Supplementary Materials: Towards Accurate Text-based Image Captioning with Content Diversity Exploration

Guanghui Xu[1,2]*, Shuaicheng Niu[1]*, Mingkui Tan[1,4], Yucheng Luo[1], Qing Du[1,4]†, Qi Wu[3]
[1]South China University of Technology, [2]Pazhou Laboratory, [3] University of Adelaide
[4]Key Laboratory of Big Data and Intelligent Robot, Ministry of Education
sexuguanghui@mail.scut.edu.cn, {mingkuitan, duqing}@scut.edu.cn, qi.wu01@adelaide.edu.au

In the supplementary, we provide more implementation details and experimental results of the proposed Anchor-Captioner. We organise the supplementary as follows.

- In Section A, we provide the detailed training and inference algorithms of Anchor-Captioner.
- In Section B, we give more discussions on the anchor-centred graph (ACG) construction strategy.
- In Section C, we conduct more ablation experiments to verify the generalisation ability of Anchor-Captioner.
- In Section D, we conduct more ablation experiments to further measure the importance of each loss term.
- In Section E, we show more visualisation results to further verify the promise of the proposed method.

A. Algorithms

Algorithm 1 Training Method of Anchor-Captioner

Require: Multimodal feature set $\{\widehat{\mathbf{V}}, \widehat{\mathbf{T}}\}$, transformer encoder Ψ, anchor predictor ϕ, sequence projection RNN, visual captioner AnCM_v, text captioner AnCM_t, overall model parameters θ.

Initialize the model parameters θ.
while not converge **do**
  Randomly sample a feature pair $(\widehat{\mathbf{V}}, \widehat{\mathbf{T}})$.
  $\mathbf{V}, \mathbf{T} = \Psi(\widehat{\mathbf{V}}, \widehat{\mathbf{T}}; \theta_a)$. *// Multimodal Embedding Fusion*
  // Anchor Proposal Module
  $\mathbf{s}_{anchor} = \text{Softmax}\,(\phi(\mathbf{T}))$ *// Predict anchor scores.*
  Choose the token with the highest score as $\mathbf{T}_{anchor}$.
  Construct anchor-centred graph $\mathcal{G}$ for $\mathbf{T}_{anchor}$ using RNN.
  // Anchor Captioning Module
  $\mathbf{h}_c = \text{AnCM}_v(\mathbf{V}, \mathbf{y}'_{c-1}; \theta_v)$ *// Obtain global visual info.*
  $\widehat{\mathcal{G}}, \widehat{\mathbf{y}}_c = \text{AnCM}_t(\mathcal{G}, \mathbf{h}_c, \text{LM}(\mathbf{y}_{c-1}); \theta_t)$ *// refine.*
  Obtain the rough caption $\mathcal{Y}' = \{y'_c\}$ using Eqn. (8).
  Obtain the fine-grained caption $\mathcal{Y} = \{y_c\}$ using Eqn. (10).
  Update θ using overall training loss (Eqn. 11)
end while

Algorithm 2 Inference of Anchor-Captioner

Require: multimodal features $(\widehat{\mathbf{V}}, \widehat{\mathbf{T}})$, transformer encoder Ψ, anchor predictor ϕ, sequence projection RNN, visual captioner AnCM_v, text captioner AnCM_t; the number of sampled ACGs (K), pretrained model parameters θ.

// Multimodal Embedding Fusion
$\mathbf{V}, \mathbf{T} = \Psi(\widehat{\mathbf{V}}, \widehat{\mathbf{T}}; \theta_a)$.
// Anchor Proposal Module
$\mathbf{s}_{anchor} = \text{Softmax}\,(\phi(\mathbf{T}))$ *// Predict anchor scores.*
Choose top-K tokens as the anchors $\{\mathbf{T}^k_{anchor}\}^K_{k=1}$.
Construct ACGs $\{\mathcal{G}_k\}$ for the anchors using RNN.
// Anchor Captioning Module
$\mathbf{h}_c = \text{AnCM}_v(\mathbf{V}, \mathbf{y}'_{c-1}; \theta_v)$ *// Obtain global visual info.*
for $k = 1, \ldots, K$ **do**
  $\widehat{\mathcal{G}}_k, \widehat{\mathbf{y}}^k_c = \text{AnCM}_t(\mathcal{G}_k, \mathbf{h}_c, \text{LM}(\mathbf{y}^k_{c-1}); \theta_t)$
  Obtain the rough caption $\mathcal{Y}'_k = \{y'_c\}_k$ using Eqn. (8).
  Obtain the k-th caption $\mathcal{Y}_k = \{y_c\}_k$ using Eqn. (10).
end for
Obtain K rough captions and K fine-grained captions

In this section, we provide the detailed training and inference algorithms of our method in Algorithms 1 and 2. Given an input image, we first fuse visual and text features to obtain multimodal embeddings. Then, we apply the anchor proposal module (AnPM) to choose and group texts to construct a series of anchor-centred graphs (ACGs), where each ACG denotes a group of relevant OCR tokens that are used to generate a specific caption. Last, we employ the visual- captioner (AnCM_v) to generate a rough caption and then use ACGs as guidance to refine the generated caption by the text-captioner (AnCM_t). In particular, we adopt the top-1 ACG for the training while using top-K ACGs to generate K diverse captions in the inference.

*Authors contributed equally.
†Corresponding author

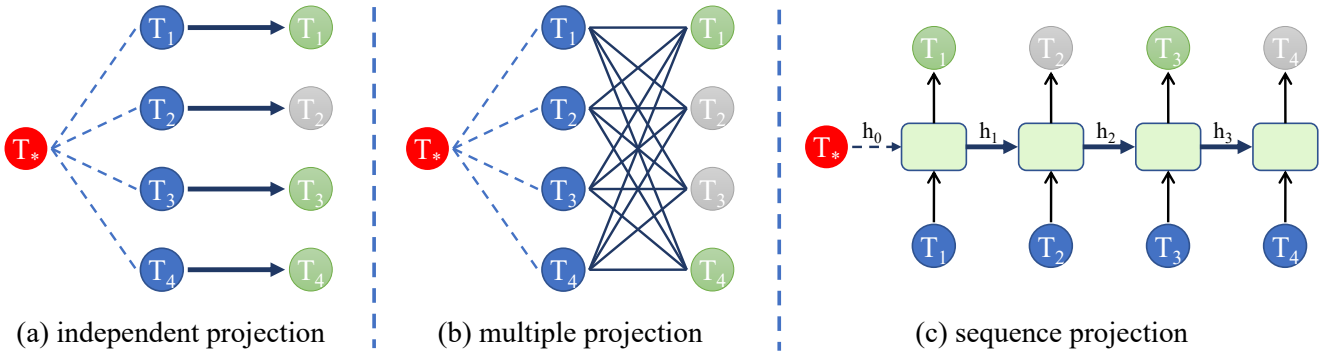


Figure S1. Illustrations of different ACG construction strategies. The red circle denotes the anchor, and the blue circles denote the candidate OCR tokens to be selected. After performing prediction, the green one indicates that the current token and the anchor belong to the same group of ACG, while the gray is the opposite. It means that the ACGs in (a-c) are $\{\mathrm{T}_*, \mathrm{T}_1, \mathrm{T}_3, \mathrm{T}_4\}$, $\{\mathrm{T}_*, \mathrm{T}_1, \mathrm{T}_4\}$ and $\{\mathrm{T}_*, \mathrm{T}_1, \mathrm{T}_3\}$, respectively.

B. More details about ACG

As shown in Figure S1, to construct an anchor-centred graph (ACG), we mainly consider three kinds of construction strategies, i.e., independent projection (fully connected layer), multiple projection (4-layer Transformer module) and sequence projection (RNN module). Specifically, based on a given anchor $\mathbf{T}_{anchor}$, the independent projection directly predicts a correlation score for each token without considering others, while the multiple projection considers global information via self-attention mechanism. As shown in Figure S1, the multiple projection makes prediction for a token (e.g.,$\mathbf{T}_1$) using rich neighbourhood information. In particular, the sequence projection considers the relationships among $\mathbf{T}_{anchor}$, all OCR tokens and history predictions. The order of OCR tokens is determined by the ranking of confidence scores (in descending order) obtained from the OCR model. In a word, the three strategies use different information to construct ACGs, where the independent and multiple projection mainly consider local and global information, and the sequence projection perceives previous predictions through hidden state. To some extent, the sequence projection will be more reasonable, because the choice between different tokens to construct an ACG is not completely independent.

To train AnPM, we first parse ground-truth (gt) captions into candidate sets, where gt-anchor is the most frequently described token and gt-ACG are the tokens appeared in the same caption. Here, we do not have semantic labels of the visual objects in an image, and thus we have trouble to know what objects are included in the ground-truth captions. In this work, we only consider the OCR token and its relative tokens to construct ACGs. We will consider training AnPM to propose sub-regions as RPN in the future study.

C. More ablation studies about generalisation

To further demonstrate the generalisation ability of our method, we conduct ablation experiments on COCO captioning and TextCaps dataset. COCO captioning is a famous large-scale dataset for general image captioning and TextCaps dataset is recently proposed to enhance the captioning ability of existing methods, especially the reading ability. As discussed in the main paper, general image captioning methods mainly focus on visual objects and overall scenes in images, while ignoring text information that is of critical importance for comprehensively understanding images. In this sense, it is necessary to study generalisation ability of existing methods on COCO captioning and TextCaps dataset. To this end, we exploit different settings to conduct experiments. From the results in Table S1, we draw the following main observations: 1) When only training models on COCO captioning (rows 1-2 and 7-8), our model achieves comparable performance as M4C-Captioner. 2) When training models on TextCaps dataset (rows 3-4 and 9-10), our model outperforms M4C-Captioner in terms of two evaluation settings. 3) When jointly training models using both COCO and TextCaps dataset (rows 5-6 and 11-12), our model improves the CIDEr score from 87.5% to 96.3% on COCO and achieves 5% absolute improvement on TextCaps. 4) Unfortunately, as shown in rows 5-6 and 11-12, training on 'COCO+TextCaps' leads to worse performance than only using COCO/TextCaps (rows 1,4,7,10). It means that simply improving the sampling ratio of these two datasets can not handle the domain shift problem, which is already a quite challenging task. However, combining COCO and TextCaps datasets for training is more suitable for complex real scenarios. In this way, the well-trained model is able to 'watch' visual objects and 'read' texts in images.

#	Method	trained on	evaluated on	BLEU	METEOR	ROUGE_L	SPICE	CIDEr
1	M4C	COCO	COCO	34.3	27.5	56.2	20.6	112.2
2			TextCaps	12.3	14.2	34.8	9.2	30.3
3		TextCaps	COCO	8.3	15.1	34.2	8.0	17.3
4			TextCaps	23.3	22.0	46.2	15.6	89.6
5		COCO+TextCaps	COCO	27.1	24.1	51.6	17.4	87.5
6			TextCaps	21.9	22.0	45.0	15.6	84.6
7	Ours	COCO	COCO	34.6	27.3	56.1	20.2	110.3
8			TextCaps	12.6	13.8	35.2	8.8	29.2
9		TextCaps	COCO	8.9	15.5	34.7	8.3	18.4
10			TextCaps	24.7	22.5	47.1	15.9	95.5
11		COCO+TextCaps	COCO	30.5	25.2	53.6	18.4	96.3
12			TextCaps	23.6	22.2	46.2	15.7	90.0

Table S1. More experiments about generalisation. We train our model and M4C-Captioner on TextCaps and COCO captioning training split and then evaluate the models on the different validation split. Specifically, 'COCO+TextCaps' denotes that a model uses both COCO captioning and TextCaps dataset for joint training. In practice, since the scale of COCO is much larger than TextCaps, we set the sampling rate to 1:8 to sample TextCaps as frequently as COCO.

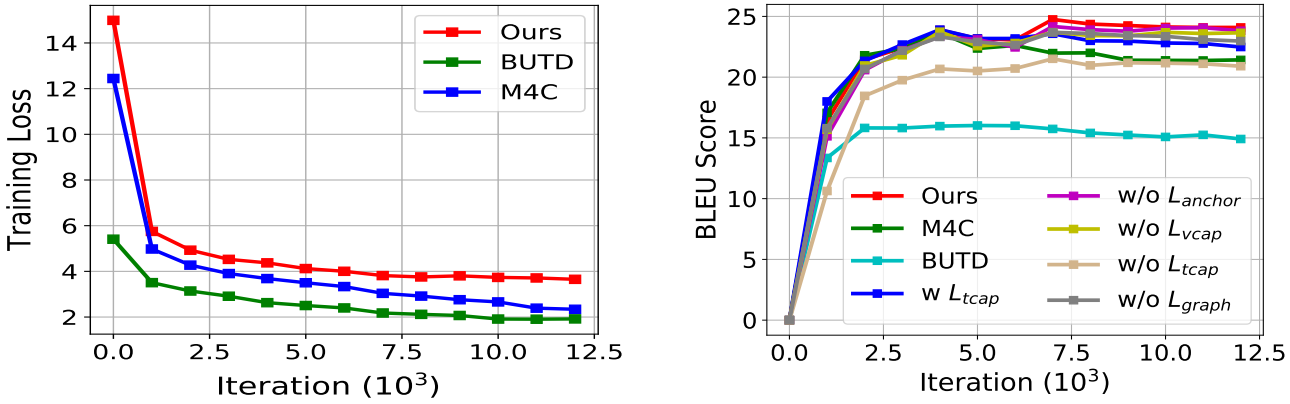


Figure S2. **Left**: The training loss of different methods on the TextCaps training set. **Right**: The BLEU scores under different iterations on the TextCaps validation set. The 'M4C' denotes M4C-Captioner model. The 'w/o' denotes 'Ours' without a specific loss term. For instance, 'w/o $\mathcal{L}_{tcap}$' denotes our method without $\mathcal{L}_{tcap}$ while 'w $\mathcal{L}_{tcap}$' denotes our method only using $\mathcal{L}_{tcap}$.

D. More ablation studies about losses

As shown in Figure S2, we also compare our method with baselines in terms of the training loss and the BLEU score. Since our total training loss contains four terms $\mathcal{L}_{\{anchor,graph,vcap,tcp\}}$, the value of our training loss is little higher than the compared methods. Our method tends to converge after 7k iteration and achieves the highest BLEU score on the validation set. Compared with the considered methods, our method has better generalisation ability to overcome the overfitting problem. To further measure the importance of losses in our method, we conduct several ablation studies, such as removing each loss term. In particular, we can train our model in an end-to-end miner (only using $\mathcal{L}_{tcap}$). From the results, the importance of the losses can be formulated as: $\mathcal{L}_{tcap}>\mathcal{L}_{vcap}>\mathcal{L}_{graph}>\mathcal{L}_{anchor}$.

E. More visualisation analysis

In this section, to further measure the qualities of our method's generation, we provide more visualisation results on TextCaps validation set. Specifically, we fist show more successful results of our method in Figure S3. Then, demonstrate the controllability of our method in Figure S4. Last but not least, in Figure S5, we provide some typical failure cases to evaluate our method more objectively.

As shown in Figure S3, compared with M4C-Captioner, our method is able to describe images from different views and cover more OCR tokens, represented as 'Ours-*'. In particular, our proposed AnCM is a progressive captioning module that AnCM$_v$ first adopts visual information to generate a global caption and then AnCM$_t$ refines the caption based on the text

information of ACGs. Note that, the refining process not only to simply replace <unk> token but also to revise the entire caption in terms of language syntax and contents. In addition, extensive experiments in the main paper also demonstrate the effectiveness of our method.

Based on anchor-centred graphs (ACGs), our method is able to generate multiple controllable image captions. To demonstrate the controllability of our method, we provide more visualisation about the generated captions aligned with the ACGs . As shown in Figure S4, the generated caption of AnCM$_t$ is aligned with the ACGs. We also can see that the generated captions always contain anchors. One possible reason is that our model takes the most important OCR token as an anchor while the other tokens in ACG are used to aggregate information to the anchor. And thus the generated caption is supposed to be at least anchor-related.

As shown in Figure S5, we also provide typical failure cases to further analyse the performance of our method. 1) Although some images could be correctly describe via one global caption, our model still tends to output multiple diverse captions, which might be correct but uninformative, such as '*ruler ... has number 10/20*' in (a). 2) Due to dataset bias, the model tends to generate words with high frequency in training set, such as *brand name is iphone or lg* in (b). 3) More critical, our model is sensitive to anchor-centred graphs (ACGs). As shown in (c)-(d), if the OCR recognise system fails to detect or only detects few OCR tokens in an image, our model will be degraded to existing models that only generate a global caption since we have trouble contracting different ACGs.

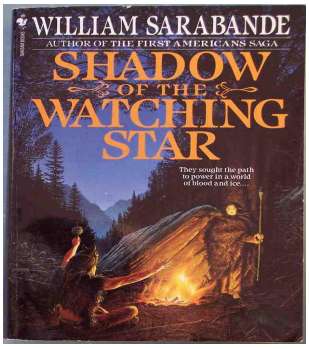
M4C: a man wearing a jersey that says igua nightrun on it
AnCM$_v$: a man in a white shirt green shirt with the number <unk> on it
Ours-1: a man in a white and green shirt with the number 101 on it
Ours-2: a man in a white and green shirt with the word igua on it
Ours-3: a man in a white and green shirt with the number 8 on it

M4C: a bottle of du du du du du du du
AnCM$_v$: a bottle of <unk> beer <unk> sits next to a glass
Ours-1: a bottle of bryggeriet du ferme sits next to a glass of beer
Ours-2: a bottle of du ferme sits next to a glass of beer
Ours-3: a bottle of beer with the word " mat " on the label

M4C: a bottle of racardi mojito mojito is on a table
AnCM$_v$: a bottle of <unk> is <unk> sits on a table day
Ours-1: a bottle of bacardi superior mojito is on a sunny day
Ours-2: a bottle of bacardi is sitting on a pole
Ours-3: a bottle of alcohol with a label that says ' superior ' on it

M4C: a phone that has the word whisper on it
AnCM$_v$: a box for a phone called the <unk>
Ours-1: a box for a phone called whisper drive
Ours-2: a box for a nokia phone has a phone number on it
Ours-3: a box for a phone called the phone

M4C: a display of pink and pink pink pink pink pink pink pink sign with the word macarons on it
AnCM$_v$: a pink sign that <unk> pink pink called <unk> & <unk>
Ours-1: a pink sign for a pink product called caramel de surveillance
Ours-2: a pink sign with a pink background that says ' myrtlle - griotte ' on it
Ours-3: a pink sign with a pink background that says ' sous ' on it

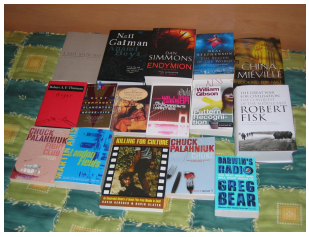


M4C: a book cover for the shadow watching watching by william sarabande
AnCM$_v$: a book by for the book by the book by
Ours-1: a book cover for the shadow of the watching star
Ours-2: a book by william sarabande titled shadow of the first saga
Ours-3: a book titled " the best of the best of america "

M4C: a yellow and yellow advertisement for a phone that says " the world ' s "
AnCM$_v$: a yellow that the words " <unk> ' s <unk> <unk> on it
Ours-1: a yellow poster with the words the world ' s favourite lock on it
Ours-2: a yellow and yellow sign that says ydr2108 / ydr3110 on it
Ours-3: a yellow and yellow poster for a yale should

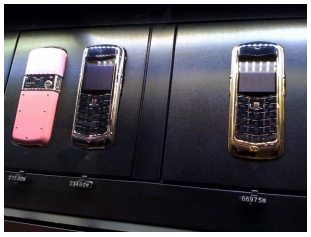
M4C: a poster for a watch that says ' better farming ' on it
AnCM$_v$: a old advertisement and white advertisement for a watch called <unk> <unk>
Ours-1: an old black and white advertisement for a watch for better farming
Ours-2: an old advertisement for a watch for better farming
Ours-3: a black and white page with the word " built " on it

M4C: a stack of books with one titled ' exempt '
AnCM$_v$: a book titled <unk> <unk> sits on a table
Ours-1: a book by mark hillery is on a table
Ours-2: a box that says ' ceo ' on it
Ours-3: a book titled " the author of the author of the book "



M4C: a table with books including one by chuck gaiman
AnCM$_v$: a collection of books including one titled them titled ' the
Ours-1: a collection of books with one of them titled ' darwin '
Ours-2: a collection of books with one of them titled ' neil gaiman '
Ours-3: a collection of books with one titled ' club ' by the author of the books

M4C: a display of magazines including one for the trammyshack
AnCM$_v$: a display of posters including one that says ' <unk> ' on it
Ours-1: a wall of posters with one that says ' big.top ' on it
Ours-2: a display of different posters including one that says ' trammyshack ' on it
Ours-3: a display of different types of different types of which are labeled as ' connectedby '



M4C: a phone with the number 6697500 on it
AnCM$_v$: a phone that has the number <unk> on it
Ours-1: a phone that has the number 6697500 on it
Ours-2: a phone that has the word blackberry on it
Ours-3: three phones are on display including one that says " 23460g "

Figure S3. More visualisation results on the TextCaps validation set. For better visualisation, the underlined word is copy from OCR tokens. The modified tokens are viewed in red colour.

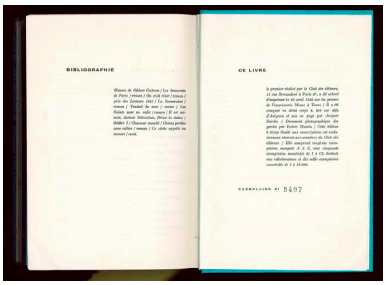
ACG: coarse sea salt from
ACMv: a bottle of <unk> <unk> sits next to a plate of food
ACMt: a bottle of sea salt is next to a plate of food
BLEU: 73.49

ACG: stadium 35 adidas
ACMv: a man wearing a jersey with the number 10 on it
ACMt: a man wearing a jersey with the number 35 on it
BLEU: 57.07

ACG: china airlines cargo
ACMv: a white airlines airplane plane is on the runway
ACMt: a china airlines cargo plane is on the runway
BLEU: 43.17

ACG: india pale ale
ACMv: a bottle of <unk> <unk> ale is next to a glass of beer
ACMt: a bottle of india pale ale is next to a glass of beer
BLEU: 71.16

ACG: tiger lager beer men
ACMv: a bottle of <unk> beer sits next to a glass of beer
ACMt: a bottle of tiger beer is next to a glass of beer
BLEU: 53.11



ACG: bibliographie ce exemplaire
ACMv: a book is open to a page that says '"
ACMt: a book is open to a page that says bibliographie
BLEU: 50.88

Figure S4. Visualisation results on controllability of our method. For each image, we show the top-1 anchor-centred graph (ACG) and the generated captions of visual-captioner (AnCM_v) and text-captioner (AnCM_t). In particular, we report the BLEU score of text-captioner's output. For better visualisation, the anchor in ACG is viewed in blue colour, the underlined word is copy from ACG and the modified tokens are viewed in red colour.

(a)

M4C: a ruler with the numbers 2002 on it
AnCM$_v$: a ruler that has the number 1 through 9 on it
Ours-1: a ruler that has the numbers 1 through t293 on it
Ours-2: a ruler that has the number 10 on it
Ours-3: a ruler that has the number 20 on it

(b)

M4C: a black iphone with the back of a black background
AnCM$_v$: a black iphone is laying face down on a white background
Ours-1: a black iphone is laying face down on a white background
Ours-2: a black iphone sits face down on a white surface
Ours-3: a black lg phone is laying face down on a white surface

(c)

M4C: a man wearing a white shirt with the number 3 on it
AnCM$_v$: a group in a white shirt with the number <unk> on it
Ours-1: a man in a white shirt with the number 8 on it is talking to another man
Ours-2: a man in a white shirt with the number 8 on it is talking to another man
Ours-3: a man in a white shirt with the number 8 on it is talking to another man

(d)

M4C: a baseball player with the number 14 on his jersey
AnCM$_v$: a baseball player with the number 5 on his jersey
Ours-1: a baseball player with the number 321 on his jersey
Ours-2: a baseball player with the number 321 on his jersey
Ours-3: a baseball player with the number 321 on his jersey

Figure S5. Some failure cases of our model on the TextCaps validation set. The <unk> denotes 'unkown' token. The underlined word is copy from OCR tokens. The modified tokens are viewed in red colour.