectiva. Donat que es considera la implementació de una tècnica existent i que ja ha estat desenvolupada i implementada com a codi obert per Google, es decideix fer servir la llibreria escrita en Python anomenada TensorFlow.
Evidentment no és una novetat la aplicació de xarxes neuronals a la traducció de textos però el cas és que encara és necessari millorar les traduccions oferides automàticament.

1.4 Planificació del Treball

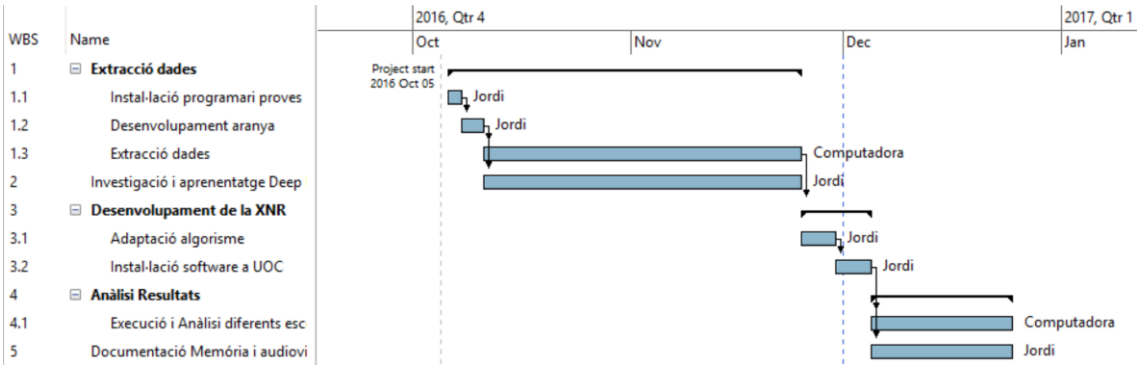
Per dur a terme el projecte es consideren necessaris els següents recursos:

- **Computadora**: Necessari per a totes les tasques involucrades.
- **Extractor de traduccions i emmagatzemador** dels diferents recursos online: Python és el llenguatge escollit.
- **Xarxa neuronal recurrent**: Les llibreries de Python creades per Google que integren el paquet TensorFlow contenen eines d'ajuda per a la creació d'aquesta.
- **Programari edició documents i audiovisuals**: Word, Powerpoint i Camtasia.

Les tasques es divideixen en les següents:

- Extracció de dades
  - Instal·lació del programari necessari
    - Python
  - Desenvolupament de aranya de pàgines web específic per a la tasca.
  - Extreure les dades.
- Desenvolupament de la Xarxa neuronal recurrent
  - Instal·lació de software
    - Python: Instal·lat prèviament
    - TensorFlow: Afegir llibreries a Python.
  - Desenvolupar algorisme. Subjecte a canvis segons es van fent progressos en l'anàlisi de resultats.
- Anàlisi de resultats.
  - Execució de diferents escenaris amb diferents tipus de dades.
  - Ajustar algorisme. És una possibilitat molt present.

El pla de treball es resum tal com segueix en el següent diagrama de gantt:



Aixi el projecte comença el 06 d'octubre de 2016 i pretén estar enllestit el 24 de desembre.
Tanmateix, al pla es considera que es pot treballar una mitja de 2 hores diaries de dilluns a divendres i 8 els caps de setmana.

Les fites no coincideixen amb les dates de lliurament de les PAC i en aquestes es preveu exposar un feedback de l'estat del projecte i aquest document actualitzat.

1.5 Breu sumari de productes obtinguts

No és senzill descriure els productes obtinguts doncs aquest projecte es centra en l'estudi d'un cert tipus de xarxa neuronal recurrent. Aquest estudi implica obtenir diverses xarxes amb diferents tamanys amb l'objectiu de fer una comparativa entre elles. Fruit de la comparativa s'obtindrà una noció de quines xarxes tenen un millor rendiment i quines cel·les de memòria són més pràctiques a l'hora de fer l'entrenament.
Com a productes que es poden considerar tangibles, o entregables, es podria dir que són:

- Un data set paral·lel (dos fitxers, un per Anglès i un per a Espanyol) amb 3973890 traduccions recopilades entre de diferents data sets.
- Un set de Xarxes Neuronals Recurrents de diferents tamanys.
- Aquesta memòria que recopila la estratègia feta servir, les experiències viscudes en la execució de la estratègia i uns resultats comparatius de les diferents Xarxes.

1.6 Breu descripció dels altres capítols de la memòria

La resta de capítols de la memòria són variats. En els primers capítols es parla dels conceptes que composen el que és la pràctica en general:

- Estat de l'art de les Xarxes Neuronals
- Composició d'aquestes
- Conceptes a tenir en compte en l'entrenament. Algorismes de convergència i ràtios de aprenentatge.
- Procés de entrenament de una xarxa.
- Comparativa entre les xarxes.

2. Xarxes Neuronals Recurrents – Estat de l'art

Les Xarxes Neuronals Recurrents (XNR) són una especialitat dintre del concepte mes generalitzat anomenat Xarxes Neuronals. Com totes elles, una XNR està composta de unitats de memòria amb la diferència de que aquestes unitats estan connectades formant un cicle dirigit. Aquesta construcció dota a la xarxa de una memòria temporal que pot ser considerada dinàmica, doncs pot variar en funció dels paràmetres de entrada amb les que entreni.

La gran aportació d'aquest tipus de xarxes és el fet de disposar d'una memòria que és reutilitzada inclús a llarg termini. Això és

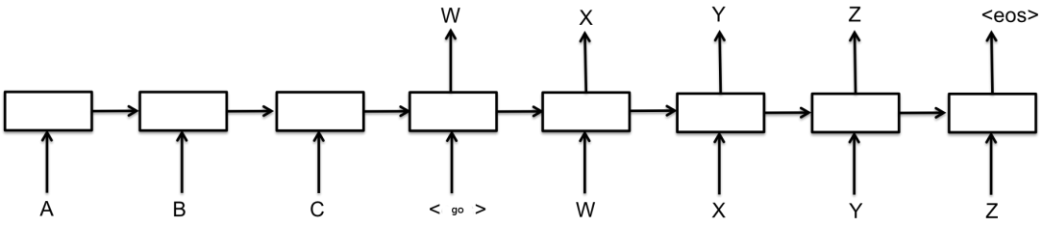
Hi ha una gran varietat d'arquitectures per a XNR. Si es fa una cerca a internet sobre "Recurrent Neural Networks" el primer resultat és un article de la Wikipèdia on hi ha unes divuit. Totes amb algun propòsit més específic o bé perquè en el seu dia van ser novetat i que ja han estat clarament superades per alguna altra. També val a dir que algunes d'aquestes arquitectures són combinacions d'altres o bé una metodologia específica. Aquestes són les divuit arquitectures que s'hi poden trobar amb un breu resum associat.

Arquitectura	**Descripció resum**
Fully recurrent network	Arquitectura bàsica creada als 80s amb els clàssics nodes de entrada, de sortida i la resta ocults. Cada node està conectat amb tots els altres. Existeixen nodes de entrada, sortida i ocults. També disposen de variació en els nodes en funcio del temps. Tots els casos següents són una sub-variació d'aquesta.
Recursive neural networks	Xarxa representada normalment en un graf diferenciable. Disenyades per a l'aprenentatge de representacions de structures distribuides com termes lógics. Aplicades al processament de llenguatge natural.
Hopfield network	Avui dia de interés histórico. La aplicació era l'estudi de sequencies de patrons. Té la garantía de que la xarxa convergirà cap a una xarxa raonable.
Elman networks and Jordan networks	Anomenades "Simple Recurrent Networks", basades en la arquitectura básica, són útils en la predicció de seqüències.
Echo state network	Xarxa poc conectada a la capa oculta (normalment un 1%). També pot reproduir patrons temporals específics.
Neural history compressor	Una de les primers arquitectures que prediu el próxim resultat basat en els resultats anteriors. Té la habilitat de "reconstruir" les capes mes profundes per a obtenir la entrada original. Es pot distinguir de un "conscient" i un "subconscient". El subconscient rarament varia en llargs entrenaments i pot ser forçada a predir tot i tenir dades no predictibles.
Long short-term memory	La xarxa que avui dia està en ús i que elimina el problema del "vanishing gradient". La xarxa està dotada de unitats

	de “oblit” que prevenen la propagaciò de errors com i que la xarxa “exploti” o hi hagi “desvaneixements”. Aplicada a reconeixement de veu, reconeixement de caracters, traducció i d’altres àrees donades les seves propietats.
Gated recurrent unit	Aquesta és un tipus de cel·la de memoria que es podría aplicar a una xarxa. No una arquitectura en sí.
Bi-directional RNN	Arquitectura que concatena dues Xarxes. L’aprenentatge es fa “from left to right” i “from right to left”. Això fa bona la xarxa en la predicció de “el Passat” i “el futur” del context de l’element. Sembla provat que es especialment útil combinat amb LSTM.
Continuous-time RNN	Model de xarxes biológiques. Aplicada al camp de evolució robotica. Per exemple, visió, cooperación i comportament cognitiu mínim.
Hierarchical RNN	Descomposició de una xarxa en subxarxes.
Recurrent multilayer perceptron	Xarxes en cascada on cada una s’entrena dels resultatsn anteriors. Excepte la última.
Second order RNN	Ús de pesos de alt ordre en comptes de els normals. Això permet disposar de una màquina de estats finits. LSTM en seria un cas, tot i no haber trobat un mapa finit en cap cas.
Multiple timescales recurrent neural network (MTRNN) model	Imitació de la activitat de un cervell fins a un cert punt. Organització propia de la xarxa per a les diferents activitats que hi puguin ocórrer.
Pollack's sequential cascaded networks	*Sense informació*
Neural Turing machines	Extensió de les xarxes neuronals a recursos de memòria externs amb els que es pot interactuar.
Neural network pushdown automata	Similar a la Neural Turin Machine, però amb l’objectiu de entrenar per a controlar.
Bidirectional associative memory	Xarxa en el que en l’entrenament es fa server una matriu de dades i la seva transposada. Sembla que tenen certa rellevància en aplicacions del món real.

Donat que aquest treball es basa en un exemple proporcionat pels desenvolupadors de “Tensorflow” amb alguna de les seves recomanacions aplicades, la arquitectura no pot ser una altra que la que fan servir. Aquest exemple cobreix les arquitectures “Long-Short Term Memory” i “Gated Recurrent Unit”. En poques paraules, la diferencia sembla ser el tipus de unitat de memòria que es fa servir que no pas un algorisme totalment diferent.

La xarxa en estudi és considerada una “Sequence 2 Sequence” que en realitat són dues xarxes neuronals recurrents on una primera realitza la codificació i la segona la decodificació. Aquest model ha estat provat com a apte per a la traducció de textos com es pot veure en l’article Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.
Com a exemple descriptiu de la xarxa es pot fer servir la següent imatge:



On “A, B i C” passen al codificador (una RNN) i “W, X Y, Z” s’obtenen com a part del procés del decodificador la segona (RNN).

3. Cel·les de memòria

Com s’ha introduït en el capítol anterior, una XNR està composta per unitats que se’n diuen de memòria. En aquest cas es tractarà només de les cel·les LSTM i GRU. Val a dir que les cel·les GRU són una variació de les cel·les LSTM. D’aquí que comparteixen objectiu, el de representar una neurona artificial, i que com a gran propietat a destacar inclouen un mecanisme de oblit que les permet no activar-se. Aquesta propietat millora les prediccions de la següent paraula en descartant les que podrien denominar-se menys probables.

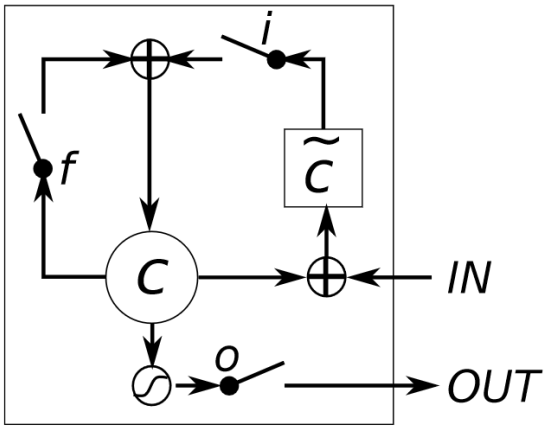
3.1 LSTM

LSTM és l’acrònim per a “Long-Short Term Memory”. Aquesta unitat, amb conjunció de un algorisme “gradient descent” apropiat va ser desenvolupada per millorar les unitats de memòria que representaven problemes amb xarxes que aplicaven “Back Propagation Through Time” i “Real-Time Recurrent Networks” on els errors de propagació o bé es feien enormes o bé es perdia la evolució temporal. El treball va ser realitzar per Hochreiter i Schmidsuber el 1997 i el treball complet es pot trobar aquí.

L’article explica molt detalladament les millores aplicades a aquesta arquitectura però, el que es creu que realment es una innovació, és la introducció de “unitats amb portes”. Aquestes portes són, bàsicament, una de entrada, una de sortida i una porta anomenada “porta del oblit”; que segons els paràmetres de entrada i/o sortida es permet l’accés i/o actualització del contingut de la unitat.

Aquesta estratègia permet evitar propagació de errors i/o pertorbar la memòria innecessàriament.

Aquesta cel·la es pot trobar representada tal com:



(a) Long Short-Term Memory

3.2 GRU

Les “Gate Recurrent Unit” són una evolució de les cel·les LSTM. El seu disseny proporciona exactament el mateix comportament tot i que es considera que el cost de còmput és raonablement menor.
Les següents figures resumeixen els càlculs que ha de fer cadascuna de les unitats LSTM i GRU:

LSTM (Long Short-Term Memory)

$$\boldsymbol{f}_t = \sigma\left(W_f\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_f\right), \tag{4a}$$
$$\boldsymbol{i}_t = \sigma\left(W_i\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_i\right), \tag{4b}$$
$$\boldsymbol{o}_t = \sigma\left(W_o\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_o\right), \tag{4c}$$
$$\tilde{\boldsymbol{c}}_t = \tanh\left(W_c\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_c\right), \tag{4d}$$
$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}}_t, \tag{4e}$$
$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t). \tag{4f}$$

GRU (Gated Recurrent Unit)

$$\boldsymbol{z}_t = \sigma\left(W_z\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_z\right), \tag{5a}$$
$$\boldsymbol{r}_t = \sigma\left(W_r\left[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_r\right), \tag{5b}$$
$$\tilde{\boldsymbol{h}}_t = \tanh\left(W_h\left[\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}, \boldsymbol{x}_t\right] + \boldsymbol{b}_h\right), \tag{5c}$$
$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t) \odot \boldsymbol{h}_{t-1} + \boldsymbol{z}_t \odot \tilde{\boldsymbol{h}}_t. \tag{5d}$$

Com es pot apreciar, el nombre de cálculs es pot considerar molt menor en GRU desde que cadascuna de les unitats que pertanyen a la xarxa han de fer aquests càlculs cada cop que es proveeix de una entrada.

3.3 Quina unitat de memòria és millor?

Hi ha uns quants articles en els que es fa la comparació entre les dues unitats. Si alguna cosa es pot entendre d’aquests és que aquestes dues unitats estàn empatades. En alguns casos LSTM es millor i en d’altres ho es la GRU. Inclús noves propostes de unitats es proposen igualar el comportament de LSTM (que també és el de GRU) simplement reduint el nombre de cómputs per a realitzar entrenaments en menys temps.

Així que donat que no queda clar quina pot ser millor opció en aquest cas, el que es pretén serà fer una comparativa d’ambdós.

4. Procés d'aprenentatge

Aquest capítol tracta de resumir els factors mes crítics en el procés escollit per a la realització de les proves. Aquests punts són:

- Dades per a l'aprenentatge
- Configuració de l'algorisme de optimització.

4.1 Dades per a l'aprenentatge

En l'exemple que proveeix l'equip de Tensorflow fa referència a un data set usat per a l'entrenament que conté 22 milions de frases. Després de 1 epoch (això és que almenys totes les frases s'han introduït al sistema una vegada) es considera que el sistema es apte per a ser provat amb resultats acceptables.

En aquest procés es pot considerar que hi ha tres jocs de dades.

Un és el de entrenament que és el que finalment acaba moldejant la xarxa neuronal. En aquest cas es disposen de 3973890 frases per a l'entrenament, fruit de la recopilació de tres datasets.

Un altre és el lot de test. Aquest sol ser mes reduït, amb unes 2000 frases i que es fa servir per a calcular la perplexitat de la iteració en curs. En aquest cas es fa servir el lot del data set de l'Europarliament amb uns dos mil registres.

Aquests data sets necessiten ser convertits a números per a poder ser processats matemàticament. Així que d'aquests fitxers en sorgeixen una altres tals com un fitxer de vocabulari amb les K paraules mes comuns i ordenades per freqüència a les frases proporcionades i una primera traducció de les frases originals a nombres. Com a exemple es pren la primera frase:

Ressumption	Of	The	Session
16579	7	4	1258

Val a dir que les paraules "The" i "Of" tenen número més baix perquè són més comuns.

Primeres paraules mes comuns en anglés:

0	_PAD (espai, intern)
1	_GO (traduiex, intern)
2	_EOS (end of sentence, intern)
3	_UNK (unknown, es pot veure en paraules no traduïbles)
4	The
5	,
6	.
7	Of
8	And
9	To
10	In
11	A

12	that

Val a dir que per a cobrir el major nombre de possibilitats per a frases simples s'ha cregut convenient usar un vocabulari de 200000 paraules per a cada idioma. En espanyol hi ha, més de 600000 paraules tenint en compte les conjugacions.

Aquestes dades seran processades fent servir el concepte dels ´mini lots´. Donat que aquestes xarxes requereixen una quantitat ingent de dades, no és possible carregar-les totes a memòria i, a més, crear les estructures paral·leles necessàries per a dur a terme l'entrenament. Aquestes serien, un mapa de paraules a identificadors numèrics i un mapa de les frases amb els identificadors en ordre invers. Això es deu a que si l'entrenament es fa amb les frases invertides els resultats resulten ser més òptims.

Així, els mini lots permeten fer ús parcial dels recursos de hardware disponibles de forma que es puguin processar un nombre considerablement reduït de sentencies sense perdre qualitat en l'objectiu global. També permet establir un recovery point una vegada el mini lot ha estat processat, de forma que és possible reanudar l'entrenament a partir de un punt.

Cal remarcar que en el transcurs de l'entrenament, la xarxa neuronal captarà i intentarà extreure factors comuns de les dades de entrada. Sigui la que sigui. Així, si la entrada son frases de text en castellà i les de sortida en anglés, la xarxa neuronal serà bona en aquest aspecte. Sempre comptat que quantes mes dades hi hagi, millor. Aquestes dades poden estar influenciades per el context del que tracta. En aquest cas s'han fet proves amb el data set de l'Euro Parliament, en el que disposen de 2 milions de frases i després de l'entrenament encertava algunes de les paraules que havia de traduir sempre que fossin de caire polític. Per a altres paraules relacionades amb la cuina, per exemple, es troba que les paraules eren majoritàriament desconegudes i per tant no traduïbles.

4.2 Bucketing i padding

Donada la variació de la longitud de les frases de entrada i de sortida l'equip de tensorflow ha introduït aquest concepte per a homogeneïtzar les longituds dels grafs. L'objectiu és el de no haver de crear grafs per a cada parella de frases ja que seria enorme amb molts subgrafs similars.

Primer de tot s'ha de considerar que les dades s'emmagatzemen amb la frase de origen, un codi especial i la frase traduïda. La frase de origen està invertida i sembla que d'aquesta manera s'obtenen millors resultats segons l'estudi de Ilya Sutskever, Oriol Vinyals i Quoc V. Le

Aquesta tècnica defineix un nombre de buckets tals com:

```
buckets = [(5, 10), (10, 15), (20, 25), (40, 50)]
```

Això ve a representar que per a les frases de origen es consideren les longituds de la esquerra (5, 10, 20, 40) i per als de destí els de la dreta (10, 15, 25, 50).

En el cas de que la frase origen sigui de 6 i la de destí de 8, el primer bucket serà escollit. Els espais lliures s'omplen amb la paraula _PAD. En cas de que una de les frases càpiga en un bucket major, aquest serà el fet servir. Això es que si la frase de destí és de 18, s'haurà de fer servir el bucket (20, 25) per a ambdós frases i omplint amb _PAD els espais lliures.

Una representació de exemple:

Original: [sesiones, de, período, del, Reanudación]
Destí: [GO Resumption, of, the, session, EOS, PAD, PAD, PAD, PAD, PAD]

4.3 Algorismes de optimització

Aquests algorismes, junt amb el tipus de cel·la, són una de les parts mes importants per al que serà al final la xarxa neuronal. Cada cop que s'introdueix una frase, l'algorisme recorre les cel·les modificant els pesos d'aquestes. Recordem que les cel·les amb les que tractem tenen la habilitat de no ser modificades o ser ignorades en funció dels pesos de entrada. Aquesta es una de les avantatges de les xarxes basades en LSTM, que preveuen la propagació de errors.

Es pot dir que la funció de l'algorisme es fer convergir les dades cap a altres grups de dades. I això de forma en que cada cop aquestes hauran de convergir en valors mes reduïts. Això està relacionat amb el ràtio d'aprenentatge, que és el valor que es va reduint per a forçar la evolució de la xarxa.

Com a algorismes de optimització també se'n poden trobar unes quantes famílies també:

- Batch Gradient Descent
- Esthocastich Gradient Descent
- Mini batch Gradient descent.

En aquest cas es fa ús algorisme de la familia "Mini batch gradient desdent" anomenat "Adagard" per simples raons. Està recomanat per els desenvolupadors de Tensorflow i ja el tenen implementat. (Aquesta configuració no ve per defecte)

4.4 Perplexitat

La perplexitat és el valor resulta de avaluar les frases de test amb l'estat de la xarxa actual. Aquest és un valor que ha de tendir a zero i per tant quant mes baix sigui millor.

En els resultats de la prova es mostra la perplexitat en el moment en que s'atura el test. La intenció és que almenys els dos primers buckets tinguin valors menors a 10.

4.5 Ràtio d'aprenentatge

Cada cop que es fa servir l'algorisme de optimitazció, s'especifica (diguem) com de llunyans o propers han de estar els resultats mes o menys semblants. Aquest valor es modifica cada cop que es perceb que la perplexitat no convergeix cap a zero reduïnt el seu valor. I l'efecte que té és que la perplexitat, efectivament baixa. Això provoca que les dades convergeixin cap als seus espais, hipoteticament, de forma més precisa.

4.6 Factor decreixent del ràtio

Si més no, és posible especificar en quant ha de decrèixer el ratio, sempre suposant que será multiplicat per un factor (0.5 per exemple).

Així, cada cop que es percebi que la perplexitat no millora els resultats anteriors, el ràtio d'aprenentatge decreixerà proporcionalment amb aquest factor, provocant la convergencia de les dades.

5. Procés entrenament

Les xarxes neurones profundes son considerades aquelles que tenen almenys 3 capes de profunditat. La primera capa és considerada de la de input, la segona es diu 'hidden' i la tercera seria la de output.

S'han recopilat uns gairebé 4 milions de frases entre els datasets de Tatoeba, l'Euro parliament i el crawling que s'ha realitzat a pàgines que ofereixen traduccions.

La configuració de l'algorisme, com bé s'ha introduït, farà servir l'algorisme Adagard, proposat per TensorFlow.

Els mini-lots serán de 64 frases, i els "checkpoints" cada 512 iteracions. La qual cosa permetrà identificar amb anticipació si és necessari reduir el ratio d'aprenentatge. Com a contrapartida pot fer que l'entrenament pugui durar mes temps donat que es fan comprovacions mes sovint.

El ratio d'aprenentatge i el factor es deixaran els que vénen per defecte i que són 0.5 i 0.9 respectivament.

Està previst que es pugui fer ús de grans xarxes (donat que a la meva computadora la mes gran amb tres capes es de 378 unitats per capa i no era possible fer servir mes de les dos milions de frases de l'Euro parliament). Aquestes xarxes estan previstes ser:

- 3*1024
- 4*1024
- 8*512

Es probable que cadascun d'aquests tets trigui dies a partir del segon test per a convergir apropiadament. Tot dependrà de la potència de còmput que es pugui fer servir.

Un cop realitzat el primer test s'establirà un set de frases assolibles per a la conjunció de datasets per tal d'evaluar si es perceben millores substancials en xarxes de major tamany.

Tanmateix el nombre de paraules amb la que es decideix entrenar la xarxa es rellevant. En un inici les proves es van realitzar amb les 40000 paraules mes freqüents per a cada idioma. En aquest cas és necessari ampliar el nombre de paraules si el que es vol és no perdre qualitat en la traducció.

6. Proves realitzades i resultats obtinguts

Abans de entrar als detalls dels resultats de es mostra una comparativa entre aquests. Aquestes comparatives estaran basades en la perplexitat doncs és la mesura que indica a quin punt de convergència s'ha arribat. Val a dir que les proves realitzades i les comparatives es basen en aquesta taula de jocs de proves:

Tamany Xarxa	LSTM	GRU
8x512	✓	✓
3x1024	✓	✓
4x1024	✓	✓
3x1200	✓	✗ (Memòria insuficient)

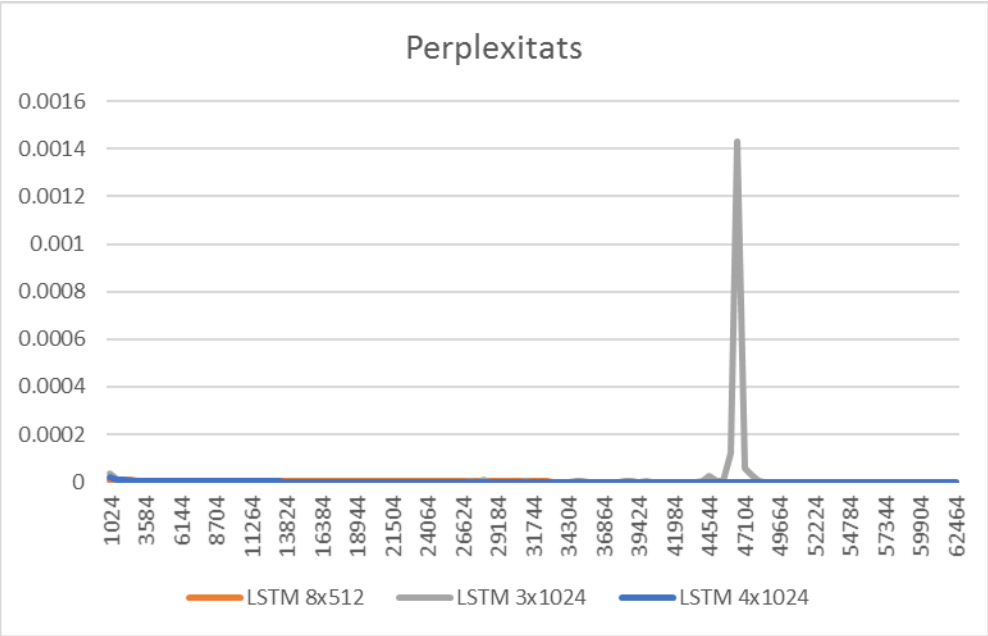
Les primeres dues comparatives seràn les xarxes del mateix tipus. Això és comparar les xarxes LSTM i en segon lloc les GRU.

Les següents comparatives seràn entre les diferents xarxes del mateix tamany on una és LSTM i l'altre es GRU.

Totes les dades mostrades s'han normalitzat a valors entre 0 i 1 per tal de reduir les distàncies que es poden mostrar en les gràfiques. S'ha provat de normalitzar estadísticament però no el resultat no era millor.

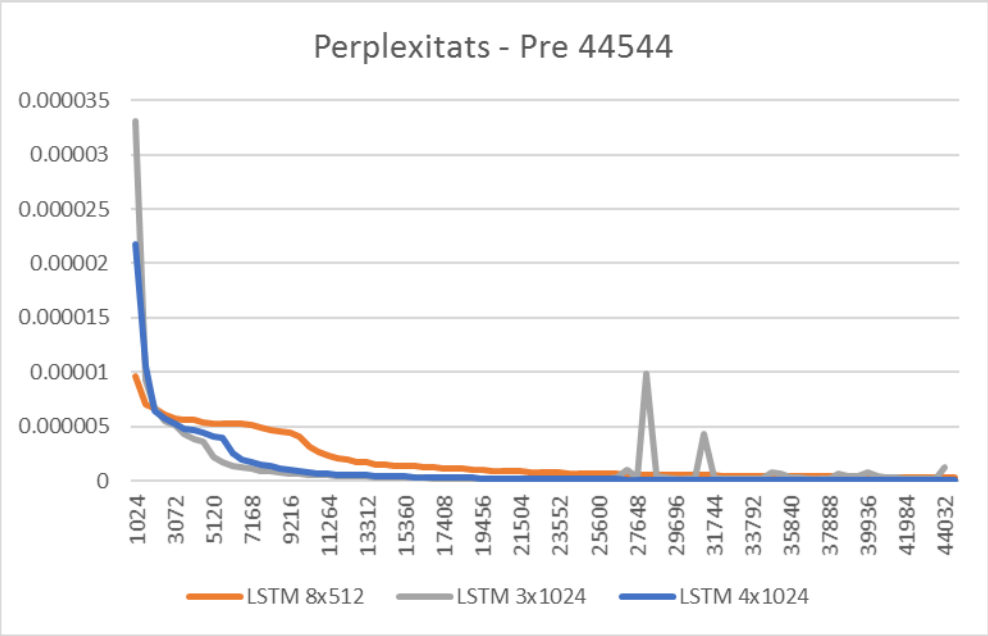
Comparativa LSTM

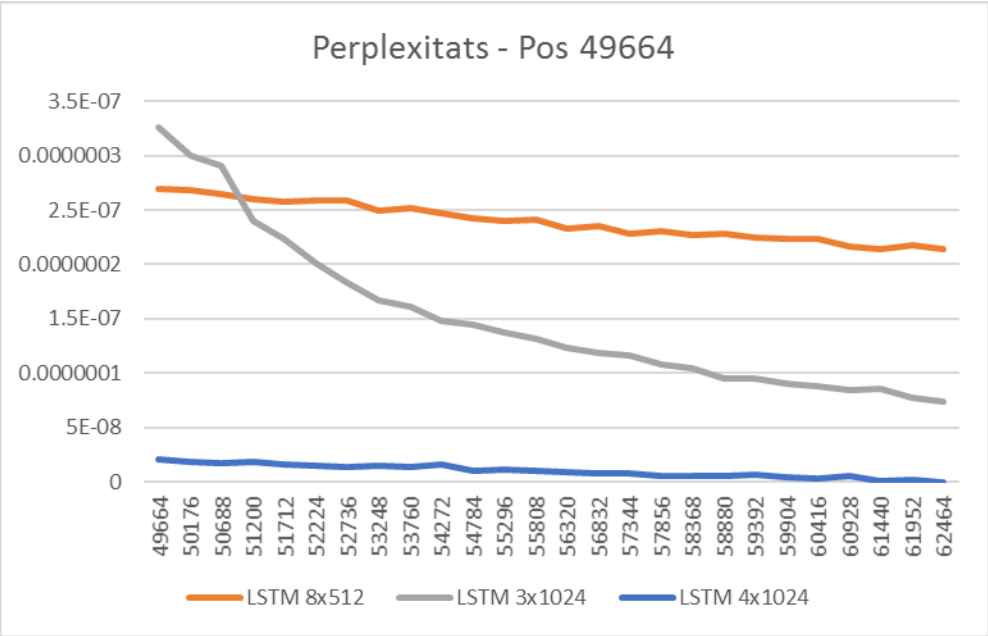
En un gràfic comparatiu de les xarxes es pot apreciar el següent:



Donat que el al voltant del pas 47104 hi ha un valor gran (la xarxa s'ha reconfigurat fortament), no es pot arribar a veure que succeeix ni quins valors son millors o pitjors en la resta.

Per a observar el comportament s'ha dividit en les següents dues gràfiques:



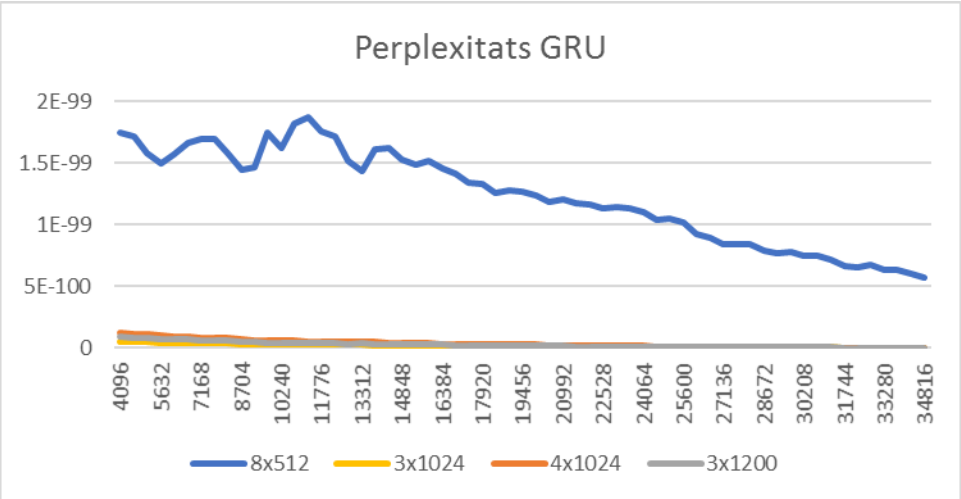


Ara es pot apreciar millor que la xarxa 3x1024 és la que pitjor comença, hi ha moments en que s'ha de reconfigurar donat que els resultats no convergeixen i es requereix un ajust però no acaba amb un valor prou baix, millor a 8x512 i pitjor a 4x1024. Tot i semblar grans diferències, en la última iteració els valors no noramlitzats eren:

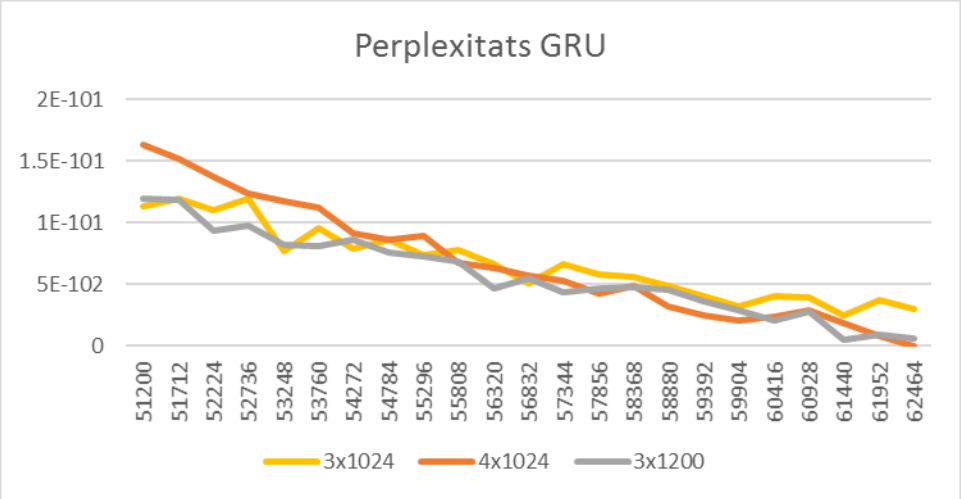
- 8x512: 6.31
- 3x1024: 6.24
- 4x1024: 4.08

Comparativa GRU

Les perplexitats GRU mostren un comportament un tant diferent però no gaire. Per variar, els resultats per als diferents tamanys difereixen molt en una gràfica degut a la xarxa de 8x512.



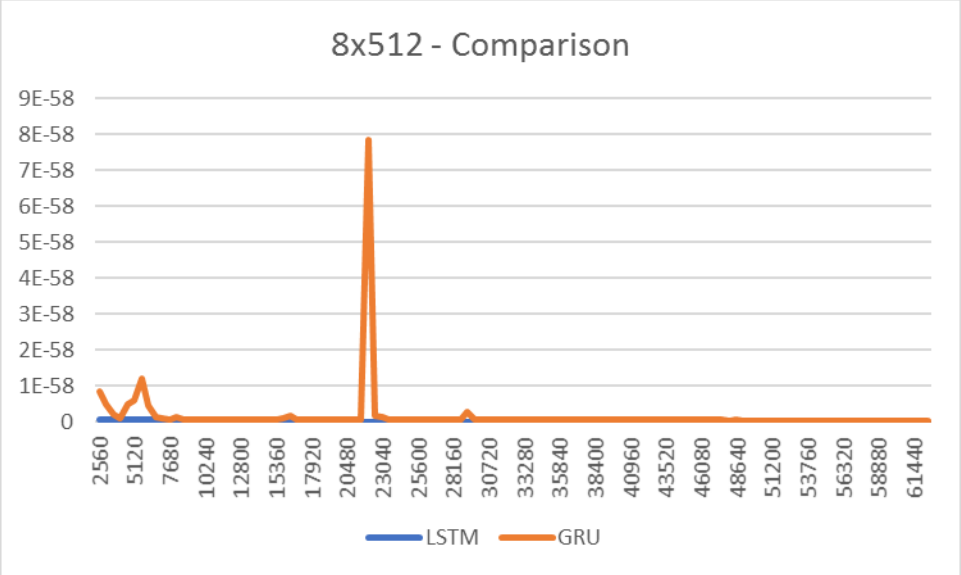
Un cop més és necessari anar la cua dels resultats a veure quina és la evolució de la perplexitat traient els resultats de 8x512.



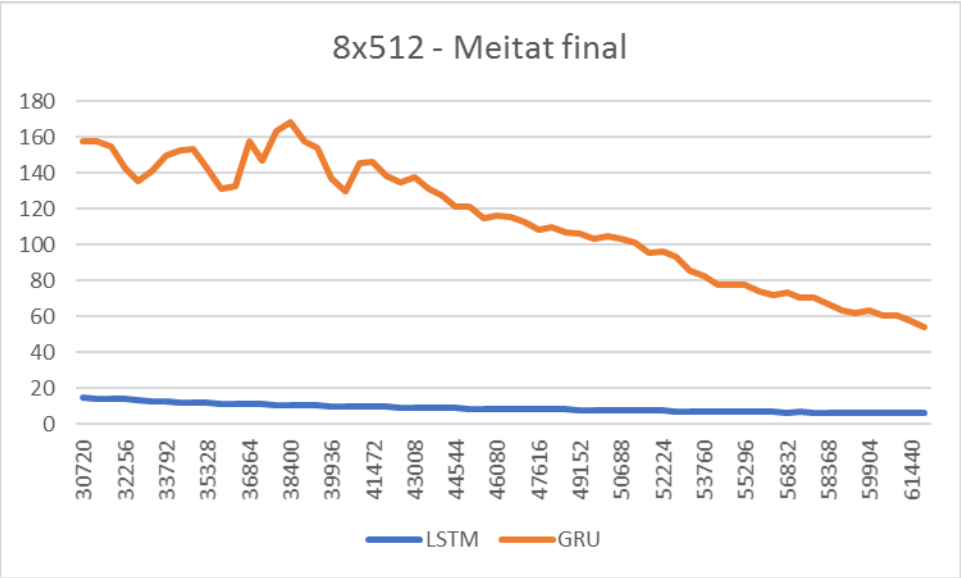
A partir d'aquest resultats veiem que hi ha certa tendencia a convergir cap al mateix punt per part de totes les xarxes. A simple ull, sembla que la xarxa 4x1024 té major pendent i dóna a entendre que amb més iteracions podría destacar entre les altres dues. La xarxa 3x1200 ha semblat convergir millor durant tot el test fins al final en el que empata amb 4x1024.

Comparativa 8x512

Aquesta es la gràfica resultant i un cop més els valors no es poden apreciar a un simple cop de vista.



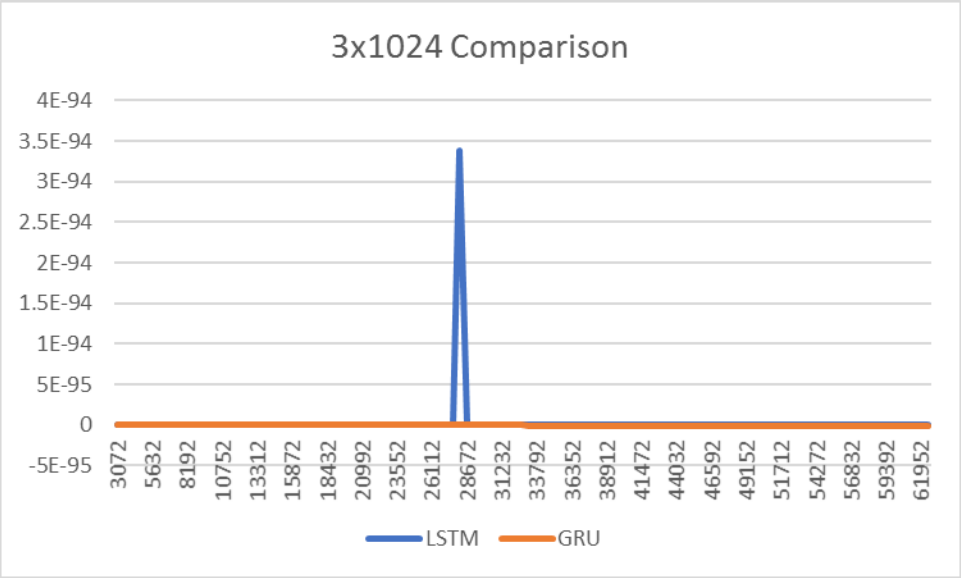
En aquesta següent es mostra la meitat final del procés amb els resultats no normalitzats on es pot apreciar millor:



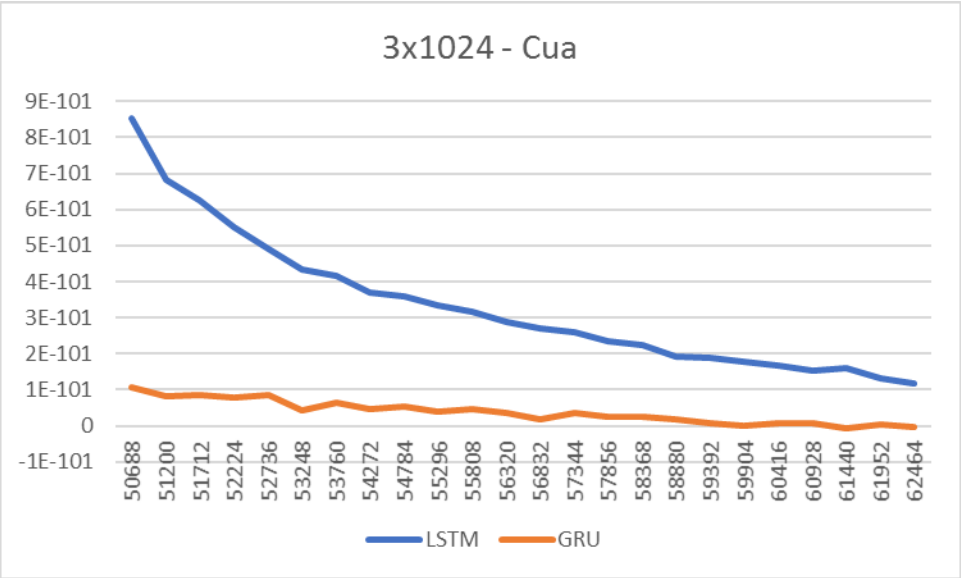
El comportament de la LSTM ha mostrat ser més consistent en tot el procés, no només al final. La xarxa GRU mostra tenir una tendéncia a zero que molt possiblement amb mes dades i més iteracions hauria de arribar a valors molt més propers dels de la LSTM.

Comparativa 3x1024

Tornem a tenir una gràfica suficientment dispar com per a poder avaluar les dades de forma general.



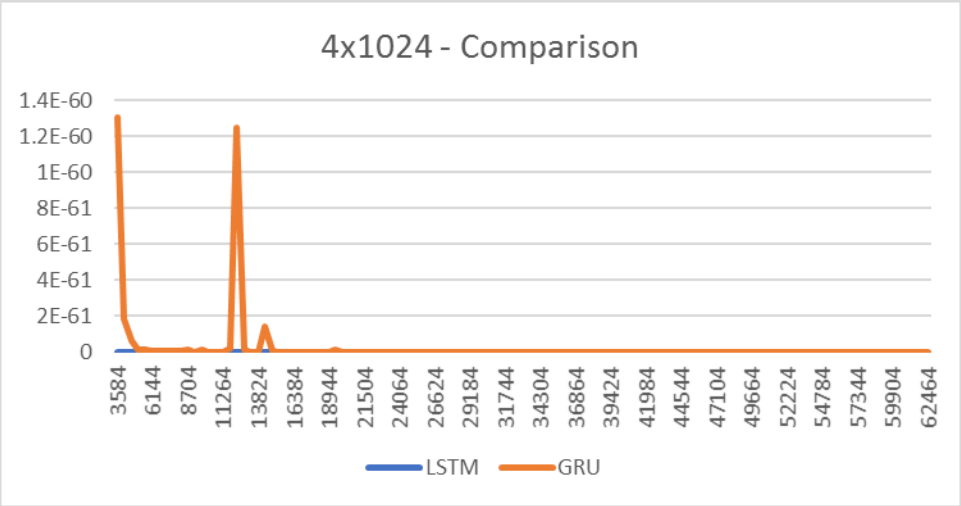
Si es para atenció a com és la gràfica per allà cap al final el que es pot veure és el següent:



En aquest cas sembla que la xarxa GRU ha mostrat major convergència durant tot el procés i més estabilitat que no pas la LSTM on es pot veure desde la primera grafica que hi ha perplexitats que es disparen.

Comparativa 4x1024

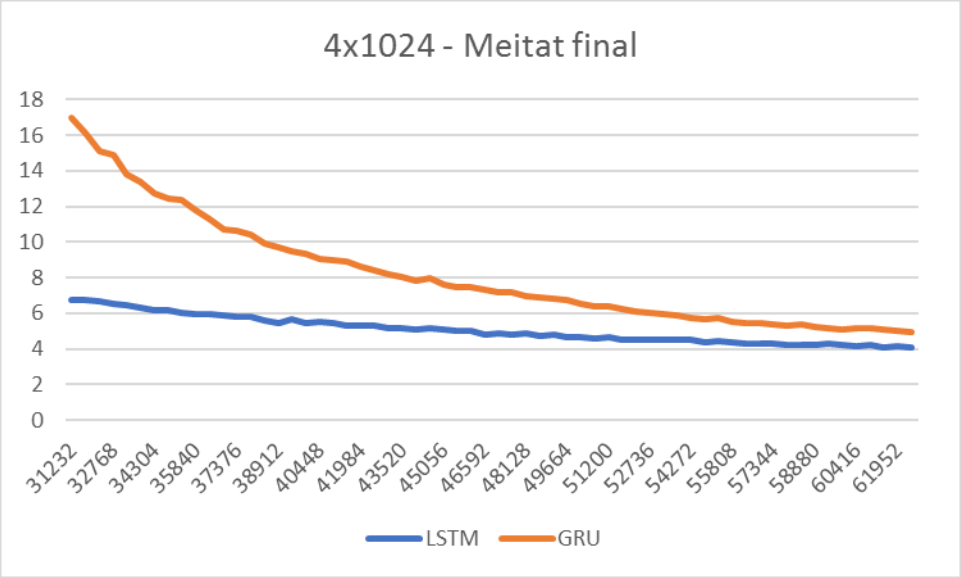
Com vé sent habitual la gràfica general no mostra molta de la informació mes rellevant donat a que hi ha valors alts en certs punts del procés:



Cal mencionar que en aquesta gràfica no es mostren les primeres iteracions degut a que els valors per general són astronómics, concretament els de GRU. Com a exemple en la primera iteració:

- Perplexitat – GRU: *212465313075409852735806743505082338453649354317595626966910133862 4.00*
- Perplexitat – LSTM: *1066622.04*

Així que es torna a mostrar la meitat final del procés amb els resultats no normalitzats on si que es pot apreciar la tendència:



Un altre cop, LSTM ha estat sempre per sota de GRU i amb mes estabilitat durant tot el procés però en aquest cas, en les iteracions finals sembla que GRU pot arribar a igualar LSTM en el valor 4 que comença a seblar assimptótic.

Resultats generals

Les xarxes 4x1024 LSTM, 4x1024 GRU i 3x1200 GRU tenen un nombre similar de perplexitat després de les 122 iteracions de 512 passos de 64 frases cadascún. A l'annex es pot trobar els resultats de provar aquestes xarxes amb diferents sentències i el resultat és que cap d'elles es prou bona per destacar davant de les altres. Algunes tenen més encert que d'altres en algunes frases. Com a element destacable, cap xarxa es bona amb paraules simples. Potser seria convenient fer ús de un diccionari i introduïr paraula a paraula les posibles traduccions.

En qualsevol cas, per a frases, el context normalment és encertat però la precisió de les paraules escollides per la xarxa no són prou bones.

Es creu que aquestes xarxes serien molt millors si es dispossessin de un dataset mes gran i un set de test menys polític i més gran.

7. Conclusions

Val la pena reincidir en que aquest treball és un inici a una tecnologia que cada cop guanya mes pes avui dia, constantment e evolució, i que es desconeixia totalment res similar vist o professionalment o acadèmicament.

En profunditzar en el que podria ser el temari s'ha trobat una infinitat de informació que és molt densa i no es poden trobar exemples senzills per a entendre el funcionament. Tot requereix un altíssim nivell de coneixements matemàtics i milions de entrades de informació!

Així, a poc a poc, es va anar llegint i entenent el funcionament teòric de una xarxa com la feta servir a la pràctica. Dedicant-hi molt de temps a llegir articles i exemples per, finalment, destriar el que realment era necessari fer i/o entendre l'exemple en el qual es basa la pràctica.

Addicionalment, un cop la teoria sembla ser consistent, resulta que els recursos necessaris no són a l'abast de qualsevol per a fer-ho en temps raonables. Un punt en el que va ser necessari insistir va ser el de fer servir una computadora amb GPU doncs els resultats eren 20 vegades mes ràpids que fent ús de la memòria i CPU tradicionals. I tot això en xarxes molt, però que molt, mes petites a les vistes en aquesta pràctica.

Els objectius eren molt més ambiciosos inicialment. Abans de començar a fer res, es va poder analitzar que era possible configurar ratis de convergència inicials, aplicar diferents datasets, fer servir dos tipus de cel·les, diferents tamanys, algorisme de convergència i ús de diferents tokenitzadors. Tal és el cas que només s'han pogut aplicar els diferents tamanys, els que permetia la GPU disponible, i els tipus de cel·les. Tot això ja ha implicat haver de tenir un recurs disponible funcionant al 100% dues setmanes. El temps just per a poder lliurar la pràctica amb els resultats presentats.

En cas d'haver tingut més temps, s'haurien pogut realitzar altres jocs de proves amb diferents paràmetres i/o funcionalitats (algorisme de convergència Gradient Descent, no Adagard i un tokenitzador anomenat NLTK)

La planificació inicial només va encertar en les tasques a dur a terme. El temps que s'havia de dedicar a les tasques realment ha estat un altre. Tasques com instal·lació de software es van fer sense inconvenients i en temps. Altres tasques com "Extracció de dades" van prendre gairebé 3 mesos quan hi havia previst 25 dies. No era possible preveure que les pàgines d'on s'obtenien les dades tenien protecció per a fer crawling i per tant es va haver de desenvolupar alternatives molt mes lentes.

De les dades fetes servir en l'entrenament la única que podria ser de una qualitat no tan bona es la obtinguda fent servir crawling. Aquesta son 1.9 milions de sentencies que no han estat revisades.

Per a dur a terme amb èxit la pràctica s'ha hagut de invertir temps addicional i procurar obtenir un recurs amb una GPU prou potent per a poder executar xarxes neuronals mes grans.

Una cosa que recomano a qualsevol persona que vulgui realitzar una practica com aquesta, és la de disposar temps complet per a fer-la. De un dia per a l'altre i havent treballat durant 8 hores, fa que es pensi més que s'hi actuï. I una pràctica com aquesta és molt exigent en els detalls. Detalls que m'hauria agradat especificar millor.

En resum estic satisfet dels resultats obtinguts. Era coneixedor de que serien necessaris més de 4 milions de sentencies per a poder obtenir uns resultats acceptables. La varietat de possibilitats que ofereix el llenguatge castellà és pràcticament infinita amb les possibles conjugacions aplicables. Obtenir un data set mes gran per veure quan les traduccions comencen a ser realment bones seria un repte personal a assolir.

També, aquest tipus de xarxes neuronals es poden aplicar a molts altres camps i que potser són menys exigents com ara "speech processing". El fet de marcar com a entrada un arxiu de so i com a sortida un text (o viceversa) realment atrau la meva atenció i seria un pas a poder fer traduccions automàtiques en temps real. Això últim és ambiciós d'arrel des de que la millor xarxa vista en aquesta pràctica no és prou bona encara.

8. Glossari

GPU – Graphics Processor Unit: Unitat semblant a la CPU amb major rendiment i que normalment s'usa per a còmputs gràfics.
GRU – Gated Recurrent Unit: Versió de les cel·les LSTM amb menys cost en el còmput.
LSTM – Long Short Term Memory: Tipus de cel·la que composa una xarxa neuronal recurrent.
RNN – Recurrent Neural Network: Traducció de XNR.
XNR – Xarxa Neuronal Recurrent: Tipus de xarxa artificial composta per cel·les de memòria. Això vol dir que les cel·les retenen la informació en el temps i la seva sortida pot ser la entrada de ella mateixa en un moment donat.

9. Bibliografia

https://en.wikipedia.org/wiki/Deep_learning
https://en.wikipedia.org/wiki/Artificial_neural_network
https://en.wikipedia.org/wiki/Backpropagation
https://en.wikipedia.org/wiki/DeepMind
https://en.wikipedia.org/wiki/Multilayer_perceptron
https://en.wikipedia.org/wiki/Universal_approximation_theorem
https://en.wikipedia.org/wiki/Bayesian_inference
https://en.wikipedia.org/wiki/DeepDream
https://en.wikipedia.org/wiki/ImageNet
https://en.wikipedia.org/wiki/Hidden_Markov_model
https://en.wikipedia.org/wiki/Recursive_neural_network
https://en.wikipedia.org/wiki/Recurrent_neural_network
https://www.coursera.org/learn/machine-learning/lecture/9zJUs/mini-batch-gradient-descent
http://www.statmt.org/wmt15/translation-task.html
https://www.tensorflow.org/versions/r0.11/tutorials/seq2seq/index.html
https://github.com/tensorflow/tensorflow/releases
https://github.com/eBay/Sequence-Semantic-Embedding
https://arxiv.org/pdf/1406.1078v3.pdf
http://deeplearning.cs.cmu.edu/pdfs/Hochreiter97_lstm.pdf
https://www.coursera.org/learn/machine-learning/lecture/9zJUs/mini-batch-gradient-descent
https://pdfs.semanticscholar.org/2d9e/3f53fcdb548b0b3c4d4efb197f164fe0c381.pdf

10. Annexos

Annex 1 – Experiments

Experiment 8x512 LSTM

Paràmetre	**Configuració**
Fitxer de training	tfm.es-en.en; tfm.es-en.es
Fitxer de test	tfm_test.en; tfm_test.es
Nombre de paraules Espanyol	200000
Nombre de paraules Anglés	200000
Ràtio d'aprenentatge	0.5 inicial
Factor decreixent del ratio	0.99
Tamany RNN	8 capes de 512 celles LSTM
Tamany batch	64
Cel·la	LSTM
Algorisme Gradient Descent	Adagard
Máxim dades entrenament	4000000
Checkpoint	Cada 512 passos.

Resultats:

Temps entrenament:	19 hores 4 minuts.
Dades últim checkpoint	global step 73216 learning rate 0.3699 step-time 0.69 perplexity 6.08 eval: bucket 0 perplexity 2.63 eval: bucket 1 perplexity 4.08 eval: bucket 2 perplexity 7.20 eval: bucket 3 perplexity 9.75
Observacions	Aquest test ha ha estat entrenat fins a la iteració 73216 accidentalment.

Test

Hola	Zimbabwe
Perro	_UNK
Gato	_UNK
Ratón	_UNK
Murciélago	Heckling
Había una vez un circo	There was one day a _UNK !
Había cierto rencor en sus palabras	I was a bad guy in his words
Mañana iremos al parque de atracciones	I ' ll go on the _UNK of _UNK
En un lugar de la Mancha de cuyo nombre no	In a _UNK of the name of the name I don ' t

quiero acordarme	like to be a
Érase una vez, en un lugar muy lejano	Let us go , at a very high level
La economía está cada vez peor	The economy is rising
Quién mató a Rogger Rabbit?	Who ' s a _UNK ?
Pablito clavó un clavito. Qué clavito clavó Pablito?	Can you use a _UNK , " _UNK _UNK ?
Me duele el codo	I ' m seeing the guitar
Me duele la cabeza	that the head
Eran mas de cien y menos de mil	Some more than 0 and less
Quinientas mil personas assistieron al mayor evento jamás visto	_UNK 000 people in the world even worse than
Quién és el presidente de los Estados Unidos?	Who ' s the President of the United Kingdom ?
Tenemos que poner orden en la asamblea	We must put a right in the Bureau
Hay que poner orden en la asamblea	A follow-up must be made in the Bureau
Me dijeron que fuera directo al médico	I told you ' t have any credit
Me dijeron que fuera directamente a urgencias	I told you to be directly arrested
Felices fiestas!	Get a visit !
Feliz Navidad!	Christmas day !
Se hizo un corrillo alrededor de su mesa	A _UNK of your book
Había montones de personas mirando el accidente de tráfico	Some people have lost the accident accident
Toda mala época llega a su fin	Your time lost after your first .
Cuando dos problemas se juntan es mas difícil encontrar una solución	When two challenges are made , it is difficult to find a change !
Cuándo te lo dijeron?	Where did you say ?
En esta época del año, invierno, las temperaturas son muy bajas	In this period of the year , winter , temperatures are very expensive
Alguien debe de hacerse responsable por las acciones ocurridas	Who must be responsible for the immediate actions
Quién és el responsable de la organización de los eventos en la empresa?	Is it responsible for the responsibility of the organization in the field of the company ?
Ayer estuvimos comiendo juntos en el jardín	Last night we live in the mountain
Quién poco coco come, poco coco compra	You ' re eating eating candy , you ' re eating food !
A quién madruga, Dios le ayuda	And who , God , will give you
A Dios rogando y con el mazo dándo	the Church and the _UNK _UNK
Para cuándo podremos ver un personaje carismàtico?	How can we see a _UNK _UNK ?

Experiment 8x512 GRU

Paràmetre	**Configuració**
Fitxer de training	tfm.es-en.en; tfm.es-en.es
Fitxer de test	tfm_test.en; tfm_test.es
Nombre de paraules Espanyol	200000
Nombre de paraules Anglés	200000

Ràtio d'aprenentatge	0.5 inicial
Factor decreixent del ratio	0.99
Tamany RNN	8 capes de 512 celles
Tamany batch	64
Cel·la	GRU
Algorisme Gradient Descent	Adagard
Máxim dades entrenament	4000000
Checkpoint	Cada 512 passos.

Resultats:

Temps entrenament:	16 hores 17 minuts
Dades últim checkpoint	global step 62464 learning rate 0.3118 step-time 0.71 perplexity 54.34 eval: bucket 0 perplexity 22.12 eval: bucket 1 perplexity 20.16 eval: bucket 2 perplexity 47.58 eval: bucket 3 perplexity 73.69
Observacions	

Test

Hola	Welcome
Perro	Welcome
Gato	Welcome
Ratón	_UNK
Murciélago	Welcome
Había una vez un circo	You have a _UNK of the _UNK
Había cierto rencor en sus palabras	Tom ' s I I I I ll !
Mañana iremos al parque de atracciones	_UNK the _UNK of the _UNK
En un lugar de la Mancha de cuyo nombre no quiero acordarme	If you you I I I I I I I I I I not
Érase una vez, en un lugar muy lejano	Tom you you you you you you you your _UNK .
La economía está cada vez peor	The European Union of the _UNK of the _UNK
Quién mató a Rogger Rabbit?	How you you you you you you you ?
Pablito clavó un clavito. Qué clavito clavó Pablito?	How you the _UNK , _UNK , _UNK ?
Me duele el codo	I you m you you you you you you you
Me duele la cabeza	I you re you you you you you you you
Eran mas de cien y menos de mil	_UNK and _UNK and _UNK and _UNK
Quinientas mil personas assistieron al mayor evento jamás visto	_UNK the _UNK of the _UNK of the _UNK
Quién és el presidente de los Estados Unidos?	Are you the Council of the Council of the United States ?
Tenemos que poner orden en la asamblea	We have a _UNK of the _UNK of the _UNK
Hay que poner orden en la asamblea	That is the _UNK of the _UNK
Me dijeron que fuera directo al médico	I you re you you you you you you you your own .
Me dijeron que fuera directamente a urgencias	I you m to the heart and you you you your own

Felices fiestas!	I ' s !
Feliz Navidad!	I ' re !
Se hizo un corrillo alrededor de su mesa	Tom was the _UNK of the _UNK of the _UNK
Había montones de personas mirando el accidente de tráfico	A and s _UNK , the _UNK of the _UNK
Toda mala época llega a su fin	A ' s _UNK , you you you !
Cuando dos problemas se juntan es mas difícil encontrar una solución	In the case , we we have be a same of not be the case !
Cuándo te lo dijeron?	How you you you you you you ?
En esta época del año, invierno, las temperaturas son muy bajas	In the case , the _UNK of the _UNK of the _UNK , the _UNK ,
Alguien debe de hacerse responsable por las acciones ocurridas	A of the _UNK of the _UNK of the _UNK
Quién és el responsable de la organización de los eventos en la empresa?	How the _UNK of the _UNK of the _UNK of the world of the world ?
Ayer estuvimos comiendo juntos en el jardín	_UNK your _UNK of the _UNK
Quién poco coco come, poco coco compra	_UNK , _UNK , _UNK , _UNK , _UNK
A quién madruga, Dios le ayuda	I you not I I I I I I not
A Dios rogando y con el mazo dándo	The _UNK , the _UNK , the _UNK ,
Para cuándo podremos ver un personaje carismàtico?	How you you you you you you you you you you you your ?

Experiment 3x1024 LSTM

Paràmetre	**Configuració**
Fitxer de training	tfm.es-en.en; tfm.es-en.es
Fitxer de test	tfm_test.en; tfm_test.es
Nombre de paraules Espanyol	200000
Nombre de paraules Anglés	200000
Ràtio d'aprenentatge	0.5 inicial
Factor decreixent del ratio	0.99
Tamany RNN	3 capes de 1024 celles
Tamany batch	64
Cel·la	LSTM
Algorisme Gradient Descent	Adagard
Máxim dades entrenament	4000000
Checkpoint	Cada 512 passos.

Resultats:

Temps entrenament:	16 hores 17 minuts.
Dades últim checkpoint	global step 62464 learning rate 0.3699 step-time 0.69 perplexity 6.24 eval: bucket 0 perplexity 4.42 eval: bucket 1 perplexity 2.63

	eval: bucket 2 perplexity 6.33 eval: bucket 3 perplexity 10.08
Observacions	

Test

Hola	VOTES
Perro	VOTES
Gato	VOTES
Ratón	Votes
Murciélago	Votes
Había una vez un circo	This was a case of a
Había cierto rencor en sus palabras	I have been in a certain word .
Mañana iremos al parque de atracciones	Click to the park of the park
En un lugar de la Mancha de cuyo nombre no quiero acordarme	On the one of the _UNK I do not want to be
Érase una vez, en un lugar muy lejano	Once once again , a very long place
La economía está cada vez peor	The economy is increasingly worse .
Quién mató a Rogger Rabbit?	Did you call me to _UNK ?
Pablito clavó un clavito. Qué clavito clavó Pablito?	_UNK _UNK _UNK _UNK ?
Me duele el codo	I love the _UNK
Me duele la cabeza	I got the head
Eran mas de cien y menos de mil	More than 00 and less than 0 years of the aid
Quinientas mil personas assistieron al mayor evento jamás visto	_UNK persons - 000 people will never be able to take the greatest community
Quién és el presidente de los Estados Unidos?	The President is the president of the United States ?
Tenemos que poner orden en la asamblea	We need to put on the floor to the
Hay que poner orden en la asamblea	The order to be taken in the Chamber
Me dijeron que fuera directo al médico	I would like to tell you that I would be asked to answer the doctor
Me dijeron que fuera directamente a urgencias	I would like to say that directly directly is directly directly .
Felices fiestas!	You ' re talking !
Feliz Navidad!	Well !
Se hizo un corrillo alrededor de su mesa	A few minutes around your table were around
Había montones de personas mirando el accidente de tráfico	People had been killed from the accident accident
Toda mala época llega a su fin	All the bad food problems are coming to their own
Cuando dos problemas se juntan es mas difícil encontrar una solución	When two problems are difficult , it is difficult to find a solution
Cuándo te lo dijeron?	Do you like to tell you ?
En esta época del año, invierno, las temperaturas son muy bajas	In this year this year , winter , winter , and winter
Alguien debe de hacerse responsable por las acciones ocurridas	The responsibility should be responsible for action by the actions concerned
Quién és el responsable de la organización de los eventos en la empresa?	How is the responsibility responsible for the organisation of business in the company ?
Ayer estuvimos comiendo juntos en el jardín	Yesterday , we built on the river in the park .
Quién poco coco come, poco coco compra	Low little , little of the _UNK , _UNK

A quién madruga, Dios le ayuda	In my love , God will help you
A Dios rogando y con el mazo dándo	God and the _UNK _UNK
Para cuándo podremos ver un personaje carismàtico?	How many can we see a _UNK ?

Experiment 3x1024 GRU

Paràmetre	**Configuració**
Fitxer de training	tfm.es-en.en; tfm.es-en.es
Fitxer de test	tfm_test.en; tfm_test.es
Nombre de paraules Espanyol	200000
Nombre de paraules Anglés	200000
Ràtio d'aprenentatge	0.5 inicial
Factor decreixent del ratio	0.99
Tamany RNN	3 capes de 1024 celles
Tamany batch	64
Cel·la	GRU
Algorisme Gradient Descent	Adagard
Máxim dades entrenament	4000000
Checkpoint	Cada 512 passos.

Resultats:

Temps entrenament:	14 hores 15 minuts.
Dades últim checkpoint	global step 62464 learning rate 0.3774 step-time 0.66 perplexity 5.19 eval: bucket 0 perplexity 3.37 eval: bucket 1 perplexity 3.12 eval: bucket 2 perplexity 6.04 eval: bucket 3 perplexity 8.30
Observacions	

Test

Hola	Votes
Perro	_UNK
Gato	_UNK
Ratón	Votes
Murciélago	_UNK
Había una vez un circo	There ' s a time after
Había cierto rencor en sus palabras	There was a certain word in his words
Mañana iremos al parque de atracciones	We will be at the park of the park
En un lugar de la Mancha de cuyo nombre no quiero acordarme	Instead of a name of name , I should not like to name
Érase una vez, en un lugar muy lejano	Once again , a very place

La economía está cada vez peor	The economy is increasingly worse .
Quién mató a Rogger Rabbit?	Did you have been killed ?
Pablito clavó un clavito. Qué clavito clavó Pablito?	_UNK _UNK _UNK _UNK _UNK _UNK _UNK _UNK _UNK _UNK ?
Me duele el codo	My hand is the
Me duele la cabeza	I ' m my head
Eran mas de cien y menos de mil	Most of them were less than 0 thousand thousand thousand thousand thousand thousand thousand thousand
Quinientas mil personas assistieron al mayor evento jamás visto	A thousand thousand thousand thousand thousand thousand thousand thousand thousand thousand thousand thousand thousand thousand
Quién és el presidente de los Estados Unidos?	Is the President of the United States ?
Tenemos que poner orden en la asamblea	We need to put a joint order at the lines
Hay que poner orden en la asamblea	We must be put to the joint body
Me dijeron que fuera directo al médico	I told that you were told to go to the doctor
Me dijeron que fuera directamente a urgencias	I told that it was directly related to
Felices fiestas!	May !
Feliz Navidad!	Happy Christmas !
Se hizo un corrillo alrededor de su mesa	A glass of the table was made below
Había montones de personas mirando el accidente de tráfico	There were people who were watching the traffic traffic traffic
Toda mala época llega a su fin	Every time of time is achieved
Cuando dos problemas se juntan es mas difícil encontrar una solución	When two problems are difficult to solve the solution .
Cuándo te lo dijeron?	When you told it ?
En esta época del año, invierno, las temperaturas son muy bajas	During the temperatures of this year , temperatures , temperatures are very low .
Alguien debe de hacerse responsable por las acciones ocurridas	Someone must be responsible for actions taken
Quién és el responsable de la organización de los eventos en la empresa?	Who will be responsible for the organisation of organisation involved in the organization ?
Ayer estuvimos comiendo juntos en el jardín	I had yesterday to eat at the garden
Quién poco coco come, poco coco compra	_UNK little , buy a little drink
A quién madruga, Dios le ayuda	Those who are suffering , God is helping
A Dios rogando y con el mazo dándo	God and the _UNK _UNK
Para cuándo podremos ver un personaje carismàtico?	When we can see a _UNK ?

Experiment 4x1024 LSTM

Paràmetre	**Configuració**
Fitxer de training	tfm.es-en.en; tfm.es-en.es
Fitxer de test	tfm_test.en; tfm_test.es
Nombre de paraules Espanyol	200000
Nombre de paraules Anglés	200000

Ràtio d'aprenentatge	0.5 inicial
Factor decreixent del ratio	0.99
Tamany RNN	4 capes de 1024 celles
Tamany batch	64
Cel·la	LSTM
Algorisme Gradient Descent	Adagard
Máxim dades entrenament	4000000
Checkpoint	Cada 512 passos.

Resultats:

Temps entrenament:	18 hores 20 minuts.
Dades últim checkpoint	global step 62464 learning rate 0.4008 step-time 0.84 perplexity 4.08 eval: bucket 0 perplexity 3.02 eval: bucket 1 perplexity 2.87 eval: bucket 2 perplexity 5.35 eval: bucket 3 perplexity 6.56
Observacions	

Test

Hola	Applause
Perro	_UNK
Gato	Heckling
Ratón	Heckling
Murciélago	_UNK
Había una vez un circo	A single game
Había cierto rencor en sus palabras	There was some of them in their words
Mañana iremos al parque de atracciones	We will go to the park of _UNK
En un lugar de la Mancha de cuyo nombre no quiero acordarme	In a place of the name of which I am not supposed to know
Érase una vez, en un lugar muy lejano	You once again , at a very large point
La economía está cada vez peor	The economy is becoming worse !
Quién mató a Rogger Rabbit?	Who was killed to _UNK _UNK ?
Pablito clavó un clavito. Qué clavito clavó Pablito?	_UNK _UNK _UNK ? _UNK _UNK ?
Me duele el codo	My forehead hurt
Me duele la cabeza	My head gets up
Eran mas de cien y menos de mil	They were more than 000 and less thousand
Quinientas mil personas assistieron al mayor evento jamás visto	_UNK a few people die to the last never ever
Quién és el presidente de los Estados Unidos?	Are you the president of the United States ?
Tenemos que poner orden en la asamblea	We need to set the order in the session
Hay que poner orden en la asamblea	The order of order needs to be established in the session
Me dijeron que fuera directo al médico	I was told that I was directly to the doctor
Me dijeron que fuera directamente a urgencias	I told you to go directly to contact

Felices fiestas!	Long games !
Feliz Navidad!	Christmas Christmas !
Se hizo un corrillo alrededor de su mesa	You got a lot of your table
Había montones de personas mirando el accidente de tráfico	There were a number of people watching traffic traffic
Toda mala época llega a su fin	All the bad times came to the end
Cuando dos problemas se juntan es mas difícil encontrar una solución	When a couple of problems are found , it is difficult to find a solution
Cuándo te lo dijeron?	Did they told you ?
En esta época del año, invierno, las temperaturas son muy bajas	In this year of the year , winter , temperatures are very low
Alguien debe de hacerse responsable por las acciones ocurridas	Anyone must be responsible for the actions committed
Quién és el responsable de la organización de los eventos en la empresa?	Is the responsibility of the organization of the business in the company ?
Ayer estuvimos comiendo juntos en el jardín	We were in the first days in the garden
Quién poco coco come, poco coco compra	_UNK , _UNK , _UNK
A quién madruga, Dios le ayuda	To whom you will , God will give you
A Dios rogando y con el mazo dándo	A God _UNK and _UNK _UNK
Para cuándo podremos ver un personaje carismàtico?	How can we see a _UNK ?

Experiment 4x1024 GRU

Paràmetre	**Configuració**
Fitxer de training	tfm.es-en.en; tfm.es-en.es
Fitxer de test	tfm_test.en; tfm_test.es
Nombre de paraules Espanyol	200000
Nombre de paraules Anglés	200000
Ràtio d'aprenentatge	0.5 inicial
Factor decreixent del ratio	0.99
Tamany RNN	4 capes de 1024 celles
Tamany batch	64
Cel·la	GRU
Algorisme Gradient Descent	Adagard
Máxim dades entrenament	4000000
Checkpoint	Cada 512 passos.

Resultats:

Temps entrenament:	26 hores 30 minuts.
Dades últim checkpoint	global step 62464 learning rate 0.4049 step-time 0.79 perplexity 4.93 eval: bucket 0 perplexity 3.71 eval: bucket 1 perplexity 3.61

	eval: bucket 2 perplexity 6.05 eval: bucket 3 perplexity 6.96
Observacions	

Test

Hola	_UNK
Perro	_UNK
Gato	_UNK
Ratón	_UNK
Murciélago	_UNK
Había una vez un circo	There was a while
Había cierto rencor en sus palabras	There was a certain word in her words
Mañana iremos al parque de atracciones	We will return to the _UNK de _UNK
En un lugar de la Mancha de cuyo nombre no quiero acordarme	On a place of the ' _UNK ' of the ' I do not want to call me
Érase una vez, en un lugar muy lejano	First time , in a very busy place
La economía está cada vez peor	The economy is at every moment .
Quién mató a Rogger Rabbit?	Who was the _UNK ?
Pablito clavó un clavito. Qué clavito clavó Pablito?	Do you have a _UNK ?
Me duele el codo	I ' m tired of the foot
Me duele la cabeza	I felt your head
Eran mas de cien y menos de mil	They were less than 000 and less than 000
Quinientas mil personas assistieron al mayor evento jamás visto	_UNK 000 people will never be born as ever known
Quién és el presidente de los Estados Unidos?	Who was the president of the United States ?
Tenemos que poner orden en la asamblea	We must put the Charter to be put out
Hay que poner orden en la asamblea	The sitting must be put up .
Me dijeron que fuera directo al médico	I told me to be direct to the doctor
Me dijeron que fuera directamente a urgencias	I told me to be directly directly to the emergency
Felices fiestas!	May !
Feliz Navidad!	Christmas !
Se hizo un corrillo alrededor de su mesa	A group was set around her table
Había montones de personas mirando el accidente de tráfico	There were hundreds of people who were looking on traffic traffic
Toda mala época llega a su fin	Every time of time is to go to the end
Cuando dos problemas se juntan es mas difícil encontrar una solución	When two problems are to be resolved , it is difficult to find a solution
Cuándo te lo dijeron?	Do you tell him ?
En esta época del año, invierno, las temperaturas son muy bajas	During this period of year , temperatures are very high levels .
Alguien debe de hacerse responsable por las acciones ocurridas	People must be responsible for the action of the following action
Quién és el responsable de la organización de los eventos en la empresa?	Would you be responsible for the organization of the organization ?
Ayer estuvimos comiendo juntos en el jardín	We were working together in the garden
Quién poco coco come, poco coco compra	_UNK is a little bit less than a drink
A quién madruga, Dios le ayuda	Who , God will help you
A Dios rogando y con el mazo dándo	God and with the _UNK _UNK

Para cuándo podremos ver un personaje carismàtico?	We can see a _UNK ' s ' ?

Experiment 3x1200 GRU

Paràmetre	**Configuració**
Fitxer de training	tfm.es-en.en; tfm.es-en.es
Fitxer de test	tfm_test.en; tfm_test.es
Nombre de paraules Espanyol	200000
Nombre de paraules Anglés	200000
Ràtio d'aprenentatge	0.5 inicial
Factor decreixent del ratio	0.99
Tamany RNN	3 capes de 1200 celles
Tamany batch	64
Cel·la	GRU
Algorisme Gradient Descent	Adagard
Máxim dades entrenament	4000000
Checkpoint	Cada 512 passos.

Resultats:

Temps entrenament:	40 hores 38 minuts.
Dades últim checkpoint	global step 62464 learning rate 0.3625 step-time 0.90 perplexity 4.98 eval: bucket 0 perplexity 3.72 eval: bucket 1 perplexity 3.70 eval: bucket 2 perplexity 6.47 eval: bucket 3 perplexity 9.39
Observacions	

Test

Hola	_UNK
Perro	_UNK
Gato	_UNK
Ratón	_UNK
Murciélago	_UNK
Había una vez un circo	There was a certain once once
Había cierto rencor en sus palabras	There is no doubt in his words
Mañana iremos al parque de atracciones	Tomorrow we will be located on the park park
En un lugar de la Mancha de cuyo nombre no quiero acordarme	On behalf of the name of the name of the name of which I don ' t want to dwell
Érase una vez, en un lugar muy lejano	Once again , in a very great way
La economía está cada vez peor	The economy is increasingly worse
Quién mató a Rogger Rabbit?	_UNK _UNK _UNK _UNK ?

Pablito clavó un clavito. Qué clavito clavó Pablito?	Will you choose a _UNK , what is a _UNK _UNK ?
Me duele el codo	My leg is cut up
Me duele la cabeza	I hit your head
Eran mas de cien y menos de mil	More than one thousand thousand thousand
Quinientas mil personas assistieron al mayor evento jamás visto	We have a thousand thousand thousand thousand people to date the world
Quién és el presidente de los Estados Unidos?	Is the United States ' s president ?
Tenemos que poner orden en la asamblea	We must order order in order
Hay que poner orden en la asamblea	We must order order to order the order .
Me dijeron que fuera directo al médico	I told me that he was direct to the doctor
Me dijeron que fuera directamente a urgencias	I told me that he was directly directly related to
Felices fiestas!	Long holiday !
Feliz Navidad!	Christmas Christmas !
Se hizo un corrillo alrededor de su mesa	A man around his table
Había montones de personas mirando el accidente de tráfico	There were people with the accident of traffic traffic
Toda mala época llega a su fin	All bad times is at times for their beginning
Cuando dos problemas se juntan es mas difícil encontrar una solución	When two problems are running , it is easier to find a solution
Cuándo te lo dijeron?	Did you told him ?
En esta época del año, invierno, las temperaturas son muy bajas	In this year of winter , winter , winter are very high
Alguien debe de hacerse responsable por las acciones ocurridas	One of the actions for the actions for the actions for the actions of the
Quién és el responsable de la organización de los eventos en la empresa?	Is the responsible organization for the organization in the company ?
Ayer estuvimos comiendo juntos en el jardín	We died yesterday in the garden
Quién poco coco come, poco coco compra	_UNK little little little little little little little little little little
A quién madruga, Dios le ayuda	You who is giving God , God is giving him
A Dios rogando y con el mazo dándo	A God and with the _UNK _UNK
Para cuándo podremos ver un personaje carismàtico?	How can we see a character of having a character ?

Annex 2 – Feina realitzada fins a la data

En un inici es van mirar tots els llocs web que hi havien per explicar el que eres les Xarxes Neuronals i el DeepLearning. Sobretot es va intentar estudiar el que es pot trobar a la Wikipedia. En realitat, aquest pot ser el mes gran dels errors que he comés ja que la información no és estructurada, si es comparen pàgines com al de Recurrent Neural Networks amb la de Recursive Neural Networks, la informació no és consistent (i és que no ho són, però es que quan no saps, és molt complicat extreure la información correcta).

També es van començar a mirar exemples a webs i eines que faciliten el desenvolupament de xarxes neuronals. En aquest moment es va trobar l'exemple de traductor a Tensorflow i semblava que encara hi havia possibilitats per a la millora segons ells mateixos. Així que amb el poc que havia aprés més

aquest exemple que només calia adaptar a un altre idioma i procurarse datasets es va decidir fer el traductor Espanyol - Anglés.

Així que en primer lloc es va decidir obtenir un llarg dataset. Com mes gran millor. La estrategia era obtenir el major nombre de paraules en espanyol posible i fer una cerca a un conegut traductor online per cadascuna de les paraules. Aquestes paraules es van trobar a la web http://www.listapalabras.com/, que suposadament té 650000 paraules vàlides en espanyol incloent conjugacions. La llista de paraules va ser fácil de obtenir. No gaire mes tard, es va fer un crawler per a obtenir les traduccions per cadascuna de les paraules. Aquest és un altre tema apart doncs la página de traduccions disposa de proteccions per a evitar el crawling així que es va haver de fer servir la xarxa TOR per poder simular ser una adreça diferent cada cop que la adreça era invalidada per el servidor. Això feia que el procés de crawling anés molt a poc a poc. Es pot dir que el crawling va començar el 6 d'octubre i va acabar el 24 de Novembre (exactament fa 4 dies). El crawling va estar funcionant de mitja unes 20 hores diaries (menys quan es feien les proves per esbrinar el funcionament de la xarxa). El resultat es que es van obtenir unes 9 milions de sentències, amb la llàstima de que 7800000 eren repetides. El cas es que per les conjugacions, moltes vegades es retornava les mateixes frases. Així que en la majoria de vegades, comptava amb repeticions. També tenia el problema afegit de que si la paraula era prou extranya, l'ordre de de les sentencies en comptes de ser espanyol- angles, era angles-espanyol. Així que amb aquestes dues premises es va escriure un programa que eliminava les repeticions i re-assignava ordenava les frases si aquestes no tenien un raonable de paraules en espanyol basada en la llista de listapalabras.

Mentre es feia el crawling, molts dies es van passar intentant configurar la GPU de la màquina donat que era impraticable fer servir la RAM i la CPU. El cas es que un cop es va poder instalar una Ubuntu amb les llibreries de nvidia es va poder comprobar que la GPU es com 20 vegades més rápida, però més limitada de memoria. Així que era inevitable obtenir més potencia de computació com fos, o no seria capaç de tenir un traductor acceptable (segons els càlculs que vaig fer, un dataset de 22 milions de frases podía trigar 9 dies a ser processat amb una GPU fins que la perplexitat era menor a 10).

No tot va ser tan directe, doncs quan es provava l'exemple proposat per Tensorflow, podien passar 5 hores i ni tan sols havia començat a entrenar la xarxa. Un cop més, quan no es sap que es fa i no es diposa de documentació adecuada, es molt difícil saber el que està passant.

Finalment, parlant amb el consultor, vaig poder obtenir un recurs de la UOC per a poder executar xarxes de major tamany i amb més frases. En aquest cas el recurs comptava amb una tarja gràfica que si permetia execucions com les vistes als exemples proporcionat per l'equip de Tensorflow.

A partir de llavors, la màquina ha estat corrent gairebé sense descans per tal de poder executar 7 jocs de proves que han trigat entre 1 i dos dies en ser executats, sense grans complicacions.

Annex 3 – Programes

Això només es una llista dels programes que s'han fet a mes del traductor de Tensorflow adaptat i millorat.

Adjunt amb aquest document es poden trobar els següents fitxers:

Fitxer-Carpeta	Descripció
Translate_es_en	Adaptació del codi de tensorflow per a llegir altres fitxers que no siguin els de l'europarliament. També s'ha modificat l'algorisme de optimització per Adagard i forçat a que el ratio d'aprenentatge millori en cas de que la perplexitat empitjori en l'ultima iteració. No crec necessari esperar a que tres iteracions siguin pitjor. La convergencia pot no dur'se a terme.
CsvConsistentLanguage.py	Aquest programa només llegeix el fitxer de crawling i crea un de espanyol i un de angles asseguant que no hi ha repeticions i que les frases están en el fitxer correcte (donat que poden estar invertides)
linguee_crawler_and_storage.py	Crawler que obté sentencies de una página eb de traduccions. Per executarlo es necessari tenir i el fitxer proporcionat per listapalabras. TorBrowser instalat i en execució.
listapalabras_crawler_and_storage.py	Fa crawling de la pàgina web listapalabras.com i desa les paraules un en fitxer csv.
SpanishEnglish.py	Per al dataset de tatoeba es necessari posar les frases en fitxers separats. Aquest programa navega pel dataset obtenint les traduccions i desant-les en dos fitxers. Per a anglés i espanyol.