Multimodal Named Entity Recognition for Short Social Media Posts

Seungwhan Moon[1,2], **Leonardo Neves**[2], **Vitor Carvalho**[3]
[1] Language Technologies Institute, Carnegie Mellon University
[2] Snap Research
[3] Intuit
seungwhm@cs.cmu.edu, lneves@snap.com, vitor_carvalho@intuit.com

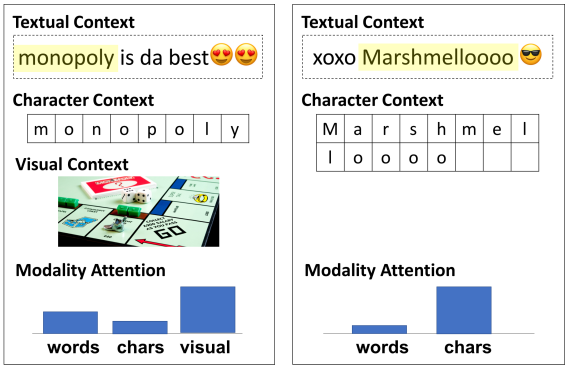
Abstract

We introduce a new task called Multimodal Named Entity Recognition (MNER) for noisy user-generated data such as tweets or Snapchat captions, which comprise short text with accompanying images. These social media posts often come in inconsistent or incomplete syntax and lexical notations with very limited surrounding textual contexts, bringing significant challenges for NER. To this end, we create a new dataset for MNER called SnapCaptions (Snapchat image-caption pairs submitted to public and crowd-sourced stories with fully annotated named entities). We then build upon the state-of-the-art Bi-LSTM word/character based NER models with 1) a deep image network which incorporates relevant visual context to augment textual information, and 2) a generic *modality-attention* module which learns to attenuate irrelevant modalities while amplifying the most informative ones to extract contexts from, adaptive to each sample and token. The proposed MNER model with modality attention significantly outperforms the state-of-the-art text-only NER models by successfully leveraging provided visual contexts, opening up potential applications of MNER on myriads of social media platforms.

1 Introduction

Social media with abundant user-generated posts provide a rich platform for understanding events, opinions and preferences of groups and individuals. These insights are primarily hidden in unstructured forms of social media posts, such as in free-form text or images without tags. Named entity recognition (NER), the task of recognizing named entities from free-form text, is thus a critical step for building structural information, allowing for its use in personalized assistance, recommendations, advertisement, etc.

While many previous approaches (Lample et al., 2016; Ma and Hovy, 2016; Chiu and Nichols, 2015; Huang et al., 2015; Lafferty et al., 2001) on NER have shown success for well-formed text in recognizing named entities via word context resolution (*e.g.* LSTM with word embeddings) combined with character-level features (*e.g.* CharLSTM/CNN), several additional challenges remain for recognizing named entities from extremely short and coarse text found in social media posts. For instance, short social media posts often do not provide enough textual contexts to resolve polysemous entities (*e.g.* "monopoly is da best 😍", where 'monopoly' may refer to a board game (named entity) or a term in economics). In addition, noisy text includes a huge number of unknown tokens due to inconsistent lexical notations and frequent mentions of various newly trending entities (*e.g.* "xoxo Marshmellooo 😎", where 'Marshmellooo' is a mis-spelling of a known entity 'Marshmello', a



(a) (b)

Figure 1: **Multimodal NER + modality attention**. (a) Visual contexts help recognizing polysemous entity names ('Monopoly' as in a board game versus an economics term). (b) Modality attention successfully suppresses word embeddings of ~~a~~ unknown token ('Marshmellooo' with erroneously trailing 'o's), and focuses on character-based context (*e.g.* capitalized first letter, and lexical similarity to a known named entity ('Marshmello', a music producer)) for correct prediction.

Proceedings of NAACL-HLT 2018, pages 852–860
New Orleans, Louisiana, June 1 - 6, 2018.

music producer), making word embeddings based neural networks NER models vulnerable.

To address the challenges above for social media posts, we build upon the state-of-the-art neural architecture for NER with the following two novel approaches (Figure 1). First, we propose to leverage auxiliary modalities for additional context resolution of entities. For example, many popular social media platforms now provide ways to compose a post in multiple modalities - specifically image and text (*e.g. Snapchat* captions, *Twitter* posts with image URLs), from which we can obtain additional context for understanding posts. While "monopoly" in the previous example is ambiguous in its textual form, an accompanying snap image of a board game can help disambiguate among polysemous entities, thereby correctly recognizing it as a named entity.

Second, we also propose a general modality attention module which chooses per decoding step the most informative modality among available ones (in our case, word embeddings, character embeddings, or visual features) to extract context from. For example, the modality attention module lets the decoder attenuate the word-level signals for unknown word tokens (*e.g.* "Marshmellooooo" with trailing 'o's) and amplifies character-level features intsead (*e.g.* capitalized first letter, lexical similarity to other known named entity token 'Marshmello', etc.), thereby suppressing noise information ("UNK" token embedding) in decoding steps. Note that most of the previous literature in NER or other NLP tasks combine word and character-level information with naive concatenation, which is vulnerable to noisy social media posts. When an auxiliary image is available, the modality attention module determines to amplify this visual context *e.g.* in disambiguating polysemous entities, or to attenuate visual contexts when they are irrelevant to target named entities, *e.g.* selfies, etc. Note that the proposed modality attention module is distinct from how attention is used in other sequence-to-sequence literature (*e.g.* attending to a specific token within an input sequence). Section 2 provides the detailed literature review.

Our contributions are three-fold: we propose (1) an LSTM-CNN hybrid multimodal NER network that takes as input both image and text for recognition of a named entity in text input. To the best of our knowledge, our approach is the first work to incorporate visual contexts for named entity recognition tasks. (2) We propose a general *modality attention* module that selectively chooses modalities to extract primary context from, maximizing information gain and suppressing irrelevant contexts from each modality (we treat words, characters, and images as separate modalities). (3) We show that the proposed approaches outperform the state-of-the-art NER models (**both** with and without using additional visual contexts) on our new MNER dataset *SnapCaptions*, a large collection of informal and extremely short social media posts paired with unique images.

2 Related Work

Neural models for NER have been recently proposed, producing state-of-the-art performance on standard NER tasks. For example, some of the end-to-end NER systems (Passos et al., 2014; Chiu and Nichols, 2015; Huang et al., 2015; Lample et al., 2016; Ma and Hovy, 2016) use a recurrent neural network usually with a CRF (Lafferty et al., 2001; McCallum and Li, 2003) for sequence labeling, accompanied with feature extractors for words and characters (CNN, LSTMs, etc.), and achieve the state-of-the-art performance mostly without any use of gazetteers information. Note that most of these work aggregate textual contexts via concatenation of word embeddings and character embeddings. Recently, several work have addressed the NER task specifically on noisy short text segments such as Tweets, etc. (Baldwin et al., 2015; Aguilar et al., 2017). They report performance gains from leveraging external sources of information such as lexical information (*e.g.* POS tags, etc.) and/or from several preprocessing steps (*e.g.* token substitution, etc.). Our model builds upon these state-of-the-art neural models for NER tasks, and improves the model in two critical ways: (1) incorporation of visual contexts to provide auxiliary information for short media posts, and (2) addition of the modality attention module, which better incorporates word embeddings and character embeddings, especially when there are many missing tokens in the given word embedding matrix. Note that we do not explore the use of gazetteers information or other auxiliary information (POS tags, etc.) (Ratinov and Roth, 2009) as it is not the focus of our study.

Attention modules are widely applied in several deep learning tasks (Xu et al., 2015; Chan

et al., 2015; Sukhbaatar et al., 2015; Yao et al., 2015). For example, they use an attention module to attend to a subset within a single input (a part/region of an image, a specific token in an input sequence of tokens, etc.) at each decoding step in an encoder-decoder framework for image captioning tasks, etc. (Rei et al., 2016) explore various attention mechanisms in NLP tasks, but do not incorporate visual components or investigate the impact of such models on noisy social media data. (Moon and Carbonell, 2017) propose to use attention for a subset of discrete source samples in transfer learning settings. Our modality attention differs from the previous approaches in that we attenuate or amplifies each modality input as a whole among multiple available modalities, and that we use the attention mechanism essentially to map heterogeneous modalities in a single joint embedding space. Our approach also allows for reuse of the same model for predicting labels even when some of the modalities are missing in input, as other modalities would still preserve the same semantics in the embeddings space.

Multimodal learning is studied in various domains and applications, aimed at building a joint model that extracts contextual information from multiple modalities (views) of parallel datasets.

The most relevant task to our multimodal NER system is the task of multimodal machine translation (Elliott et al., 2015; Specia et al., 2016), which aims at building a better machine translation system by taking as input a sentence in a source language as well as a corresponding image. Several standard sequence-to-sequence architectures are explored (*e.g.* a target-language LSTM decoder that takes as input an image first).

Other previous literature include study of Canonical Correlation Analysis (CCA) (Dhillon et al., 2011) to learn feature correlations among multiple modalities, which is widely used in many applications. Other applications include image captioning (Xu et al., 2015), audio-visual recognition (Moon et al., 2015), visual question answering systems (Antol et al., 2015), etc.

To the best of our knowledge, our approach is the first work to incorporate visual contexts for named entity recognition tasks.

3 Proposed Methods

Figure 2 illustrates the proposed multimodal NER (MNER) model. First, we obtain word embeddings, character embeddings, and visual features (Section 3.1). A Bi-LSTM-CRF model then takes as input a sequence of tokens, each of which comprises a word token, a character sequence, and an image, in their respective representation (Section 3.2). At each decoding step, representations from each modality are combined via the modality attention module to produce an entity label for each token (3.3). We formulate each component of the model in the following subsections.

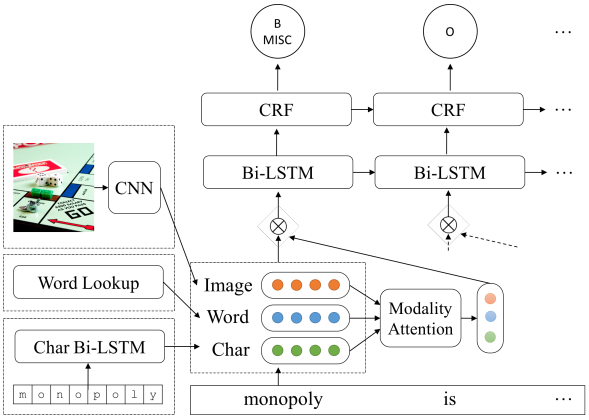


Figure 2: The main architecture for our multimodal NER (MNER) network with modality attention. At each decoding step, word embeddings, character embeddings, and visual features are merged with modality attention. Bi-LSTM/CRF takes as input each token and produces an entity label.

Notations: Let $\mathbf{x} = \{\mathbf{x}_t\}_{t=1}^{T}$ a sequence of input tokens with length T, with a corresponding label sequence $\mathbf{y} = \{\mathbf{y}_t\}_{t=1}^{T}$ indicating named entities (*e.g.* in standard BIO formats). Each input token is composed of three modalities: $\mathbf{x}_t = \{\mathbf{x}_t^{(w)}, \mathbf{x}_t^{(c)}, \mathbf{x}_t^{(v)}\}$ for word embeddings, character embeddings, and visual embeddings representations, respectively.

3.1 Features

Similar to the state-of-the-art NER approaches (Lample et al., 2016; Ma and Hovy, 2016; Aguilar et al., 2017; Passos et al., 2014; Chiu and Nichols, 2015; Huang et al., 2015), we use both word embeddings and character embeddings.

Word embeddings are obtained from an unsupervised learning model that learns co-occurrence statistics of words from a large external corpus, yielding word embeddings as distributional semantics (Mikolov et al., 2013). Specifically, we use pre-trained embeddings from GloVE (Pennington et al., 2014).

Character embeddings are obtained from a Bi-LSTM which takes as input a sequence of characters of each token, similarly to (Lample et al., 2016). An alternative approach for obtaining character embeddings is using a convolutional neural network as in (Ma and Hovy, 2016), but we find that Bi-LSTM representation of characters yields empirically better results in our experiments.

Visual embeddings: To extract features from an image, we take the final hidden layer representation of a modified version of the convolutional network model called Inception (GoogLeNet) (Szegedy et al., 2014, 2015) trained on the ImageNet dataset (Russakovsky et al., 2015) to classify multiple objects in the scene. Our implementation of the Inception model has deep 22 layers, training of which is made possible via "network in network" principles and several dimension reduction techniques to improve computing resource utilization. The final layer representation encodes discriminative information describing what objects are shown in an image, which provide auxiliary contexts for understanding textual tokens and entities in accompanying captions.

Incorporating this visual information onto the traditional NER system is an open challenge, and multiple approaches can be considered. For instance, one may provide visual contexts only as an initial input to decoder as in some encoder-decoder image captioning systems (Vinyals et al., 2015). However, we empirically observe that an NER decoder which takes as input the visual embeddings at every decoding step (Section 3.2), combined with the modality attention module (Section 3.3), yields better results.

Lastly, we add a transform layer for each feature *e.g.* $\mathbf{x}_t^{(w)}, \mathbf{x}_t^{(c)}, \mathbf{x}_t^{(v)} := \sigma_w(\mathbf{x}_t^{(w)}), \sigma_c(\mathbf{x}_t^{(c)}), \sigma_v(\mathbf{x}_t^{(v)})$ before it is fed to the NER entity LSTM.

3.2 Bi-LSTM + CRF for Multimodal NER

Our MNER model is built on a Bi-LSTM and CRF hybrid model. We use the following implementation for the entity Bi-LSTM.

$$
\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_{xi}\mathbf{h}_{t-1} + \mathbf{W}_{ci}\mathbf{c}_{t-1}) \\
\mathbf{c}_t &= (1 - \mathbf{i}_t) \odot \mathbf{c}_{t-1} \\
&\quad + \mathbf{i}_t \odot \tanh(\mathbf{W}_{xc}\overline{\mathbf{x}_t} + \mathbf{W}_{hc}\mathbf{h}_{t-1}) \\
\mathbf{o}_t &= \sigma(\mathbf{W}_{xo}\overline{\mathbf{x}_t} + \mathbf{W}_{ho}\mathbf{h}_{t-1} + \mathbf{W}_{co}\mathbf{c}_t) \\
\mathbf{h}_t &= \text{LSTM}(\overline{\mathbf{x}_t}) \\
&= \mathbf{o}_t \odot \tanh(\mathbf{c}_t)
\end{aligned} \tag{1}
$$

where $\overline{\mathbf{x}_t}$ is a weighted average of three modalities $\mathbf{x}_t = \{\mathbf{x}_t^{(w)}; \mathbf{x}_t^{(c)}; \mathbf{x}_t^{(v)}\}$ via the modality attention module, which will be defined in Section 3.3. Bias terms for gates are omitted here for simplicity of notation.

We then obtain bi-directional entity token representations $\overleftrightarrow{\mathbf{h}_t} = [\overrightarrow{\mathbf{h}_t}; \overleftarrow{\mathbf{h}_t}]$ by concatenating its left and right context representations. To enforce structural correlations between labels in sequence decoding, $\overleftrightarrow{\mathbf{h}_t}$ is then passed to a conditional random field (CRF) to produce a label for each token maximizing the following objective.

$$
\mathbf{y}^* = \underset{\mathbf{y}}{\text{argmax}}\, p(\mathbf{y}|\overleftrightarrow{\mathbf{h}}; \mathbf{W}_{\text{CRF}}) \tag{2}
$$

$$
p(\mathbf{y}|\overleftrightarrow{\mathbf{h}}; \mathbf{W}_{\text{CRF}}) = \frac{\prod_t \psi_t(\mathbf{y}_{t-1}, \mathbf{y}_t; \overleftrightarrow{\mathbf{h}})}{\sum_{\mathbf{y}'} \prod_t \psi_t(\mathbf{y'}_{t-1}, \mathbf{y'}_t; \overleftrightarrow{\mathbf{h}})}
$$

where $\psi_t(\mathbf{y}', \mathbf{y}'; \overleftrightarrow{\mathbf{h}})$ is a potential function, $\mathbf{W}_{\text{CRF}}$ is a set of parameters that defines the potential functions and weight vectors for label pairs $(\mathbf{y}', \mathbf{y}')$. Bias terms are omitted for brevity of formulation.

The model can be trained via log-likelihood maximization for the training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}$:

$$
\mathcal{L}(\mathbf{W}_{\text{CRF}}) = \sum_i \log p(\mathbf{y}|\overleftrightarrow{\mathbf{h}}; \mathbf{W}) \tag{3}
$$

3.3 Modality Attention

The modality attention module learns a unified representation space for multiple available modalities (*e.g.* words, characters, images, etc.), and produces a single vector representation with aggregated knowledge among multiple modalities, based on their weighted importance. We motivate this module from the following observations.

A majority of the previous literature combine the word and character-level contexts by simply concatenating the word and character embeddings at each decoding step, *e.g.* $\mathbf{h}_t = \text{LSTM}([\mathbf{x}_t^{(w)}; \mathbf{x}_t^{(c)}])$ in Eq.1. However, this naive concatenation of two modalities (word and characters) results in inaccurate decoding, specifically for unknown word token embeddings (*e.g.* an all-zero vector $\mathbf{x}_t^{(w)} = \mathbf{0}$ or a random vector $\mathbf{x}_t^{(w)} = \boldsymbol{\epsilon} \sim U(-\sigma, +\sigma)$ is assigned for any unknown token $\mathbf{x}_t$, thus $\mathbf{h}_t = \text{LSTM}([\mathbf{0}; \mathbf{x}_t^{(c)}])$ or $\text{LSTM}([\epsilon; \mathbf{x}_t^{(c)}])$). While this concatenation approach does not cause significant errors for well-formatted text, we observe that it induces performance degradation for our social media post

datasets which contain a significant number of missing tokens.

Similarly, naive merging of textual and visual information (*e.g.* $\mathbf{h}_t = \text{LSTM}([\mathbf{x}_t^{(w)}; \mathbf{x}_t^{(c)}; \mathbf{x}_t^{(v)}])$) yields suboptimal results as each modality is treated equally informative, whereas in our datasets some of the images may contain irrelevant contexts to textual modalities. Hence, ideally there needs a mechanism in which the model can effectively turn the *switch* on and off the modalities adaptive to each sample.

To this end, we propose a general modality attention module, which adaptively attenuates or emphasizes each modality as a whole at each decoding step t, and produces a soft-attended context vector $\overline{\mathbf{x}_t}$ as an input token for the entity LSTM.

$$[\mathbf{a}_t^{(w)}, \mathbf{a}_t^{(c)}, \mathbf{a}_t^{(v)}] = \sigma\big(\mathbf{W}_m \cdot [\mathbf{x}_t^{(w)}; \mathbf{x}_t^{(c)}; \mathbf{x}_t^{(v)}] + \mathbf{b}_m\big)$$

$$\alpha_t^{(m)} = \frac{\exp(\mathbf{a}_t^{(m)})}{\sum_{m' \in \{w,c,v\}} \exp(\mathbf{a}_t^{(m')})} \quad \forall m \in \{w, c, v\}$$

$$\overline{\mathbf{x}_t} = \sum_{m \in \{w,c,v\}} \alpha_t^{(m)} \mathbf{x}_t^{(m)} \tag{4}$$

where $\alpha_t = [\alpha_t^{(w)}; \alpha_t^{(c)}; \alpha_t^{(v)}] \in \mathbb{R}^3$ is an attention vector at each decoding step t, and $\overline{\mathbf{x}_t}$ is a final context vector at t that maximizes information gain for $\mathbf{x}_t$. Note that the optimization of the objective function (Eq.1) with modality attention (Eq.4) requires each modality to have the same dimension (*e.g.* $\mathbf{x}_t^{(w)}, \mathbf{x}_t^{(c)}, \mathbf{x}_t^{(v)} \in \mathbb{R}^p$), and that the transformation via $\mathbf{W}_m$ essentially enforces each modality to be mapped into the same unified subspace, where the weighted average of which encodes discrimitive features for recognition of named entities.

When visual context is not provided with each token (as in the traditional NER task), we can define the modality attention for word and character embeddings only in a similar way:

$$[\mathbf{a}_t^{(w)}, \mathbf{a}_t^{(c)}] = \sigma\big(\mathbf{W}_m \cdot [\mathbf{x}_t^{(w)}; \mathbf{x}_t^{(c)}] + \mathbf{b}_m\big) \tag{5}$$

$$\alpha_t^{(m)} = \frac{\exp(\mathbf{a}_t^{(m)})}{\sum_{m' \in \{w,c\}} \exp(\mathbf{a}_t^{(m')})} \quad \forall m \in \{w, c\}$$

$$\overline{\mathbf{x}_t} = \sum_{m \in \{w,c\}} \alpha_t^{(m)} \mathbf{x}_t^{(m)}$$

Note that while we apply this modality attention module to the Bi-LSTM+CRF architecture (Section 3.2) for its empirical superiority, the module itself is flexible and thus can work with other NER architectures or for other multimodal applications.

4 Empirical Evaluation

4.1 SnapCaptions Dataset

The **SnapCaptions** dataset is composed of 10K user-generated image (snap) and textual caption pairs where named entities in captions are manually labeled by expert human annotators (entity types: PER, LOC, ORG, MISC). These captions are collected exclusively from snaps submitted to public and crowd-sourced stories (aka Snapchat *Live Stories* or *Our Stories*). Examples of such public crowd-sourced stories are "New York Story" or "Thanksgiving Story", which comprise snaps that are aggregated for various public events, venues, etc. All snaps were posted between year 2016 and 2017, and do not contain ~~raw images~~ or other associated information (only textual captions and obfuscated visual descriptor features extracted from the pre-trained InceptionNet are available). We split the dataset into train (70%), validation (15%), and test sets (15%). The captions data have average length of 30.7 characters (5.81 words) with vocabulary size 15,733, where 6,612 are considered unknown tokens from Stanford GloVE embeddings (Pennington et al., 2014). Named entities annotated in the SnapCaptions dataset include many of new and emerging entities, and they are found in various surface forms (various nicknames, typos, etc.) To the best of our knowledge, *SnapCaptions* is the only dataset that contains natural image-caption pairs with expert-annotated named entities.

4.2 Baselines

Task: given a caption and a paired image (if used), the goal is to label every token in a caption in BIO scheme (B: beginning, I: inside, O: outside) (Sang and Veenstra, 1999). We report the performance of the following state-of-the-art NER models as baselines, as well as several configurations of our proposed approach to examine contributions of each component (W: word, C: char, V: visual).

- Bi-LSTM/CRF (W only): only takes word token embeddings (Stanford GloVE) as input. The rest of the architecture is kept the same.
- Bi-LSTM/CRF + Bi-CharLSTM (C only): only takes a character sequence of each word token as input. (No word embeddings)

Modalities	Model	4 Entity Types (%)			Segmentation (%)		
		Prec.	Recall	F1	Prec.	Recall	F1
C	Bi-LSTM/CRF + Bi-CharLSTM	5.0	28.1	8.5	68.6	10.8	18.6
W	Bi-LSTM/CRF	38.2	53.3	44.6	82.5	50.1	62.4
W + C	(Aguilar et al., 2017)	45.9	48.9	47.4	74.0	61.7	67.3
W + C	(Ma and Hovy, 2016)	46.0	51.9	48.7	76.8	61.0	68.0
W + C	(Lample et al., 2016)	47.7	49.9	48.8	74.4	63.3	68.4
~~W + C~~	~~Bi-LSTM/CRF + Bi-CharLSTM w/ Modality Attention~~	49.4	51.7	50.5	75.7	63.3	68.9
W + C + V	Bi-LSTM/CRF + Bi-CharLSTM + Inception	**50.5**	52.3	51.4	71.9	**66.5**	**69.1**
W + C + V	Bi-LSTM/CRF + Bi-CharLSTM + Inception w/ Modality Attention	48.7	**58.7**	**52.4**	**77.4**	60.6	68.0

Table 1: NER performance on the *SnapCaptions* dataset with varying modalities (W: word, C: char, V: visual). We report precision, recall, and F1 score for both entity types recognition (PER, LOC, ORG, MISC) and entity segmentation (untyped recognition - named entity or not) tasks.

- Bi-LSTM/CRF + Bi-CharLSTM (W+C) (Lample et al., 2016): takes as input both word embeddings and character embeddings extracted from a Bi-CharLSTM. Entity LSTM takes concatenated vectors of word and character embeddings as input tokens.
- Bi-LSTM/CRF + CharCNN (W+C) (Ma and Hovy, 2016): uses character embeddings extracted from a CNN instead.
- Bi-LSTM/CRF + CharCNN (W+C) + Multi-task (Aguilar et al., 2017): trains the model to perform both recognition (into multiple entity types) as well as segmentation (binary) tasks.
- (**proposed**) Bi-LSTM/CRF + Bi-CharLSTM with modality attention (W+C): uses the modality attention to merge word and character embeddings.
- (**proposed**) Bi-LSTM/CRF + Bi-CharLSTM + Inception (W+C+V): takes as input visual contexts extracted from InceptionNet as well, concatenated with word and char vectors.
- (**proposed**) Bi-LSTM/CRF + Bi-CharLSTM + Inception with modality attention (W+C+V): uses the modality attention to merge word, character, and visual embeddings as input to entity LSTM.

4.3 Results: SnapCaptions Dataset

Table 1 shows the NER performance on the *Snap Captions* dataset. We report both entity types recognition (PER, LOC, ORG, MISC) and named entity segmentation (named entity or not) results.

Parameters: We tune the parameters of each model with the following search space (bold indicate the choice for our final model): character embeddings dimension: {25, 50, 100, **150**, 200, 300}, word embeddings size: {25, 50, 100, **150**, 200, 300}, LSTM hidden states: {25, 50, **100**, 150, 200, 300}, and $\overline{x}$ dimension: {25, 50, 100, **150**, 200, 300}. We optimize the parameters with Adagrad (Duchi et al., 2011) with batch size 10, learning rate 0.02, epsilon 10^{-8}, and decay 0.0.

Main Results: When visual context is available (W+C+V), we see that the model performance greatly improves over the textual models (W+C), showing that visual contexts are complimentary to textual information in named entity recognition tasks. In addition, it can be seen that the modality attention module further improves the entity type recognition performance for (W+C+V). This result indicates that the modality attention is able to focus on the most effective modality (visual, words, or characters) adaptive to each sample to maximize information gain. Note that our text-only model (W+C) with the modality attention module also significantly outperform the state-of-the-art baselines (Aguilar et al., 2017; Ma and Hovy, 2016; Lample et al., 2016) that use the same textual modalities (W+C), showing the effectiveness of the modality attention module for textual models as well.

Error Analysis: Table 2 shows example cases where incorporation of visual contexts affects prediction of named entities. For example, the token 'curry' in the caption *"The curry's* 🏆*"* is polysemous and may refer to either a type of food or a famous basketball player 'Stephen Curry', and the surrounding textual contexts do not provide

	Caption (target)	Visual Tags	GT	Prediction (W+C+V)	Prediction (W+C)
+	*"The curry's 🏆"*	parade, marching, urban area, ...	B-PER	B-PER	O
	"Grandma w dat lit Apple Crisp"	funnel cake, melting, frozen, ...	O	O	B-ORG
	"Okay duke dumont 😍"	DJ, guitarist, circus, ...	B,I-PER	B,I-PER	O,O
	"CSI with my hubby"	TV, movie, television, ...	B-MISC	B-MISC	B-ORG
	"Twin day at angel stadium"	stadium, arena, stampede, ...	B,I-LOC	B,I-LOC	O,O
	"LETS GO CID"	drum, DJ, drummer, ...	B-PER	B-PER	O
	"MARSHMELLOOOOOOOS"	DJ, night, martini, ...	B-PER	B-PER	O
-	*"Y'all come see me at bojangles. 😊"*	floor, tile, airport terminal, ...	B-ORG	O	B-ORG
	"If u're not watching this season of bachelorette ur doing LIFE WRONG"	monitor, suite, cubicle, ...	B-MISC	O	B-MISC

Table 2: Error analysis: **when do images help NER**? Ground-truth labels (GT) and predictions of our model with vision input (W+C+V) and the one without (W+C) for the underlined named entities (or false positives) are shown. For interpretability, visual tags (label output of InceptionNet) are presented instead of actual feature vectors used.

enough information to disambiguate it. On the other hand, visual contexts (visual tags: 'parade', 'urban area', ...) provide similarities to the token's distributional semantics from other training examples (*e.g.* snaps from "NBA Championship Parade Story"), and thus the model successfully predicts the token as a named entity. Similarly, while the text-only model erroneously predicts 'Apple' in the caption *"Grandma w dat lit Apple Crisp"* as an organization (*e.g.* Apple Inc.), the visual contexts (describing objects related to food) help disambiguate the token, making the model predict it correctly as a non-named entity (a fruit). Trending entities (musicians or DJs such as 'CID', 'Duke Dumont', 'Marshmello', etc.) are also recognized correctly with strengthened contexts from visual information (describing concert scenes) despite lack of surrounding textual contexts. A few cases where visual contexts harmed the performance mostly include visual tags that are unrelated to a token or its surrounding textual contexts.

Visualization of Modality Attention: Figure 3 visualizes the modality attention module at each decoding step (each column), where amplified modality is represented with darker color, and attenuated modality is represented with lighter color.

For the image-aided model (W+C+V; upper row in Figure 3), we confirm that the modality attention successfully attenuates irrelevant signals (*e.g.* selfies, etc.) and amplifies relevant modality-based contexts in prediction of a given token. In the example of "*disney word essential = coffee*" with visual tags *selfie, phone, person*, the modality attention successfully attenuates distracting visual signals and focuses on textual modalities, consequently making correct predictions. The named entities in the examples of "*Beautiful night atop The Space Needle*" and "*Splash Mountain*" are challenging to predict because they are composed of common nouns (space, needle, splash, mountain), and thus they often need additional contexts to correctly predict. In the training data, visual contexts make stronger indicators for these named entities (space needle, splash mountain), and the modality attention module successfully attends more to stronger signals.

For text-only model (W+C), we observe that performance gains mostly come from the modality attention module better handling tokens unseen during training or unknown tokens from the pre-trained word embeddings matrix. For example, while *WaRriOoOrs* and *Kooler Matic* are missing tokens in the word embeddings matrix, it successfully amplifies character-based contexts (*e.g.* capitalized first letters, similarity to known entities 'Golden State Warriors') and suppresses word-based contexts (word embeddings for unknown tokens *e.g.* 'WaRriOoOrs'), leading to correct predictions. This result is significant because it shows performance of the model, with an almost identical architecture, can still improve without having to scale the word embeddings matrix indefinitely.

Figure 3 (b) shows the cases where the modality attention led to incorrect predictions. For example, the model predicts missing tokens *HUUUGE* and *Shampooer* incorrectly as named entities by amplifying misleading character-based contexts (*e.g.* capitalized first letters) or visual contexts (*e.g.*

Figure 3: **Visualization of modality attention** (a) successful cases and (b) unsuccessful ones from *SnapCaptions* test data. For each decoding step of a token (column), the modality attention module amplifies the most relevant modality (darker) while attenuating irrelevant modalities (lighter). The model makes final predictions based on weighted signals from all modalities. For interpretability, visual tags (label output of InceptionNet) are presented instead of actual feature vectors used. GT: ground-truth, Pred: prediction by our model. Modalities- W: words, C: characters, V: visual.

Vocab Size	w/o M.A.	w/ M.A.
100%	48.8	**50.5**
75%	48.7	**50.1**
50%	47.8	**49.6**
25%	46.4	**48.7**

Table 3: NER performance (F1) on SnapCaptions with **varying word embeddings vocabulary size**. Models being compared: (W+C) Bi-LSTM/CRF + Bi-CharLSTM w/ and w/o modality attention (M.A.)

concert scenes, associated contexts of which often include named entities in the training dataset).

Sensitivity to Word Embeddings Vocabulary Size: In order to isolate the effectiveness of the modality attention module on textual models in handling missing tokens, we report the performance with varying word embeddings vocabulary sizes in Table 3. By increasing the number of missing tokens artificially by randomly removing words from the word embeddings matrix (original vocab size: 400K), we observe that while the overall performance degrades, the modality attention module is able to suppress the peformance degradation. Note also that the performance gap generally gets bigger as we decrease the vocabulary size of the word embeddings matrix. This result is significant in that the modality attention is able to improve the model more robust to missing tokens without having to train an indefinitely large word embeddings matrix for arbitrarily noisy social media text datasets.

5 Conclusions

We proposed a new multimodal NER (MNER: image + text) task on short social media posts. We demonstrated for the first time an effective MNER system, where visual information is combined with textual information to outperform traditional text-based NER baselines. Our work can be applied to myriads of social media posts or other articles across multiple platforms which often include both text and accompanying images. In addition, we proposed the *modality attention* module, a new neural mechanism which learns optimal integration of different modes of correlated information. In essence, the modality attention learns to attenuate irrelevant or uninformative modal information while amplifying the primary modality to extract better overall representations. We showed that the modality attention based model outperforms other state-of-the-art baselines when text was the only modality available, by better combining word and character level information.

References

Gustavo Aguilar, Suraj Maharjan, A. Pastor Lopez-Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. *ACL WNUT Workshop* .

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *ICCV*.

Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*.

William Chan, Navdeep Jaitly, Quoc V Le, and Oriol Vinyals. 2015. Listen, attend and spell. *arXiv:1508.01211* .

Jason PC Chiu and Eric Nichols. 2015. Named entity recognition with bidirectional lstm-cnns. *arXiv:1511.08308* .

Paramveer Dhillon, Dean P Foster, and Lyle H Ungar. 2011. Multi-view learning of word embeddings via cca. In *NIPS*. pages 199–207.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR* .

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR, abs/1510.04709* .

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv:1508.01991* .

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data .

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *NAACL* .

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354* .

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *NAACL*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ICLR* .

Seungwhan Moon and Jaime Carbonell. 2017. Completely heterogeneous transfer learning with attention: What and what not to transfer. *IJCAI* .

Seungwhan Moon, Suyoun Kim, and Haohan Wang. 2015. Multimodal transfer deep learning with applications in audio-visual recognition. In *NIPS MMML Workshop*.

Alexandre Passos, Vineet Kumar, and Andrew McCallum. 2014. Lexicon infused phrase embeddings for named entity resolution. *aarXiv:1404.5367* .

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *CoNLL*.

Marek Rei, Gamal KO Crichton, and Sampo Pyysalo. 2016. Attending to characters in neural sequence labeling models. *COLING* .

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV* .

Erik F Sang and Jorn Veenstra. 1999. Representing text chunks. In *EACL*.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *WMT*.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *NIPS*.

C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. 2014. Going deeper with convolutions. *CVPR* .

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *CoRR* .

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *ICML* .

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *ICCV*.