Generació automàtica de diàlegs de còmic

Alumne (Menció de computació):
Sergi Masip Cabeza
hello@sergimasip.com

Tutor:
Ernest Valveny Llobet
ernest@cvc.uab.cat

Universitat Autònoma de Barcelona (UAB) / Escola d'Enginyeria (EE)
Juny de 2022

Resum

En els últims anys, els models de generació de llenguatge han estat millorant a un ritme accelerat. Aquestes millores han arribat també als models multimodals que treballen amb text i imatges, amb els quals se solen dur a terme tasques de descriure imatges o respondre preguntes sobre elles. En aquest projecte, però, s'han fet servir models d'aquest tipus amb l'objectiu últim de generar diàlegs de còmic. Per a aconseguir-ho, s'ha fet servir la base de dades proporcionada a COMICS. Primer s'ha plantejat la tasca de predir el següent diàleg d'entre un conjunt de candidats (*Text cloze*) donat un context de 3 panells i avaluar l'eficàcia d'aquests models comparant-los amb els resultats assolits a COMICS. Llavors, s'ha entrenat el millor d'aquests models per a la generació de diàlegs. Els resultats quantitatius mostren que els models de *Text cloze* superen al model hi-LSTM de COMICS. A més, es proposa un model generatiu en aquesta tasca el qual és capaç de generar següents diàlegs amb una adequació al context limitada, tot i obtenir uns valors baixos a les mètriques i contenir errors induïts per la qualitat de la base de dades.

Paraules clau– comics, visual storytelling, generació de llenguatge, transformer, t5, vl-t5

Abstract

Recently, language generation models have improved at an accelerated pace. Along with them, multimodal models that work with text and images have also improved. These models are usually used to perform tasks such as image captioning or visual question answering. However, in this project, we are using these models to generate comic dialogues. To do this, we used the dataset provided in COMICS. First, we proposed predicting the following dialogue from a set of candidates given the 3 previous panels as a context (Text cloze) as the first task and evaluating the effectiveness of these models by comparing them with the results achieved in COMICS. Then, we trained another based on the previous one for generating dialogues. The quantitative results show that the Text cloze models outperform the hi-LSTM model used in COMICS. In addition, we propose a generative model in this task, which is able to generate subsequent dialogues that fit the context to a certain extent, despite obtaining low values in the metrics and containing errors induced by the quality of the text transcriptions.

Keywords– comics, visual storytelling, language modeling, transformer, t5, vl-t5

1 Introducció

En els últims anys, han començat a proliferar dins el món de l'aprenentatge profund models molt capaços quant al processament de text gràcies als mecanismes d'autoatenció (*self-attention*). Aquests permeten als models aprendre les relacions de cada paraula amb la resta i, així, aprendre a predir la probabilitat de la següent paraula donada una seqüència prèvia. No obstant això, no només es fan servir únicament en text, sinó que existeixen models que utilitzen autoatenció combinant imatges i textos alhora. Aquests models prenen les característiques d'una imatge (*visual features*), en comptes d'un context en forma de text, i la representació d'un text (*embedding*) per a dur a terme la tasca per a la qual siguin entrenats. D'entre les tasques que existeixen, s'hi troben la generació de comentaris o descripció donada una imatge (*Image Captioning*), o la generació d'una petita història a partir d'una imatge (*Visual Storytelling*). Són aquestes dues últimes tasques en les quals s'inspira fonamentalment aquest treball.

En aquest projecte es proposa resoldre la tasca de generació automàtica de diàlegs de còmic. Per a aconseguir-ho, com a primer pas, es planteja la tasca d'escollir l'opció correcta d'entre un conjunt de candidats (*Text cloze* o elecció múltiple) donat un context, en primera instància, textual i, posteriorment, un de multimodal (text i imatge). Veieu la figura 1. Finalment, es planteja la tasca de generació de text (*language modeling*) per tal de generar els diàlegs. Per tant, s'ha hagut de desenvolupar i implementar un model capaç d'extreure les característiques del text dels diàlegs i de les imatges dels panells, i, amb elles, de generar el següent diàleg.

Per a resoldre ambdues tasques, s'han proposat 3 models basats en l'arquitectura del T5 [19], un model de llenguatge basat en autoatenció, fent ús de paràmetres preentrenats. Les entrades del primer model consisteixen en un context format per un conjunt de panells de còmic amb els 3 diàlegs de cadascun d'ells. En el segon i el tercer model s'hi afegeixen les característiques visuals dels panells. Els dos primers mo-

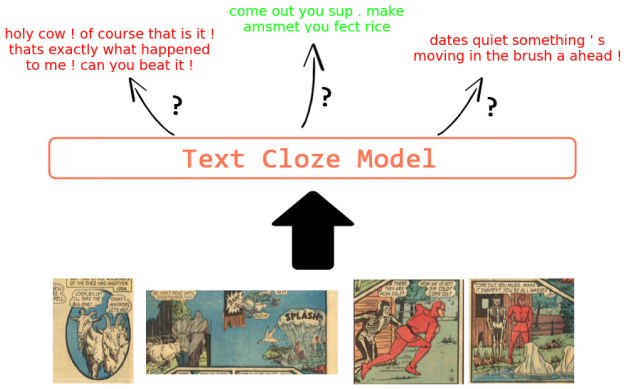

Figura 1: Exemple de la tasca d'elecció múltiple. El model, donat els tres primers panells com a context, haurà d'escollir el diàleg del següent panell d'entre el conjunt de candidats que se li presenta (en aquest cas, 3). Els candidats vermells són incorrectes mentre que el verd es correspon amb el del quart panell, és a dir, el correcte. En el cas de només text, únicament se li passen els textos de les bafarades, mentre que en el multimodal s'hi afegeixen les característiques visuals dels 4 panells també.

dels s'han entrenat en la tasca d'elecció múltiple, on se'ls ha fet escollir el següent diàleg d'entre un conjunt de diàlegs candidats. El tercer model, basat en el VL-T5 [6], una arquitectura lleugerament modificada del T5 per a acceptar entrades multimodals de text i imatge; s'ha entrenat en la tasca de generar el següent diàleg d'un còmic.

Per a avaluar l'eficàcia dels primers 2 models en la tasca d'elecció múltiple s'ha fet servir l'*accuracy* com a mètrica. Els experiments duts a terme amb aquests models són especialment interessants pel fet que permet quantificar la capacitat del model d'entendre el context, a més de la informació que es pot inferir comparant els resultats d'ambdós. Els resultats mostren que la informació visual té un impacte significatiu en el rendiment del model en la tasca objectiu. Pel que fa a l'últim model, s'han fet servir les mètriques GLEU [14] i METEOR [3] que es descriuran a la secció 6.1.

2 Treball relacionat

Existeixen diferents tasques que tracten de generar text a partir d'una imatge d'entrada. Entre les més populars hi són l'*Image Captioning* i la generació d'una resposta donada una imatge i una pregunta sobre aquesta (*Visual Question Answering* o *VQA*). Aquestes tasques es basen a generar una predicció literal i concreta de la imatge passada com a entrada. És a dir, si a una imatge hi apareixen dues persones assegudes xerrant, el model serà entrenat per a generar una descripció o una resposta tal que *"hi ha dues persones assegudes xerrant"*. En el context d'una història visual en la qual la imatge acompanya al text, però, aquesta descripció literal no serviria. El text generat hauria de tenir un estil més narratiu i creatiu, seguint l'exemple: *"el Marc i la Maria s'acabaven de conèixer, però el temps els hi va passar volant"*. En aquest treball es busca generar text amb aquest enfocament creatiu.

Per a la generació de text predominen dos tipus d'arquitectures, les Xarxes Neuronals Recurrents (RNN), i els Transformers [26]. Les primeres són un tipus de xarxa que permet fer servir les sortides anteriors com a entrades. La segona arquitectura basa el seu funcionament en mecanismes d'autoatenció i és actualment l'estat de l'art dins l'àmbit del Processament del Llenguatge Natural. Fins fa uns pocs anys, es feien servir RNN com la *Long Short Term Memory* (LSTM) [10], però amb l'arribada del Transformer, la tendència va canviar per dos motius: les RNN no es poden paral·lelitzar mentre que els Transformer sí i aquests, a més, obtenen millors resultats.

Algunes arquitectures força recents que es basen en el Transformer i el milloren d'alguna forma són el Switch-Transformer [9] i el T5 [19]. La primera reemplaça la capa densa situada després de la d'autoatenció per diferents capes expertes que s'especialitzen en una tasca en concret, la qual cosa permet entrenar models encara més massius però amb la mateixa capacitat de còmput. L'última és una arquitectura que millora el disseny base del *Transformer* fent que en totes les tasques de llenguatge natural les entrades i les sortides siguin text i tinguin un mateix format, i també canvia la representació de la posició sinusoïdal per una de relativa. Ha demostrat funcionar molt bé, amb adaptacions com en el VL-T5 [6], en tasques multimodals de text i imatge, com són *Image Captioning* o *VQA*, aconseguint resultats considerats estat de l'art o propers. Per aquesta raó, el T5 i el VL-T5 han estat la principal opció en aquest projecte.

La tasca que cerca generar text amb un enfocament narratiu a partir d'imatges és l'anomenada *Visual Storytelling*, per a la qual existeix una base de dades de Microsoft[25] que consta d'imatges anotades amb descripcions d'aquest tipus. Aquesta tasca tracta de crear una petita història inspirada en una imatge donada. L'últim treball més destacat en aquesta tasca es tracta de *Latent Memory-Augmented Graph Transformer for Visual Storytelling*[17], en el qual introdueixen un nou model basat en *Transformers* [26] anomenat *LMGT (Latent Memory-Augmented Graph Transformer for Visual Storytelling)*, millorant així l'estat de l'art del moment que es basava en RNN.

Quant a l'anàlisi visual dels còmics, existeixen alguns treballs que han tractat de classificar els estils o els autors [8][28]. En el darrer, els autors han entrenat un model basat en una Xarxa Neuronal Convolucional (CNN) per a classificar, primer, les pàgines i, segon, les vinyetes segons els autors (cada autor es considera que té un estil únic), i finalment han visualitzat les característiques visuals. A [15] s'entrena un model multitasca i multimodal basat en *Transformers* el qual s'entrena amb les característiques visuals de les imatges i els textos extrets automàticament d'aquests (OCR).

Els còmics solen implicar converses en les quals els diferents interlocutors intercanvien més d'un parell de frases. La tasca concreta que tracta de generar aquests diàlegs s'anomena *multi-turn dialogue modeling*. Mal-

grat que els models preentrenats de generació del llenguatge en general funcionen força bé, com podria ser el mateix GPT-2 [18] o, més recentment, el GPT-3 [5]; també existeixen treballs que s'especialitzen en aquesta tasca [29] [27]. En aquests, presenten arquitectures basades en Xarxes Recurrents anomenades *Gated Recurrent Units (GRU)* [7] amb autoatenció específicament dissenyades per a la tasca.

Existeixen unes poques bases de dades de còmics o manga que comptin amb anotacions adequades per al projecte, és a dir, que comptin amb les anotacions de text de les bafarades. Una d'elles és la base de dades COMICS introduïda per [11], treball en el qual fan servir una LSTM jeràrquica (hi-LSTM) per a dur a terme tasques de coherència de personatges i de predicció del següent panell (*visual cloze*) o diàleg (*text cloze*) a partir de les característiques visuals dels panells, els *embeddings* i la posició al panell (*bounding box*) del diàleg objectiu. Una altra base de dades a tenir en compte és el Manga109 [1], la qual està formada per 109 mangues en japonès amb anotacions tant visuals i de característiques dels personatges com de text. Per a aquest treball, es farà servir la primera base de dades perquè els diàlegs estan en anglès i la major part dels models de llenguatge actuals estan preentrenats en aquest idioma.

Finalment, un treball molt relacionat amb la proposta és el de *Comic strips* [20], el qual basa el seu model en el proposat a *Stories for Images-in-Sequence* [24] aplicat a seqüències de vinyetes. L'objectiu en aquest cas és el de predir les frases dels personatges que apareixen a les vinyetes seqüencialment. Per a això, han entrenat un model on codifiquen els texts i les característiques visuals de les vinyetes anteriors fent ús de dos codificadors basats en GRU diferents, i fan servir un decodificador, també basat en GRU, per a generar la següent frase.

Per a aquest projecte s'ha optat per fer servir com a referència el treball realitzat a COMICS [11] per a la tasca de *text clozing* fent ús de l'arquitectura T5 [19] i de la VL-T5 [6]. D'aquesta manera, es tindrà una referència amb la qual poder comparar l'eficàcia dels models implementats i es podrà fer un desenvolupament esglaonat del treball fins a aconseguir el model generatiu final.

3 Objectius i metodologia

La llista d'objectius d'aquest projecte és la següent:

1. Entrenar un model per a la tasca de predir el següent diàleg d'entre un conjunt de possibles diàlegs donat un context previ de n diàlegs de m panells anteriors.

2. Entrenar un model per a la tasca de predir el següent diàleg d'entre un conjunt de possibles diàlegs donat un context previ de n diàlegs de m panells anteriors utilitzant també les seves característiques visuals.

3. Adaptar el model anterior per a dur a terme la modelització del llenguatge i entrenar-lo per a la generació automàtica de diàlegs.

La metodologia que s'ha emprat per a assolir les metes es basa en la filosofia *Agile* en combinació amb un model de desenvolupament en cascada per tal d'aprofitar el millor d'ambdues metodologies. D'aquesta manera, la planificació i el desenvolupament han girat al voltant de la seqüencialització de les tasques i subtasques definides, que a causa de la forta dependència de les tasques, s'han plantejat en cascada. Tanmateix, de la filosofia *Agile* es pren que el desenvolupament s'ha dividit en 3 esprints, que coincideixen amb les 3 fites importants del treball, en els quals s'ha mantingut una comunicació constant amb el tutor. Cada setmana s'han fet reunions amb ell per tal de revisar l'estat del projecte, definir noves pautes d'actuació i corregir les desviacions que han sorgit.

Durant cada esprint, s'han desenvolupat les subtasques definides per a cadascuna de les tasques principals, les quals no han tingut una data límit concreta a causa de la naturalesa de recerca del projecte, però sí una aproximació de les hores de treball que es van estimar. Al final de cada esprint, s'ha fet l'entrega del model final de cada objectiu i de l'informe de seguiment corresponent. La planificació s'ha definit en un diagrama de Gantt que es pot visualitzar a la figura 2.

El llenguatge de programació emprat en tot el projecte ha estat *Python* amb la llibreria d'aprenentatge automàtic *Pytorch*. Per tal de gestionar les diferents versions del codi i fer un seguiment d'aquestes, s'ha utilitzat l'eina *git* (consulteu el repositori). Per a l'entrenament dels models i els experiments, s'ha fet servir un ordinador amb una targeta gràfica NVIDIA GTX2080Ti i el clúster del departament del Centre de Visió per Computador (CVC).

Per a facilitar les execucions dels experiments i minimitzar les modificacions del codi necessàries, s'ha programat un entorn de treball modular i escalable que permet definir i executar diferents models amb diferents paràmetres modificant únicament certs fitxers de configuració. S'ha pres com a referència per a algunes parts del disseny de l'entorn un treball de refactorització del codi d'un projecte de ciència de dades realitzat per un enginyer de software [2].

4 Base de dades

La base de dades emprada per als experiments és la publicada a COMICS [11], la qual compta amb més de 2,5 milions de diàlegs extrets automàticament de 3958 còmics diferents fent servir el Google Cloud Vision OCR (2016). Aquest sistema d'OCR té un rendiment pobre, comparat amb els d'avui dia, que sol introduir errors en els textos que extreu. De fet, a COMICS ja alerten que aquesta base de dades podria no funcionar gaire bé per a tasques de generació de text. Es poden visualitzar alguns exemples en la figura C.4, en la qual es pot observar com en el candidat correcte introdueix artefactes com "styrene"o "torpee up", fet que, de vegades, impossibilita la comprensió del text.

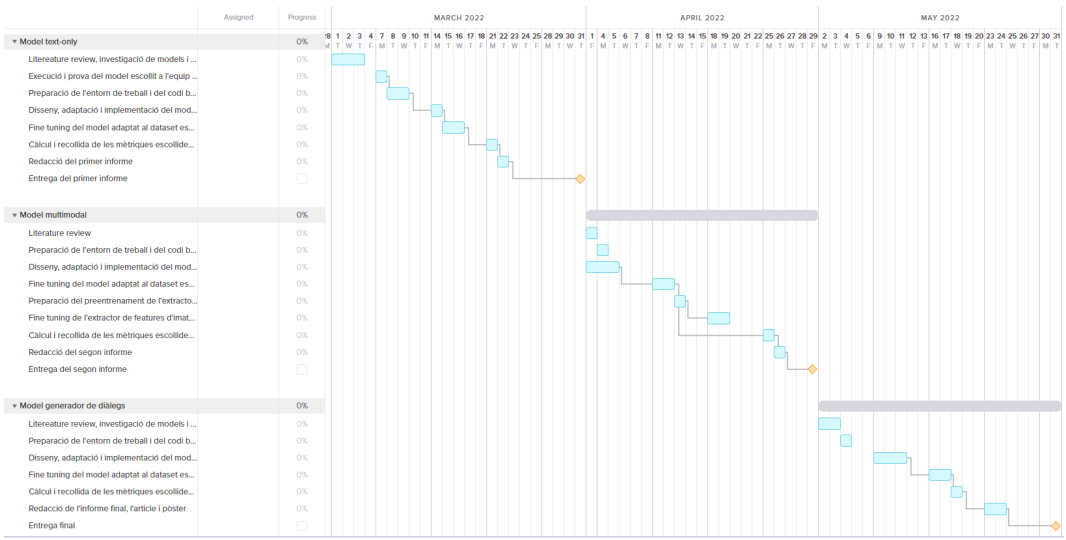

Figura 2: Diagrama de Gantt de la planificació inicial.

Els autors van processar aquestes dades i van crear 3 bases de dades, una per a cadascuna de les tasques que van proposar: elecció del diàleg següent d'entre un conjunt de candidats (*Text cloze*), elecció de la imatge del panell següent i coherència de personatge. De cadascuna d'aquestes van preparar dues variants amb el context i els candidats preseleccionats, que ells anomenen *fàcil* i *difícil*. En la fàcil, els candidats que no són correctes els escullen aleatòriament d'entre tots els diàlegs extrets de tots els còmics, mentre que en la difícil escullen els diàlegs de panells propers, que solen ser més semblants al correcte, de manera que és més difícil predir quin és el diàleg correcte. Finalment, van reservar 2929 còmics per al conjunt d'entrenament, 500 per al conjunt de validació i 500 més per al de test, mentre que els altres resten per a l'entrenament. Aquest treball s'enfocarà en la tasca de *Text cloze*.

A COMICS fan servir un (*tokenitzador*) simple, el qual separa per espais i signes de puntuació cada diàleg i n'extreuen així les paraules. Aquestes paraules s'assignen a un enter únic anomenat *token*. A partir d'aquests tokens extrets es va construir un vocabulari amb els 20.000 tokens més freqüents de tot el conjunt de dades. Aquest enfocament presenta un problema evident de pèrdua d'informació, ja que quan el *tokenitzador* troba una paraula que no hi és al vocabulari li assigna un token especial que correspon a *desconegut* (*UNK*). Per aquest motiu, els autors comenten que van descartar dels conjunts de validació i de test aquelles entrades que contenien massa tokens desconeguts.

En aquest projecte s'ha fet servir el *tokenitzador* propi de cada model, però tots es basen en algorismes que separen el contingut del text n-grames, és a dir, en seqüències de n caràcters, com és el *Byte-Pair Encoding* [22]. L'avantatge d'aquest enfocament respecte de l'utilitzat a COMICS és que ara les paraules, encara que no apareguin al vocabulari, es poden codificar com una seqüència de subparaules. No obstant això, s'ha decidit fer servir els conjunts de validació i test originals filtrats a l'hora d'avaluar els models per a poder fer la comparació.

Pel que fa a les característiques visuals que s'han afegit, s'han fet servir les característiques de 36 objectes representats amb un vector de 2048 posicions de cada panell extretes amb un model FasterRCNN [21] de la llibreria Detectron2, mentre que els autors originals de COMICS utilitzaven un únic vector de característiques de 4096 posicions per panell que el representa extret amb una VGG-16 [23]. Cal afegir que, abans d'extreure les característiques de cada panell, s'han tapat amb un rectangle negre les bafarades perquè els models no memoritzin directament el text.

5 Definició dels models

L'objectiu d'aquest projecte és el de dur a terme les tasques d'elecció múltiple de candidats i de generació del següent diàleg donat un context previ. Per a aquest, s'han considerat únicament els 3 panells previs i el panell del diàleg objectiu, i fins a 3 diàlegs com a màxim de cadascun dels panells. Per a dur a terme aquestes tasques, s'ha dividit el treball en 3 parts. En la primera, s'han provat diferents arquitectures per a fer la tasca d'elecció múltiple amb un context format per únicament el text dels diàlegs. En la segona part, s'ha fet el mateix però afegint-hi les característiques visuals i les *bounding box* de 36 objectes de cadascun de les imatges dels panells. Finalment, en la tercera part, s'ha modificat el millor model de la segona part per tal de generar el següent diàleg.

Així doncs, sigui $I = I_x \cup \{I_a\}$ el conjunt de panells utilitzats a una execució del model on $I_x = \{I_{x_1}, I_{x_2}, I_{x_3}\}$ és el subconjunt dels 3 panells considerats que corresponen al context i I_a el panell del diàleg a predir. Es defineix el conjunt de textos $t^x = \{t_1^{x_1}, t_2^{x_1}, t_3^{x_1}, ..., t_1^{x_3}, t_2^{x_3}, t_3^{x_3}\}$ a partir dels 3 primers diàlegs de cada panell del context I_x (si en té menys, se'ls hi assigna una cadena buida). Els 36 objectes considerats de cada panell del context I_x com a $o^x = \{o_1^{x_1}, o_2^{x_1}, ..., o_{35}^{x_3}, o_{36}^{x_3}\}$, els 36 objectes del panell candidat com a $o^a = \{o_1^a, ..., o_{36}^a\}$, i el text dels diàlegs candidats com a $t^a = \{t_1^a, t_2^a, t_3^a\}$ on un dels textos correspondrà al diàleg candidat correcte $t_c^a \in t^a$.

5.1 Text cloze (context de text)

L'objectiu dels models d'aquesta tasca és predir la probabilitat que cadascun dels candidats t^a sigui el correcte donats t^x i t^a com a entrada. Per tant, durant l'entrenament s'ha fet servir la *Cross Entropy loss* i per a optimitzar els paràmetres del model θ s'ha minimitzat la funció de *loss* del log-likelihood negatiu del candidat correcte t_c^a donat t^x i t^a:

$$\mathcal{L}_\theta = -logP_\theta(t_c^a|t^x, t^a)$$

D'entre les diferents proves que s'han dut a terme amb diferents models i configuracions per a aquesta tasca, en aquesta subsecció es comentaran els dissenys i configuracions dels models que han donat millors resultats d'entre les 3 arquitectures amb les quals s'ha experimentat. Les dues que més sobresurten i que es comentaran en aquest informe es basen en un model T5 [19] preentrenat.

5.1.1 T5 jeràrquic

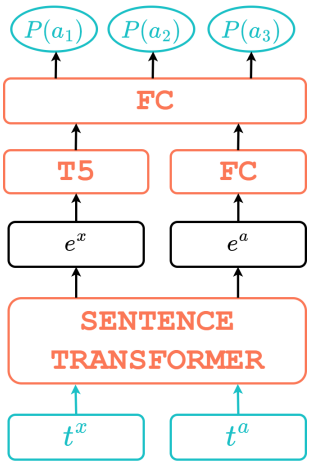

Figura 3: **Arquitectura del model T5 jeràrquic** el qual utilitza *sentence transformers* per a obtenir la representació textual dels diàlegs i seguidament passa aquesta representació per dos mòduls. Els contextos són processats al T5 mentre que els candidats són processats per la línia densa. Finalment, la sortida d'ambdós mòduls es concatena i es projecta per a predir la probabilitat de cadascun dels candidats de ser el diàleg correcte.

El model T5 jeràrquic basa la seva arquitectura (figura 3) en una actualització del model emprat a COMICS per a la tasca de *text-cloze*. L'actualització consisteix, principalment, en el reemplaçament dels dos mòduls de *LSTM* per mòduls basats en *Transformers*. En primer lloc, s'ha afegit un *sentence-transformer*[1] preentrenat que extreu una representació de cadascun dels diàlegs d'entrada t^x i dels candidats t^a. En segon lloc, s'ha introduït un T5 de mida petita amb només el bloc codificador que rep les representacions del context e^x i una capa densa (FC) que rep les representacions dels candidats e^a. La idea del codificador T5 és que rebi la representació de cada frase generada pel *sentence-transformer* i entengui les relacions entre totes elles. La capa densa, llavors, projecta les característiques dels candidats sobre la representació generada pel codificador del context. Aleshores, les sortides d'aquests dos últims mòduls es concatenen i passen per una última capa d'elecció múltiple, és a dir, una capa densa que retornarà les probabilitats per a cada candidat que sigui el correcte.

[1] Un *sentence-transformer* és un *Transformer* entrenat per a extreure una representació d'una frase sencera i no de cada paraula.

5.1.2 T5 base

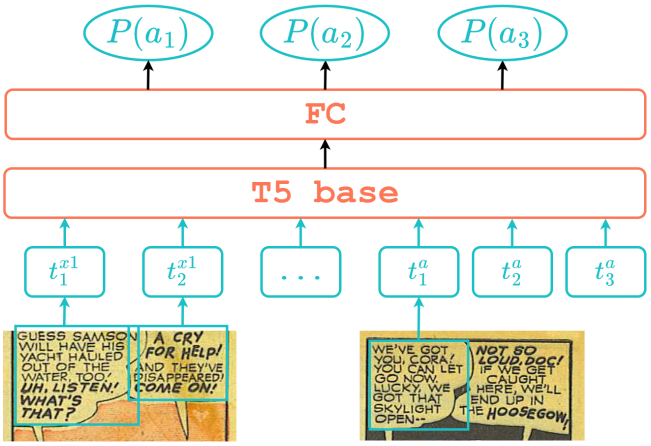

Figura 4: **Arquitectura del model T5 base** utilitzat a la modalitat de només text en la qual els diàlegs passen pel codificador d'un T5 de mida base i la representació resultant passa per una capa densa per a obtenir la probabilitat de cadascun dels candidats.

L'arquitectura del model basat en el T5 base (figura 4) consisteix, primordialment, en el bloc codificador del T5 de mida base al qual se li ha afegit una capa d'elecció múltiple. A diferència del model anterior, en aquest es manté la capa densa que extreu la representació del text que incorpora el T5. Llavors, l'entrada d'aquest model consisteix en la concatenació dels textos del context t^x i dels candidats t^a tal que $t_{input} = [t^x; t^a]$.

5.2 Text cloze (context multimodal)

El model multimodal es basa en la mateixa arquitectura que el de l'anterior secció, però aquest cop afegint-hi les característiques visuals del context i del panell a predir.

L'objectiu del model és predir la probabilitat que cadascun dels candidats t^a sigui el correcte donats t^x, o^x, o^a i t^a com a entrada. Per tant, durant l'entrenament s'ha fet servir la *Cross Entropy loss* i per a optimitzar els paràmetres del model θ s'ha minimitzat la funció de *loss* del log-likelihood negatiu del candidat correcte t_c^a donat t^x, o^x, o^a i t^a:

$$\mathcal{L}_\theta = -logP_\theta(t_c^a|t^x, o^x, o^a, t^a)$$

L'arquitectura final utilitzada en aquesta tasca és la del VL-T5 [6] (figura 5), el qual és un model de generació de llenguatge basat en el T5 estès perquè accepti una entrada multimodal de text i imatge (en forma de característiques i *bounding box* de 36 objectes extrets amb una FasterRCNN [21]). Aquest model ha estat preentrenat en diferents tasques que involucren text i imatges fent servir dues bases de dades: COCO [13] i Visual Genome [12]. Els canvis efectuats per adaptar-lo a la tasca han estat canviar el paràmetre que indica el nombre màxim d'imatges d'entrada de 2 a 4 i el del màxim nombre de tokens generats a la seqüència de sortida a 2 per tal de predir únicament l'índex del candidat correcte (l'últim token correspon al fi de seqüència *</s>*).

5.3 Generació del següent diàleg

L'arquitectura del model de generació del següent diàleg té, essencialment, la mateixa arquitectura el de l'anterior (figura 5), només s'ha canviat el mètode d'entrenament tal com s'explicarà en la secció 6 més endavant.

L'objectiu del model és generar el següent diàleg predient la seqüència de tokens y donats t^x, o^x i o^a com a entrada. Aquest següent diàleg és el que va en lloc del diàleg pres com a referència (en la tasca anterior, el considerat com a diàleg correcte t_c^a). Per tant, durant l'entrenament dels paràmetres del model θ s'ha minimitzat la funció de *loss* del log-likelihood negatiu dels tokens y donat t^x, o^x, o^a:

$$\mathcal{L}_\theta = -\sum_{i=1}^{|y|} logP_\theta(y_i|y_{<i}, t^x, o^x, o^a)$$

En aquest cas, s'ha limitat el nombre de tokens generats a la seqüència de sortida a 60 durant les fases de validació i test.

6 Experiments

En la cerca dels models de text i multimodal que funcionessin i obtinguessin un *accuracy* igual o superior a l'aconseguit pels autors de COMICS en la tasca d'elecció múltiple del següent diàleg, s'han fet un seguit d'experiments amb diferents arquitectures i paràmetres.

Quant als models de només text, a més de provar modificacions de l'arquitectura, addicionalment s'han provat diferents hiperparàmetres tals com la mida del model, el nombre de paràmetres d'algunes capes o el percentatge de les capes de *dropout*[2].

Pel que fa als hiperparàmetres d'entrenament, s'han provat diferents algorismes optimitzadors, ràtios d'aprenentatge (lr) i mides del *batch*, és a dir, el nombre de mostres amb el qual es fa la mitjana de la *loss*. Els hiperparàmetres utilitzats en l'entrenament dels models de només text, i de la resta, es poden consultar a les pàgines de l'apèndix A.

[2]El *dropout* és una tècnica regularitzadora en què, durant l'entrenament, s'ignoren alguns nodes d'una capa per ajudar el model a generalitzar millor.

A les tasques d'elecció múltiple, s'ha entrenat cada arquitectura en les dues modalitats de la base de dades (*easy i hard*) i s'han avaluat en ambdues. A la tasca de generació només s'ha entrenat un model en la base de dades corresponent a la modalitat *easy*, ja que només es té en compte el candidat correcte.

Per als entrenaments dels models de Text Cloze de només text s'ha fixat un màxim de 50 epochs i per als generatius de 10 epochs. No obstant això, seguint la metodologia de l'*early stopping*, monitorant la *loss* sobre el conjunt de validació l'entrenament s'ha aturat un cop la millora del model no era significativa. Això ha estat especialment rellevant en els experiments dels models de la tasca de *Text cloze*, mentre que en el generatiu s'ha arribat al límit de les 10 epochs.

L'entrenament del model multimodal s'ha dut a terme, en un primer moment, a partir dels hiperparàmetres utilitzats al model de només text. No obstant això, pel fet que el model basat en el T5 base no ha millorat els resultats, en l'entrenament del model multimodal final s'han fet servir els mateixos hiperparàmetres que els que han emprat els autors del VL-T5 [6] canviant el nombre d'imatges d'entrada que accepta el model 2 a 4. La resta de hiperparàmetres es pot trobar a la taula A.1).

L'entrenament del model de generació de diàlegs també ha utilitzat els hiperparàmetres del VL-T5 [6] amb els mateixos canvis que el model anterior (veieu la taula A.1) i, aquest cop, el diàleg de referència s'ha truncat a 60 tokens. En aquest model, però, a diferència dels anteriors s'han tingut en compte les mètriques GLEU [14] i METEOR [3] (s'explicaran a la subsecció 6.1) en comptes de l'*accuracy*.

Per a la generació d'una seqüència de text se solen usar tècniques de cerca i existeixen diferents alternatives que poden alterar la qualitat del text generat pel model final. En tasques de *Captioning* és comú emprar una cerca de *beams*, la qual explora els nodes d'un conjunt limitat de n nodes, amb $n = 10$. A *Visual Storytelling* [25], però, van concloure que per a la tasca de generar una història era més eficaç fer una cerca àvida. Així doncs, un cop entrenat el model generatiu, s'ha provat la cerca de beams, la cerca àvida i la cerca àvida però escollint el següent token aleatòriament d'acord amb la seva distribució de probabilitats (*sampling*).

6.1 Mètriques del model generatiu

En tasques que impliquen la generació de text condicionada a una entrada, com les tasques de traducció o el *Image Captioning*, és comú l'ús de mètriques com la BLEU [16] o la METEOR [3]. La primera calcula la *precision*. És a dir, calcula per cada n-grama (seqüència de n caràcters) que hi ha a la referència, el nombre de vegades m que apareix a la predicció (amb un límit marcat pel nombre de cops que hi és a la referència m_{max}) dividit entre el nombre de n-grames a la predicció w_t tal que $P = \frac{m}{w_t}$.

En aquestes tasques se solen fer servir més d'una frase de referència per mostra o cossos de text, però en aquest treball es treballarà amb una sola frase de

Figura 5: **Arquitectura del model VL-T5** utilitzada a la modalitat multimodal de les tasques de *text cloze* i de generació de diàlegs. En la tasca de *text cloze* el token generat correspon a l'índex del candidat correcte (0, 1 o 2). En canvi, a la tasca de generació de diàlegs els tokens corresponen al diàleg generat.

referència per mostra. Com que la BLEU en aquests casos no és tan eficaç, en els experiments s'avaluarà la GLEU [14], una mètrica que es correlaciona amb la BLEU, però que soluciona aquest problema calculant la *recall* i la *precision* i retornant el mínim d'entre aquestes dues.

La METEOR amplia i millora la BLEU de forma que fa una correspondència entre els n-grames de la referència i els de la predicció i calcula la *precision* i la *recall* de manera que m passa a ser el nombre de n-grames de la predicció que corresponen a algun de la referència. També introdueix una penalització per l'ordre dels n-grames.

Aquestes mètriques funcionen bé per a saber si un text generat s'assembla al de referència i conté les mateixes paraules. No obstant això, quan es vol explicar una història a partir d'una imatge, per exemple, es pot fer amb diferents paraules, diferent ordre i, sovint, es poden narrar diferents fets que no tenen res a veure els uns amb els altres, però que encaixen en el context. Per tant, s'ha de tenir present que els resultats que retornin indicaran com de bé el model s'ajusta a la referència, però no la seva creativitat.

7 Resultats

7.1 Text cloze (context de text)

A la taula 1 es poden observar els resultats quantitatius dels experiments més rellevants duts a terme corresponents a la tasca de *text clozing*. També s'han afegit els resultats obtinguts pels autors de COMICS als seus experiments.

En primer lloc, es va plantejar una hipòtesi inicial tal que substituint els mòduls de *LSTM* per *Transfomers* milloraria substancialment els resultats respecte a la hi-LSTM de COMICS, però no ha estat aquest el cas. El model jeràrquic ha convergit molt de pressa i ha començat a memoritzar el conjunt d'entrenament (*overfitting*) massa d'hora, assolint així una *accuracy* al conjunt de validació més baix que els altres models. La raó d'això, podria ser la complexitat excessiva del model.

Arran d'això, es va definir una altra arquitectura més simple, la definida a la secció 5.1.2 però amb una mida del T5 petita. Aquest model no va obtenir resultats satisfactoris, ja que es produïa *overfitting* quan el model arribava al 51,55% d'*accuracy* en la modalitat *easy* i al 46,54% en la *hard* (veieu la taula 1). La hipòtesi que es planteja llavors és que el model no és prou complex per a generalitzar, però sí que ho és per a memoritzar-la. En canvi, amb el T5 de mida base, que consisteix en la mateixa arquitectura, però amb una representació interna més gran i amb més capes, aconsegueix superar la hi-LSTM de COMICS amb un *accuracy* en la modalitat *easy* del 66,36% i en la *hard* del 54,89%.

Finalment, s'han fet uns experiments per veure com afecta la dificultat de cada modalitat (*easy i hard*) de la base de dades. S'hipotetitzava que un model entrenat sobre la modalitat *hard* assoliria millors resultats en la base de dades *easy* que un entrenat en aquesta, ja que el primer s'hauria entrenat amb exemples més difícils i, per tant, seria capaç de generalitzar millor. En canvi, es pot observar com els models entrenats en la base de dades *easy* rendeixen gairebé igual en la *hard* que els entrenats en aquesta, mentre que entrenats en la *hard* no aconsegueixen tan bons resultats en la *easy*.

7.2 Text cloze (context multimodal)

Per a aquesta tasca amb context multimodal, s'han fet principalment dos experiments. El primer consistia en una evolució del millor model de la secció anterior en què es passaven les característiques visuals extretes amb *BERT Pretraining of Image Transformers* (BEiT) [4] en forma de vector de 2048 posicions passat per una capa de *embedding* per tal de traslladar-lo a l'espai de característiques dels tokens de text. Aquesta primera aproximació no ha donat bons resultats i, de fet, els resultats al conjunt de validació indicaven que empitjorava lleugerament els resultats respecte del model de només text. Probablement, un vector de característiques per a cada panell és insuficient, a més a més, els pesos utilitzats en el T5-base han estat optimitzats en una base de dades formada únicament per text.

Aleshores, s'ha optat per adaptar el codi del model VL-T5 [6] a l'entorn, de forma que s'ha programat un script per a carregar la base de dades en el format que fan servir (concatenació de tot el text d'entrada separat per etiquetes especials i les característiques visuals per separat). S'han fet servir els pesos preentrenats proporcionats pels autors.

Inicialment, es creia que pel fet que l'extractor de ca-

Model	Text	Imatge	BBDD d'entrenament	BBDD testada	Accuracy
hi-LSTM	✓	✗	easy	easy	63.40
hi-LSTM	✓	✗	hard	hard	52.90
hi-LSTM	✓	✓	easy	easy	68,60
hi-LSTM	✓	✓	hard	hard	61,00
T5 jeràrquic	✓	✗	easy	easy	46,67
T5 jeràrquic	✓	✗	hard	hard	46,85
T5 small	✓	✗	easy	easy	51,55
T5 small	✓	✗	hard	hard	46,54
T5 base	✓	✗	easy	easy	66,36
T5 base	✓	✗	easy	hard	52,30
T5 base	✓	✗	hard	easy	62,88
T5 base	✓	✗	hard	hard	54,89
VL-T5	✓	✓	easy	easy	**79,67**
VL-T5	✓	✓	easy	hard	66,05
VL-T5	✓	✓	hard	easy	76,03
VL-T5	✓	✓	hard	hard	**69,76**

Taula 1: **Resultats de les execucions dels models de la tasca de *Text cloze***. Els models T5 base entrenats en les dues modalitats de la base de dades milloren lleugerament al hi-LSTM de COMICS. Pel que fa als models VL-T5, superen de forma significativa al hi-LSTM en ambdues modalitats.

racterístiques visuals usat ha estat entrenat amb imatges del món real, no s'assolirien uns resultats raonablement superiors als obtinguts a COMICS. No obstant això, s'ha aconseguit una *accuracy* 11 punts superior en la modalitat *easy* i 8 punts en la modalitat *hard* (veieu la taula 1, files 13 i 16). És destacable com la millora relativa respecte del model de només text és gairebé el doble que la que assoleixen a COMICS (+4,8 vs. +13,31 punts en la modalitat *easy* i +8,1 vs. +14,87 punts en la modalitat *hard*). Aquesta millora podria deure's al millor rendiment dels models actuals i a l'ús d'un conjunt de característiques visuals més ric que el que fan servir a COMICS.

Des d'un punt de vista d'anàlisi qualitativa, després de visualitzar diferents mostres del conjunt de test, es creu que el model falla principalment per dos motius. Un d'ells és la qualitat de l'OCR, tal com es pot observar a la figura C.2 de l'apèndix. Alguns artefactes fruit d'un error de detecció, com el "amsmet"que hi apareix al diàleg correcte de la figura, es repeteixen en altres mostres i podria ser un motiu de confusió per al model. L'altre motiu és el fet que alguns diàlegs, tot i no ser els correctes, podrien encaixar en la vinyeta a predir com és el cas de la figura C.4.

7.3 Generació del següent diàleg

Mètode de cerca	GLEU	METEOR
10 beams	1,15	3,20
Àvida	2,22	5,87
Àvida amb aleatorietat	**2,71**	**6,60**

Taula 2: **Resultats de les execucions a la base de dades *easy* del model generatiu VL-T5** amb diferents mètodes de cerca del següent token. S'ha fet servir $seed = 42$.

Per al model generatiu s'ha fet un únic entrenament amb els hiperparàmetres comentats a la secció 6 i fent servir, un cop més, els pesos proporcionats pels autors del VL-T5 [6]. Un cop acabat l'entrenament s'ha procedit amb l'avaluació del model sobre el conjunt de test.

Els resultats quantitatius obtinguts després de provar diferents mètodes de cerca es poden veure a la taula 2. Així mateix, es van analitzar qualitativament els resultats dels diferents mètodes i es va veure que en la cerca àvida amb aleatorietat els resultats qualitatius també eren millors en comparació amb la resta de mètodes.

Com es pot observar, s'aconsegueixen unes mètriques amb valors baixos, sobretot si els resultats es comparen amb els d'altres treballs relacionats. A *Visual Storytelling* [25] assoleixen una METEOR de 27,76 amb el mètode de la cerca àvida, i a *Comic strips* [20] obtenen una precisió BLEU-3 de 31,35.

En l'anàlisi qualitativa s'han detectat paraules mal escrites o fora de context en alguns casos. Això pot ser degut a la qualitat de l'OCR emprat pels autors de la base de dades. Malgrat això, com es pot observar en els exemples de l'apèndix C es pot veure que hi ha algunes mostres sense gaires o cap error. A banda de les faltes d'ortografia o gramaticals, hi ha paraules com "hey"o, fins i tot, frases com "hey, mr. hsh"que es tendeixen a repetir-se a moltes mostres. També sembla que els textos es tallen a mitja generació, fet que pot ser causat per la limitació del diàleg de referència a 60 tokens durant l'entrenament.

Pel que fa a la comprensió del context per part del model, sembla que el comprèn fins a cert punt. Destaca l'exemple de la figura C.2 que sí que obté unes mètriques més acceptables comparades amb els altres treballs. A la figura C.3 el text generat no segueix correctament la seqüència, però sí que té relació amb el context. Aquest fenomen es repeteix en altres mostres,

la qual cosa podria indicar que el model entén correctament el context, però, de vegades, li costa seguir la seqüència d'esdeveniments.

En altres ocasions, sembla que el model prioritza en excés el context visual. Per exemple, a la figura C.6 el model genera un diàleg que es pot considerar violent tot i que l'escena no ho és, potser pel fet que hi apareixen soldats a les imatges.

També s'ha explorat si les mètriques quantitatives són una bona referència i els resultats han mostrat que, en alguns casos, no. Per exemple, en la figura C.4 en un primer moment podria semblar que el text no encaixa amb l'escena, de fet té una GLEU de 2,94 i una METEOR de 0,00; però si es para atenció a l'escena i el context, sí que podria tenir lògica que el personatge estigui buscant a algú (malgrat que la frase sí que sembla memoritzada tal com s'ha comentat al paràgraf anterior). En canvi, a la figura C.8 el model obté una GLEU de 16,67 i una METEOR 40,32, però el text no té gaire coherència amb el context.

D'aquesta anàlisi qualitativa se'n poden extreure tres conclusions principals. La primera és que fent *fine tuning* d'un model de llenguatge preentrenat sobre una base de dades amb una qualitat del text pobre a causa de l'OCR, s'obté un model que hereta part d'aquests errors i que podria limitar la seva capacitat de generació. La segona és que les mètriques quantitatives no sempre s'ajusten als resultats qualitatius i que cal el judici humà per a avaluar models d'un caire més narratiu i creatiu. La tercera és que el model té força marge de millora pel que fa a comprensió del context i la seqüència d'esdeveniments.

8 Conclusions

S'han assolit els objectius principals inicialment plantejats, excepte el *fine tuning* de l'extractor de característiques visuals. En aquest projecte de recerca s'han implementat, entrenat i avaluat models basats en l'arquitectura T5 [19] per a les tasques de *Text cloze* i de generació del llenguatge. Per a aconseguir-ho, primer s'ha desenvolupat un entorn de treball que facilités els experiments i s'ha adaptat la base de dades existent a les necessitats del projecte i a les tecnologies més modernes fent ús de nous extractors de característiques visuals i de nous *tokenitzadors*. Els experiments han demostrat que la qualitat de l'OCR de la base de dades és un problema recurrent a tots els models. Tot i això, els models de *Text cloze* han superat a la hi-LSTM de COMICS [11]. El model generatiu final és capaç de generar següents diàlegs tot i que contingui errors induïts per la qualitat de la base de dades, i amb una adequació al context limitada.

Amb tot, el model té molt marge de millora i, en un futur treball, seria interessant reconstruir la base de dades original amb un OCR més modern i tornar a repetir els experiments. Altres millores podrien basar-se a escollir els candidats de forma dinàmica en la tasca de *Text cloze* i fer *fine tuning* a un extractor de característiques visuals sobre un conjunt de dades d'imatges de còmic, el qual era subobjectiu que no ha estat possible dur a terme.

Agraïments

Agraïments especials a l'Ernest Valveny Llobet per ser el meu tutor de recerca i guiar-me de la millor manera possible en la consecució d'aquest treball i a en Rubèn Pérez Tito per tot el suport, ajuda i mentoria que m'ha proporcionat al llarg de la meva estada al Centre de Visió per Computador. Sense ells, aquest treball no hauria estat possible. Agrair també a tots els companys del departament de Visió i Llenguatge del CVC i a tots els familiars i amics que m'han donat suport durant aquests mesos.

Referències

[1] Kiyoharu Aizawa et al. "Building a Manga Dataset "Manga109" With Annotations for Multimedia Applications". A: *IEEE MultiMedia* 27.2 (abr. de 2020), pàg. 8 - 18. ISSN: 1941-0166. DOI: 10.1109/mmul.2020.2987895. URL: http://dx.doi.org/10.1109/MMUL.2020.2987895 (v. la pàg. 3).

[2] ArjanCodes. *Data Science Project Refactoring*. https://github.com/ArjanCodes/2021-data-science-refactor. 2021 (v. la pàg. 3).

[3] Satanjeev Banerjee i Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". A: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics, juny de 2005, pàg. 65 - 72. URL: https://aclanthology.org/W05-0909 (v. les pàg. 2, 6).

[4] Hangbo Bao et al. *BEiT: BERT Pre-Training of Image Transformers*. 2021. DOI: 10.48550/ARXIV.2106.08254. URL: https://arxiv.org/abs/2106.08254 (v. la pàg. 7).

[5] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165 [cs.CL] (v. la pàg. 3).

[6] Jaemin Cho et al. *Unifying Vision-and-Language Tasks via Text Generation*. 2021. DOI: 10.48550/ARXIV.2102.02779. URL: https://arxiv.org/abs/2102.02779 (v. les pàg. 2, 3, 6 - 8).

[7] Kyunghyun Cho et al. *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. 2014. arXiv: 1406.1078 [cs.CL] (v. la pàg. 3).

[8] Wei-Ta Chu i Wei-Chung Cheng. "Manga-specific features and latent style model for manga style analysis". A: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2016, pàg. 1332 - 1336. DOI: 10.1109/ICASSP.2016.7471893 (v. la pàg. 2).

[9] William Fedus et al. *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. 2021. DOI: 10.48550/ARXIV.2101.03961. URL: https://arxiv.org/abs/2101.03961 (v. la pàg. 2).

[10] Sepp Hochreiter i Jürgen Schmidhuber. "Long Short-term Memory". A: *Neural computation* 9 (des. de 1997), pàg. 1735-80. DOI: 10.1162/neco.1997.9.8.1735 (v. la pàg. 2).

[11] Mohit Iyyer et al. *The Amazing Mysteries of the Gutter: Drawing Inferences Between Panels in Comic Book Narratives*. 2017. arXiv: 1611.05118 [cs.CV] (v. les pàg. 3, 9).

[12] Ranjay Krishna et al. *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*. 2016. DOI: 10.48550/ARXIV.1602.07332. URL: https://arxiv.org/abs/1602.07332 (v. la pàg. 6).

[13] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2014. DOI: 10.48550/ARXIV.1405.0312. URL: https://arxiv.org/abs/1405.0312 (v. la pàg. 6).

[14] Andrew Mutton et al. "GLEU: Automatic Evaluation of Sentence-Level Fluency". A: *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic: Association for Computational Linguistics, juny de 2007, pàg. 344-351. URL: https://aclanthology.org/P07-1044 (v. les pàg. 2, 6, 7).

[15] Nhu-Van Nguyen et al. "Manga-MMTL: Multimodal Multitask Transfer Learning for Manga Character Analysis". A: *Document Analysis and Recognition – ICDAR 2021*. Ed. de Josep Lladós et al. Cham: Springer International Publishing, 2021, pàg. 410-425. ISBN: 978-3-030-86331-9 (v. la pàg. 2).

[16] Kishore Papineni et al. "Bleu: a Method for Automatic Evaluation of Machine Translation". A: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, jul. de 2002, pàg. 311-318. DOI: 10.3115/1073083.1073135. URL: https://aclanthology.org/P02-1040 (v. la pàg. 6).

[17] Mengshi Qi et al. "Latent Memory-Augmented Graph Transformer for Visual Storytelling". A: *Proceedings of the 29th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2021, pàg. 4892-4901. ISBN: 9781450386517. URL: https://doi.org/10.1145/3474085.3475236 (v. la pàg. 2).

[18] Alec Radford et al. "Language Models are Unsupervised Multitask Learners". A: 2019 (v. la pàg. 3).

[19] Colin Raffel et al. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. 2020. arXiv: 1910.10683 [cs.LG] (v. les pàg. 1-3, 5, 9).

[20] Aniketh Janardhan Reddy i Ramesh Balaji. "Incorporating Visual and Textual Cues in Dialogue Generation : An Application to Comic Strips". A: 2019 (v. les pàg. 3, 8).

[21] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. 2015. DOI: 10.48550/ARXIV.1506.01497. URL: https://arxiv.org/abs/1506.01497 (v. les pàg. 4, 6).

[22] Rico Sennrich et al. *Neural Machine Translation of Rare Words with Subword Units*. 2015. DOI: 10.48550/ARXIV.1508.07909. URL: https://arxiv.org/abs/1508.07909 (v. la pàg. 4).

[23] Karen Simonyan i Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. 2014. DOI: 10.48550/ARXIV.1409.1556. URL: https://arxiv.org/abs/1409.1556 (v. la pàg. 4).

[24] Marko Smilevski et al. "Stories for Images-in-Sequence by Using Visual and Narrative Components". A: *ICT Innovations 2018. Engineering and Life Sciences* (2018), pàg. 148-159. ISSN: 1865-0937. DOI: 10.1007/978-3-030-00825-3_13. URL: http://dx.doi.org/10.1007/978-3-030-00825-3_13 (v. la pàg. 3).

[25] Ting-Hao et al. *Visual Storytelling*. 2016. arXiv: 1604.03968 [cs.CL] (v. les pàg. 2, 6, 8).

[26] Ashish Vaswani et al. *Attention Is All You Need*. 2017. arXiv: 1706.03762 [cs.CL] (v. la pàg. 2).

[27] Yi Xu et al. *Topic-Aware Multi-turn Dialogue Modeling*. 2020. arXiv: 2009.12539 [cs.CL] (v. la pàg. 3).

[28] Kim Young-Min. "Feature visualization in comic artist classification using deep neural networks". A: *Journal of Big Data* 6 (juny de 2019). DOI: 10.1186/s40537-019-0222-3 (v. la pàg. 2).

[29] Zhuosheng Zhang et al. *Modeling Multi-turn Conversation with Deep Utterance Aggregation*. 2018. arXiv: 1806.09102 [cs.CL] (v. la pàg. 3).

Apèndix

A Hiperparàmetres dels models principals

Hiperparàmetre	**T5 jeràrquic**	**T5 base**	**VL-T5**
Mida del batch	128	64	16
Optimitzador	adafactor	adafactor	adamw
Weight Decay	-	-	0.01
Adam epsilon	-	-	1e-06
Learning Rate	1e-6	1e-6	1e-05
Tokenitzador	all-MiniLM-L6-v2	t5-base	vlt5
Backbone	t5-small	t5-base	t5-base
Mida del vocabulari	32100	32100	32200
Mida del *embedding*	512	768	768
Mida del *embedding* de la resposta	384	-	-
Mida del *encoder*	512	768	768
Mida del *pooler*	256	256	-
dropout	0.4	0.4	0.1
dropout_p	0.2	0.2	-
Nombre de candidats	3	3	3
Tokens màxims a la resposta	30	30	30
Tokens màxims per diàleg	30	30	30
Nombre d'objectes (característiques visuals)	-	-	36
Mida de les característiques visuals	-	-	2048

Taula A.1: **Hiperparàmetres utilitzats en l'entrenament dels models T5 jeràrquic, T5 base i VL-T5 (*Text cloze* multimodal i model generatiu).** S'han fet servir els mateixos hiperparàmetres del T5 base per al T5 de mida petita canviant els valors de *tokenizer* i *backbone* per "t5-small".

B Altres experiments

Un dels primers experiments que es va dur a terme al principi del projecte va ser provar diferents maneres de construir la seqüència d'entrada del model. Inicialment, es passava una concatenació de tots els diàlegs tant del context com els candidats. Una idea que s'ha provat ha estat afegir-hi tokens especials al principi de cada diàleg amb la idea d'ajudar al model a saber quan comença cadascun dels diferents diàlegs. Una altra ha estat desordenar el vector dels candidats perquè no vinguessin en un ordre fix sempre per evitar que el model memoritzés la posició del diàleg correcte. Aquests canvis no han millorat gaire els resultats, però s'han mantingut per a la resta d'experiments.

C Exemples qualitatius

A les figures següents es mostren exemples de l'execució del model de *Text cloze* multimodal i del generatiu sobre mostres del conjunt de test de les dues versions de la base de dades (s'ha fet servir el model de *Text cloze* entrenat sobre la corresponent base de dades en cada cas). Totes les execucions s'han dut a terme amb $seed = 4$.

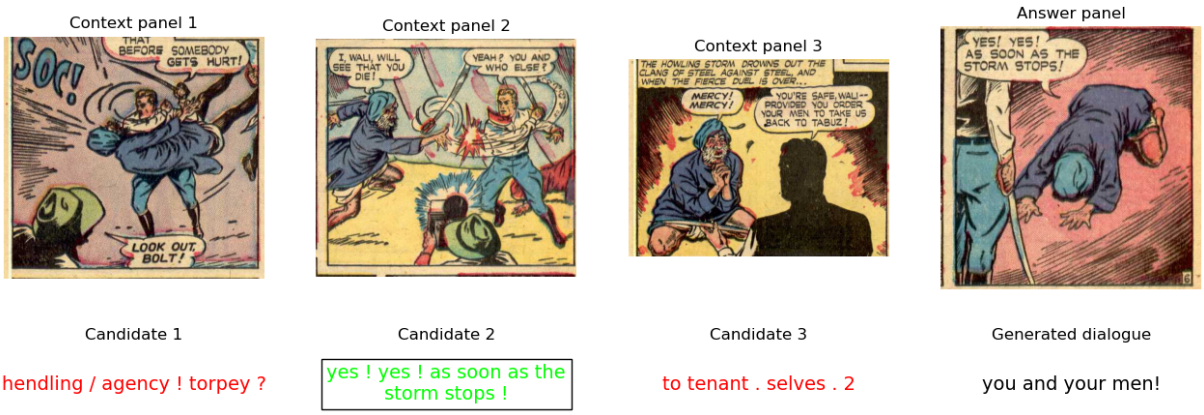

Figura C.1: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *easy*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 7. GLEU: 2,63. METEOR: 4,00.

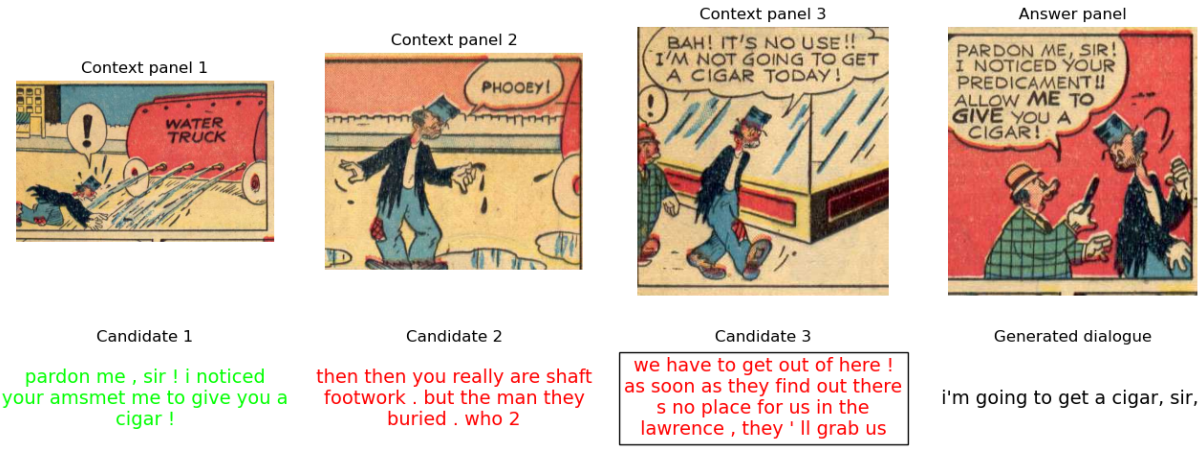

Figura C.2: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *easy*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 8. GLEU: 12,07. METEOR: 27,00.

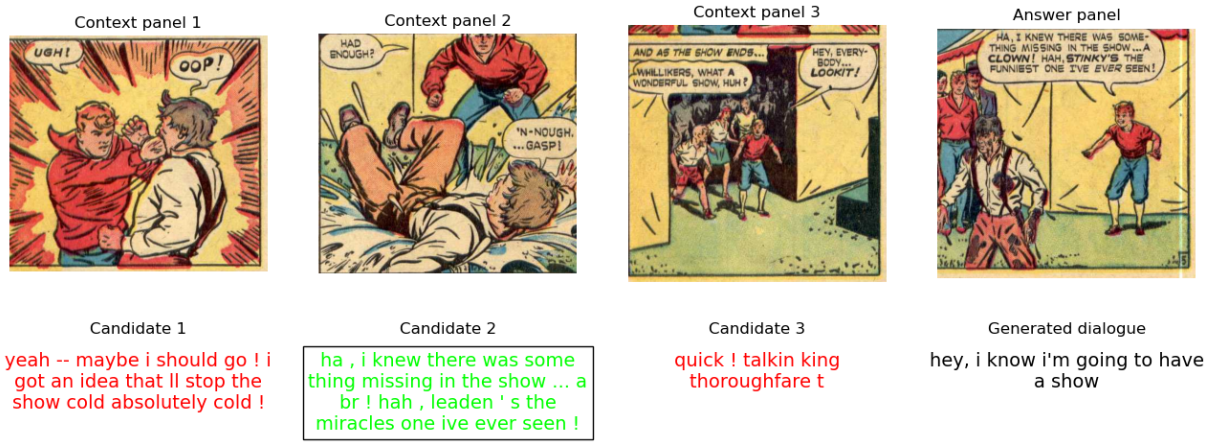

Figura C.3: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *hard*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 6. GLEU: 4,39. METEOR: 9,00.

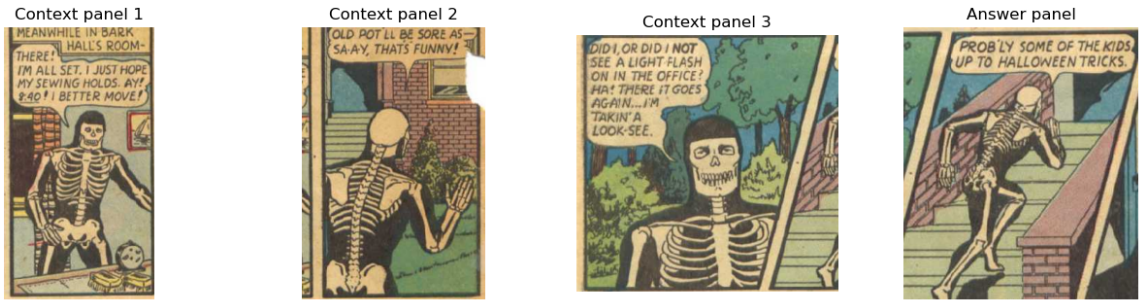

Figura C.4: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *hard*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 18. GLEU: 2,94. METEOR: 0,00.

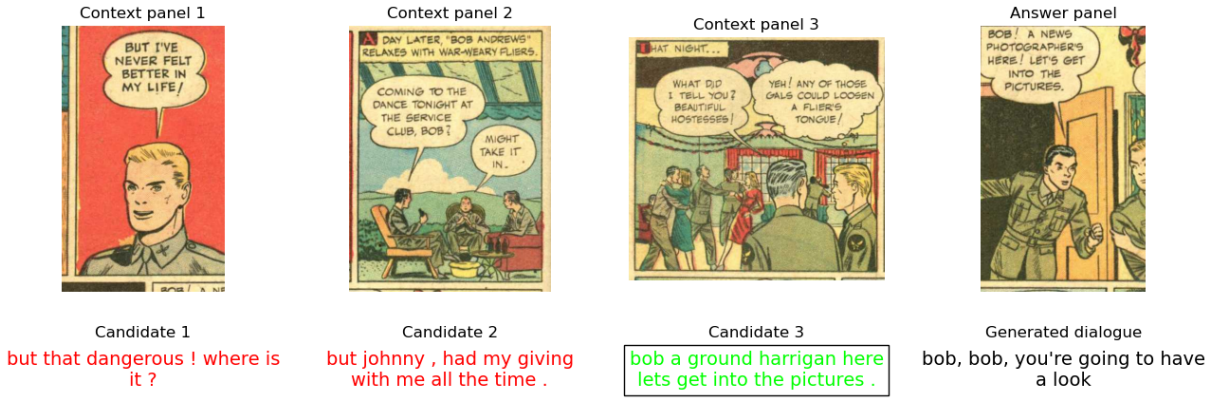

Figura C.5: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *easy*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 68. GLEU: 5,26. METEOR: 13,64.

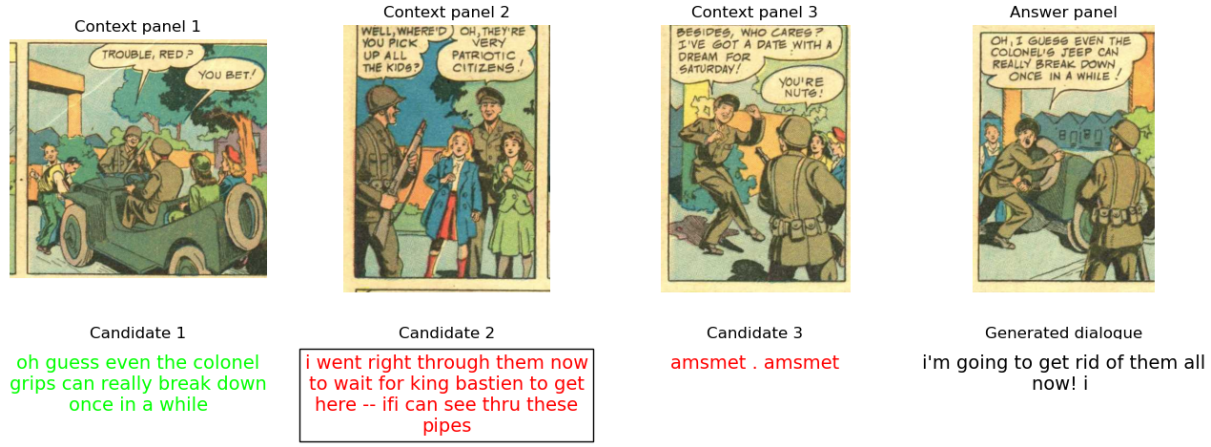

Figura C.6: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *easy*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 72. GLEU: 0,0. METEOR: 3,62.

Figura C.7: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *easy*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 75. GLEU: 5,88. METEOR: 10,00.

Figura C.8: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *easy*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 76. GLEU: 16,67. METEOR: 40,32.

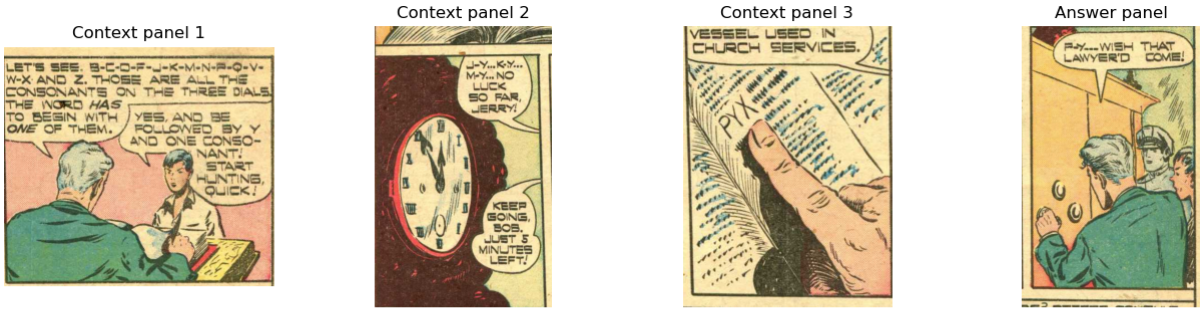

Figura C.9: Exemple d'una execució del model multimodal de *Text cloze* i del generatiu sobre la base de dades *easy*. Els 3 primers textos corresponen als candidats que es presenten al model de *Text cloze*, el color verd indica que és el candidat correcte mentre que el vermell que és incorrecte. El text envoltat per un rectangle és l'escollit pel model. El text de la dreta del tot correspon al diàleg generat pel model generatiu. La mostra executada és la número 99. GLEU: 0,00. METEOR: 0,00.