Image-to-Markup Generation with Coarse-to-Fine Attention

Yuntian Deng[1] Anssi Kanervisto[2] Jeffrey Ling[1]
Alexander M. Rush[1]

[1]Harvard University

[2]University of Eastern Finland

harvardnlp

Outline

harvardnlp

Multimodal Generation

Real text is not disembodied. It always appears in context... As soon as we begin to consider the generation of text in context, we immediately have to countenance issues of **typography** *and* **orthography** *(for the written form) and* **prosody** *(for the spoken form)... This is perhaps most obvious in the case of systems that* **generate both text and graphics** *and attempt to combine these in sensible ways.*

Dale et al. [1998]

harvardnlp

Image to Text

- Natural OCR [Shi et al., 2016, Lee and Osindero, 2016, Mishra et al., 2012, Wang et al., 2012]

`cocacola`

Image to Text

- Natural OCR [Shi et al., 2016, Lee and Osindero, 2016, Mishra et al., 2012, Wang et al., 2012]

cocacola

- Image Captioning [Xu et al., 2015, Karpathy and Fei-Fei, 2015, Vinyals et al., 2015]

A man in street racer armor is examining the tire of another racers motor bike

$$V = i\begin{bmatrix} B & 2\overline{A} \\ -2A & -B \end{bmatrix}$$

IM2LATEX-100K

$$A_{0}^{3}(\alpha^{\prime}\rightarrow 0)=2g_{d}\,\,\varepsilon_{\lambda}^{(1)}\varepsilon_{\mu}^{(2)}\varepsilon_{\nu}^{(3)}\left\{\eta^{\lambda\mu}\left(p_{1}^{\nu}-p_{2}^{\nu}\right)+\eta^{\lambda\nu}\left(p_{3}^{\mu}-p_{1}^{\mu}\right)+\eta^{\mu\nu}\left(p_{2}^{\lambda}-p_{3}^{\lambda}\right)\right\}.$$

A _ { 0 } ^ { 3 } (\alpha ^ { \prime } \rightarrow 0) = 2 g _ { d } \, \, \varepsilon ^ { (1) } _ { \lambda } \varepsilon ^ { (2) } _ { \mu } \varepsilon ^ { (3) } _ { \nu } \left \{ \eta ^ { \lambda \mu } \left (p _ { 1 } ^ { \nu } - p _ { 2 } ^ { \nu } \right) + \eta ^ { \lambda \nu } \left (p _ { 3 } ^ { \mu } - p _ { 1 } ^ { \mu } \right) + \eta ^ { \mu \nu } \left (p _ { 2 } ^ { \lambda } - p _ { 3 } ^ { \lambda } \right) \right \} .

IM2LATEX-100K

$$\left\{ \begin{array}{rcl} \delta_{\epsilon} B & \sim & \epsilon F \, , \\ \delta_{\epsilon} F & \sim & \partial \epsilon + \epsilon B \, , \end{array} \right.$$

```
\left \{ \begin {array} { r c l } \delta _ { \epsilon } B & \sim & \epsilon F \, , \\
\delta _ { \epsilon } F & \sim & \partial \epsilon + \epsilon B \, , \\ \end {array} \right .
```

harvardnlp

IM2LATEX-100K

$$\int\limits_{{\cal L}^{d}_{d-1}} f(H) d\nu_{d-1}(H) = c_{3} \int\limits_{{\cal L}^{A}_{2}} \int\limits_{{\cal L}^{L}_{d-1}} f(H) [H, A]^{2} d\nu_{d-1}^{L}(H) d\nu_{2}^{A}(L).$$

```
\int \limits _ { { \cal L } ^ { d } _ { d - 1 } } f ( H ) d \nu _ { d - 1 } ( H ) = c _ { 3 } \int \limits _ { { \cal L } ^ { A } _ { 2 } }
\int \limits _ { { \cal L } ^ { L } _ { d - 1 } } f ( H ) [ H , A ] ^ { 2 } d \nu _ { d - 1 } ^ { L } ( H ) d \nu _ { 2 } ^ { A } ( L ) .
```

IM2LATEX-100K

$$J = \left(\begin{array}{cc} \alpha^{t} & \tilde{f}_{2} \\ f_{1} & \tilde{A} \end{array} \right) \left(\begin{array}{ll} 0 & 0 \\ 0 & L \end{array} \right) \left(\begin{array}{cc} \alpha & \tilde{f}_{1} \\ f_{2} & A \end{array} \right) = \left(\begin{array}{ll} \tilde{f}_{2} L f_{2} & \tilde{f}_{2} L A \\ \tilde{A} L f_{2} & \tilde{A} L A \end{array} \right)$$

```
J = \left ( \begin {array} { c c } \alpha ^ { t } & \tilde { f } _ { 2 } \\ f _ { 1 } & \tilde { A } \end {array} \right ) \left ( \begin {array}
{ l l } 0 & 0 \\ 0 & L \end {array} \right ) \left ( \begin {array} { c c } \alpha & \tilde { f } _ { 1 } \\ f _ { 2 } & A \end {array} \right )
= \left ( \begin {array} { l l } \tilde { f } _ { 2 } L f _ { 2 } & \tilde { f } _ { 2 } L A \\ \tilde { A } L f _ { 2 } & \tilde { A } L A \end {array}
\right )
```

harvardnlp

IM2LATEX-100K

$$\lambda_{n,1}^{(2)} = \frac{\partial \overline{H}_0}{\partial q_{n,0}} \, , lambda_{n,j_n}^{(2)} = \frac{\partial \overline{H}_0}{\partial q_{n,j_n-1}} - \mu_{n,j_n-1} \, , \quad j_n = 2, 3, \cdots, m_n - 1 \, .$$

\lambda _ { n , 1 } ^ { (2) } = \frac { \partial \overline { H } _ 0 } { \partial q _ { n , 0 } } \ , \ \, l a m b d a _ { n , j _ n } ^ { (2) } = \frac { \partial \overline { H } _ 0 } { \partial q _ { n , j _ n - 1 } } - \mu _ { n , j _ n - 1 } \ , \ \ j _ n = 2 , 3 , \cdots , m _ n - 1 \ .

IM2LATEX-100K

$$(P_{ll'} - K_{ll'})\phi'(z_q)|\chi >= 0$$

(P _ { l l ' } - K _ { l l ' }) \phi ' (z _ { q }) | \chi > = 0

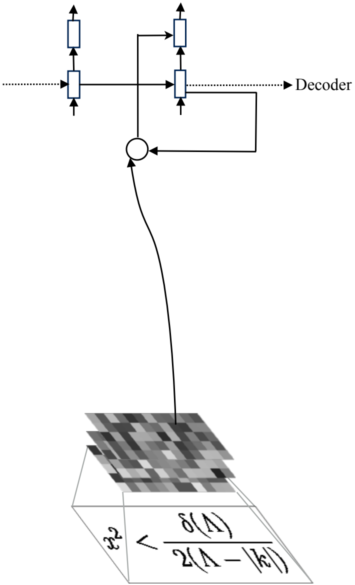
harvardnlp

IM2LATEX-100K

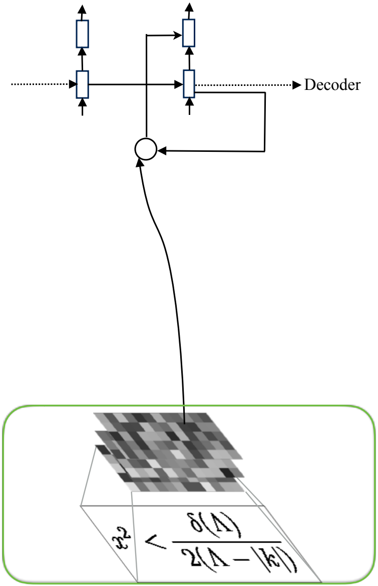
#	img size	median #char	min #char	max #char
103,556	1654×2339	98	38	997

- Originally developed for OpenAI requests for research
- LaTeX sources of arXiv papers on high energy physics from 2003 KDD cup [Gehrke et al., 2003]
- Extracted with regular expressions
- Rendered in a vanilla LaTeX environment

Attention-based Image Captioning (Xu et al. 2015)

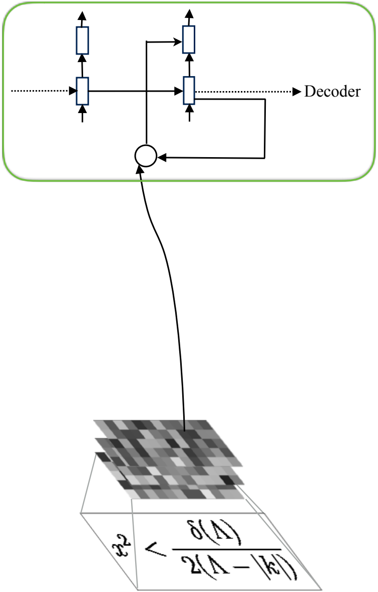


harvardnlp



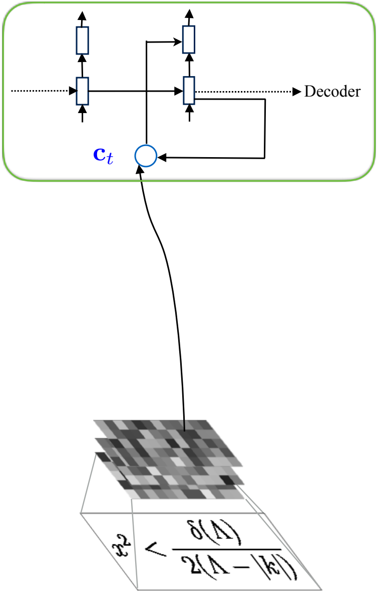
- Encoder: CNN

harvardnlp



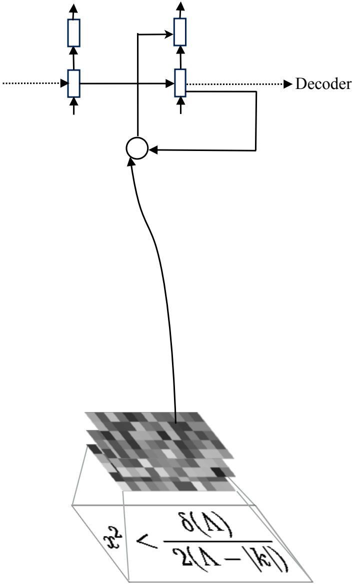
- Encoder: CNN
- Decoder: RNN with attention

harvardnlp



- Encoder: CNN
- Decoder: RNN with attention

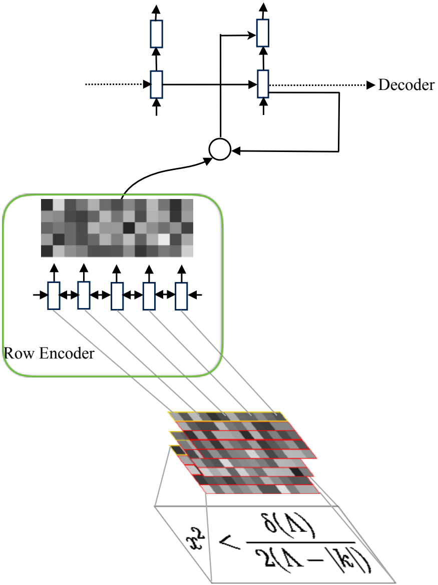
harvardnlp



- Encoder: CNN
- Decoder: RNN with attention
- Objective: maximize log-likelihood

harvardnlp

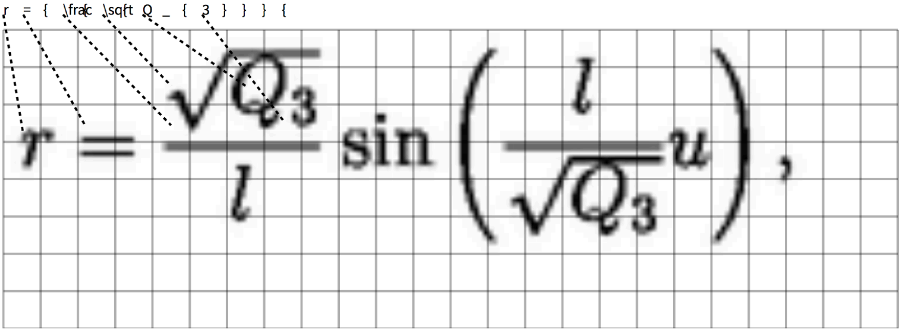
Model Extensions



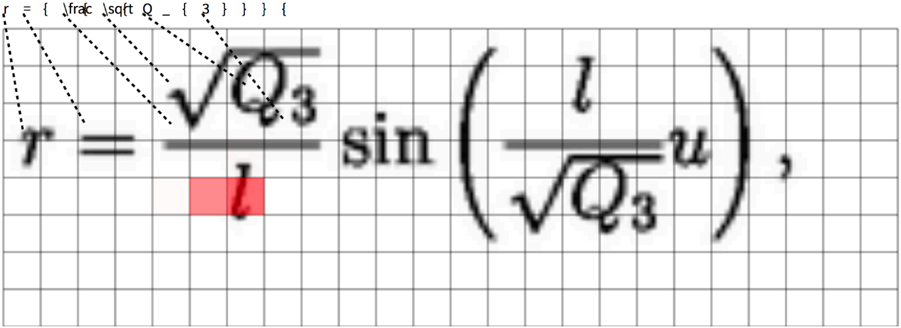
- Row Encoder: RNN over each row of feature map
- Parameters shared across rows
- Row embeddings to initialize RNN

harvardnlp

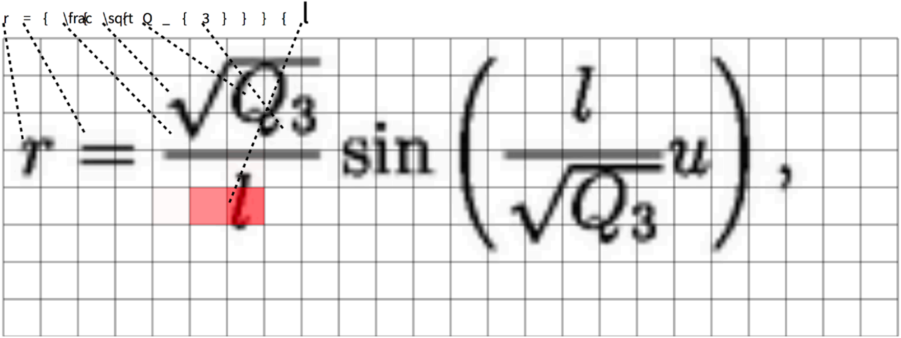
Attention



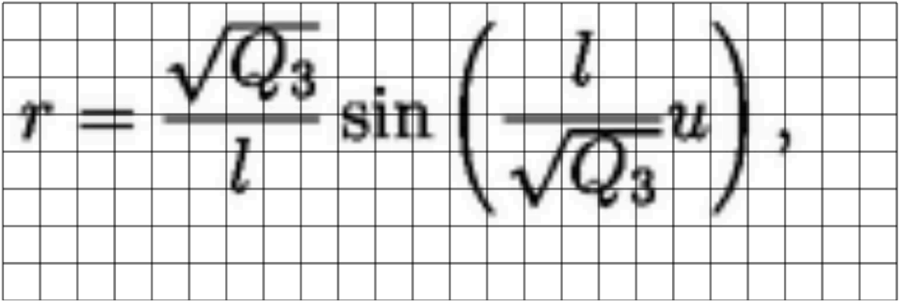
Attention



Attention

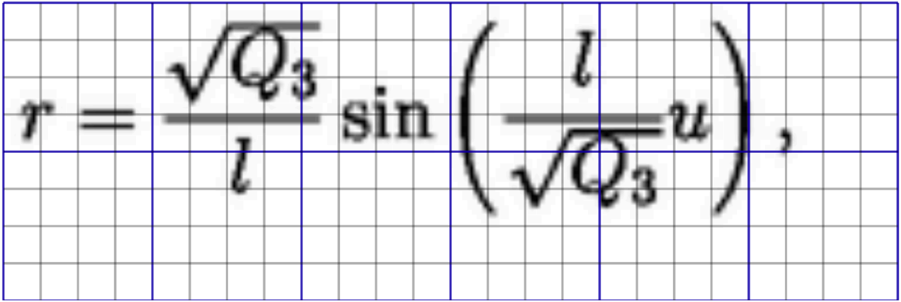


harvardnlp

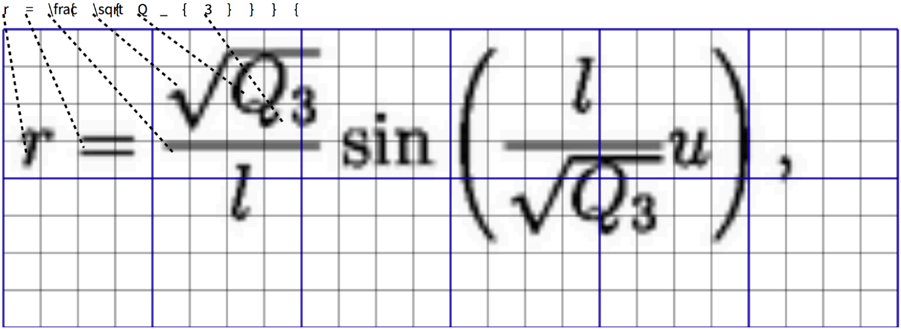


harvardnlp

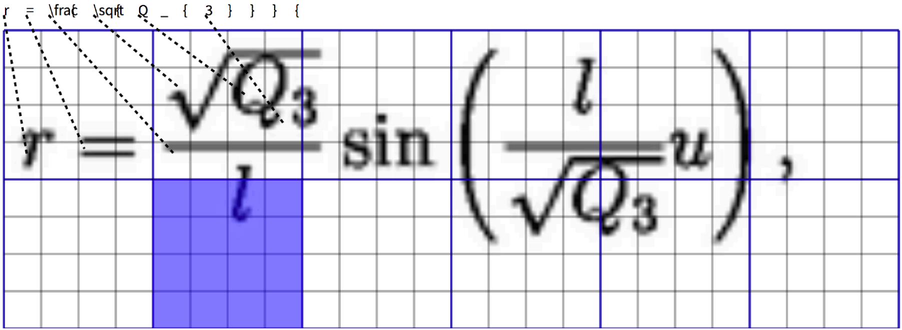
Coarse-to-Fine Attention



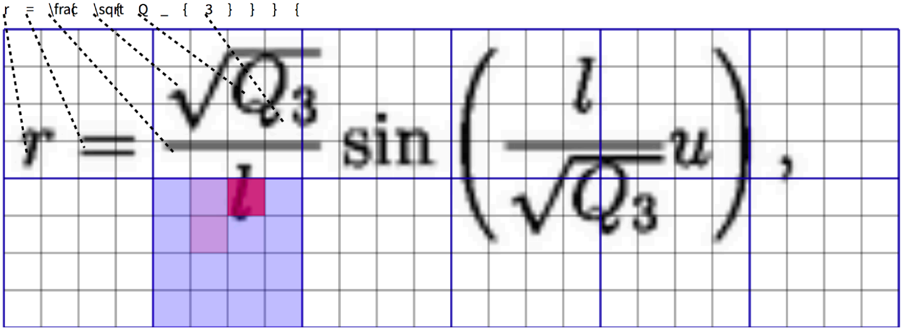
Coarse-to-Fine Attention



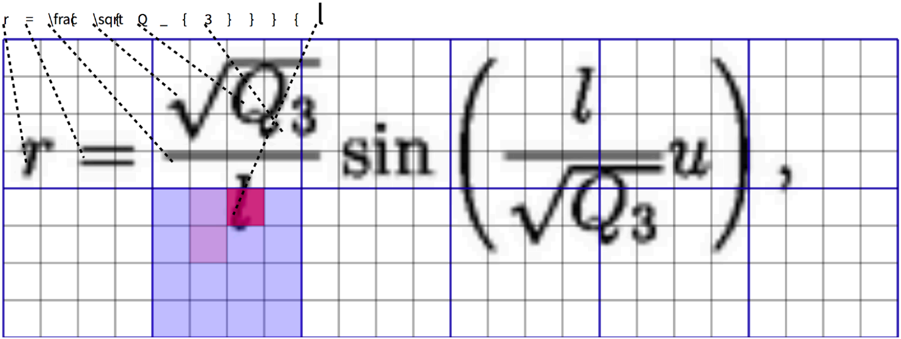
Coarse-to-Fine Attention



harvardnlp

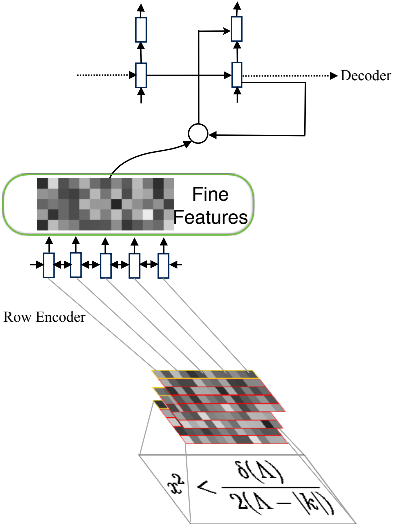

r = \frac { \sqrt { Q _ { 3 } } } {

Coarse-to-Fine Attention



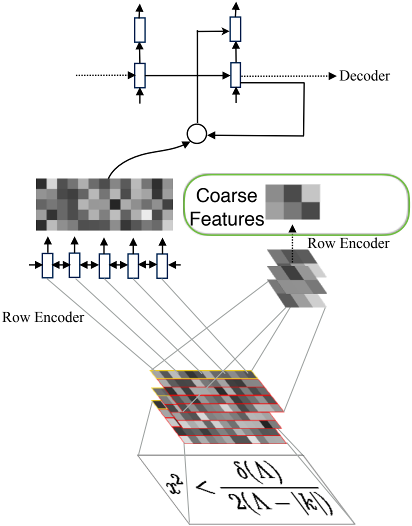
harvardnlp

Coarse-to-Fine Attention

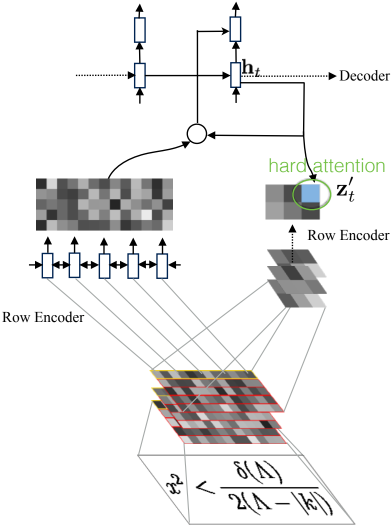


harvardnlp

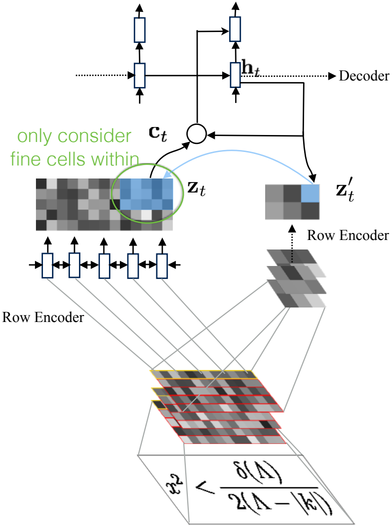
Coarse-to-Fine Attention



Coarse-to-Fine Attention



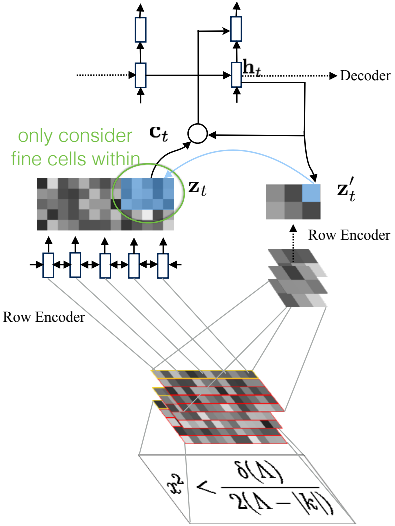
Coarse-to-Fine Attention



$$p(z_t) = \sum_{z'_t} p(z'_t) p(z_t | z'_t)$$

harvardnlp

Coarse-to-Fine Attention



$$p(z_t) = \sum_{z'_t} p(z'_t) p(z_t | z'_t)$$

Coarse-to-Fine Variants

- REINFORCE: hard attention [Xu et al., 2015] to select a **single** coarse cell, the presented model
- SPARSEMAX: use sparse activation function Sparsemax [Martins and Astudillo, 2016] instead of Softmax to select **multiple** coarse cells

harvardnlp

Experiment Details

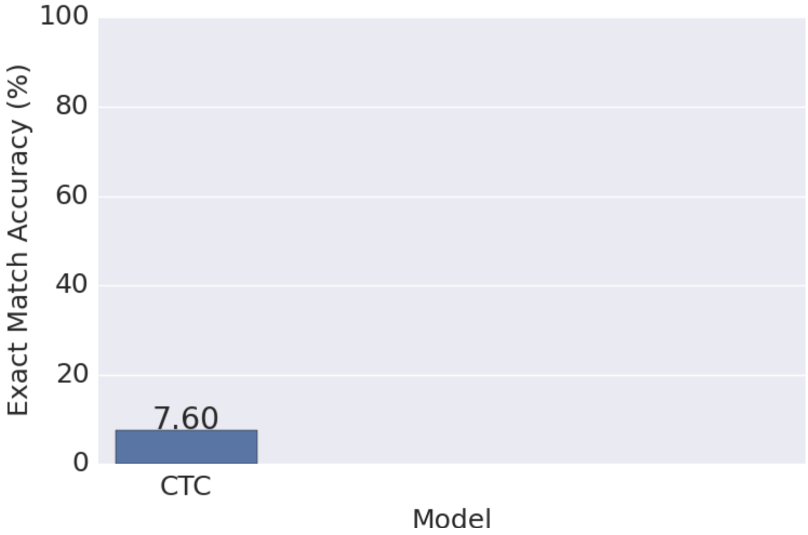
- Tokenization & Normalization:
  `P_{ll'}^1-K^2_{ll}`

  $\Downarrow$

  `P _ { l l ^ { \prime } } ^ { 1 } - K _ { l l } ^ { 2 }`
- Evaluation: exact image match accuracy (rendered prediction versus original image)
- Implementation: Torch [Collobert et al., 2011], based on OpenNMT [Klein et al., 2017]

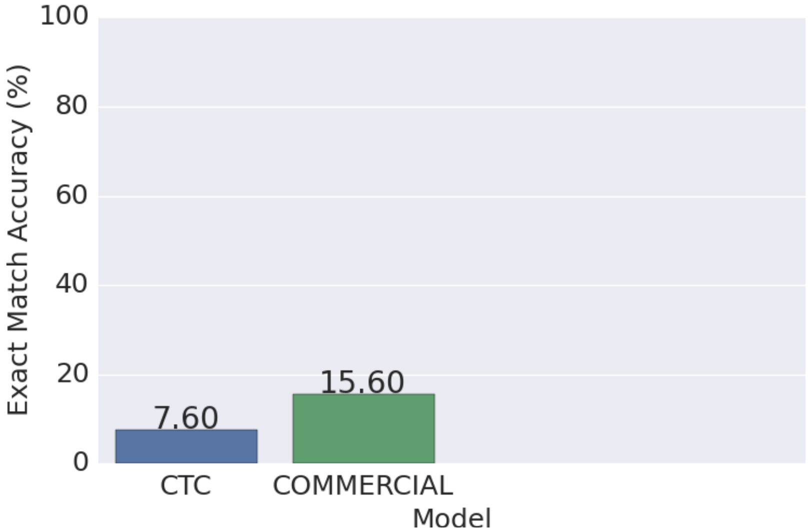
harvardnlp

Baseline Results



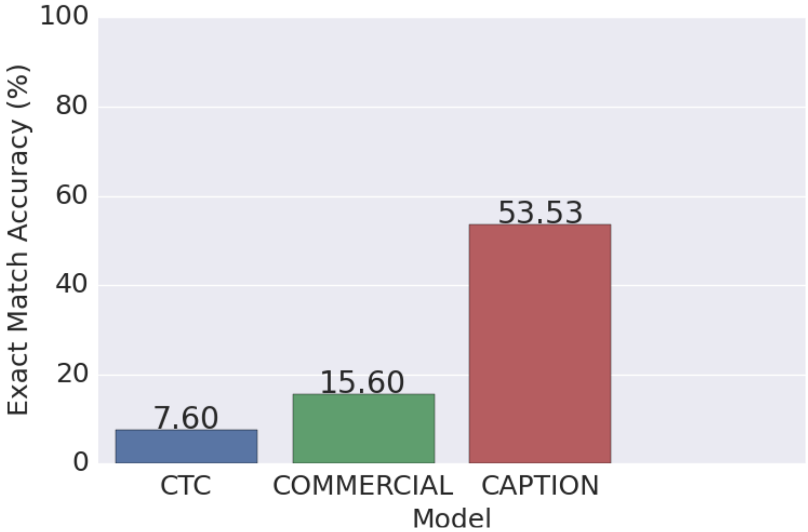
harvardnlp

Baseline Results

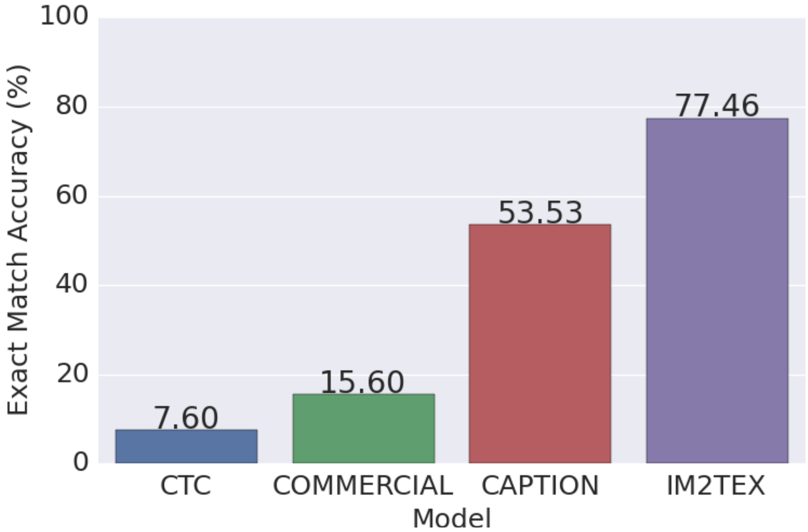


harvardnlp

Baseline Results

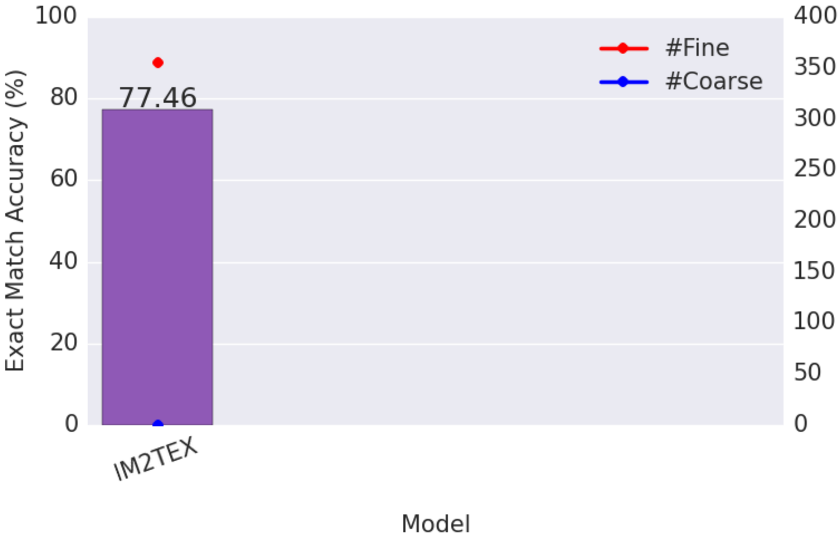


Baseline Results



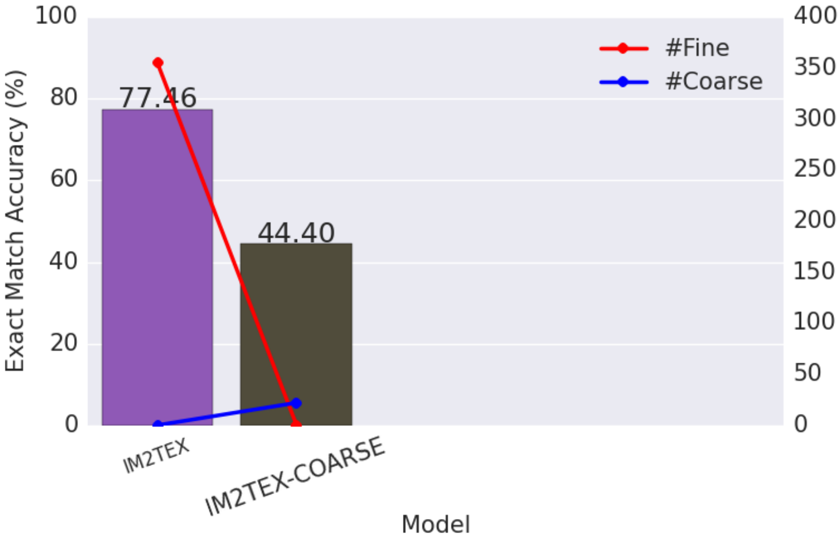
harvardnlp

Main Results



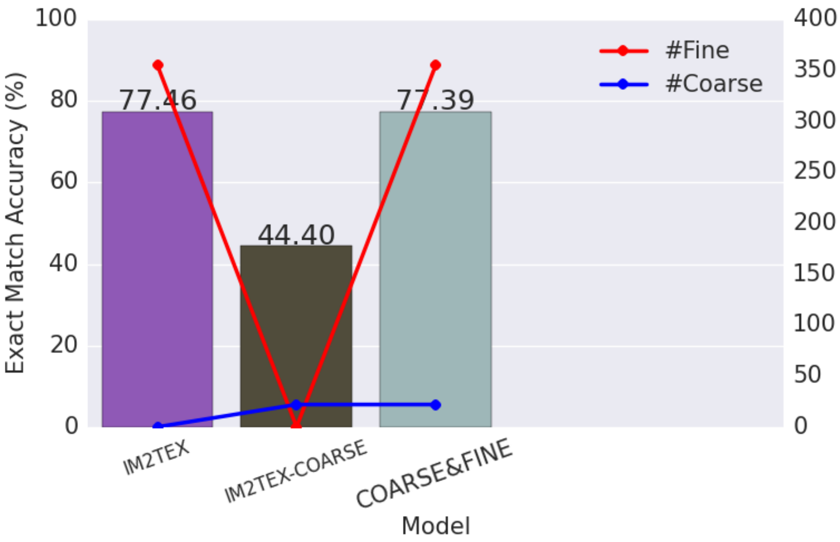
harvardnlp

Main Results



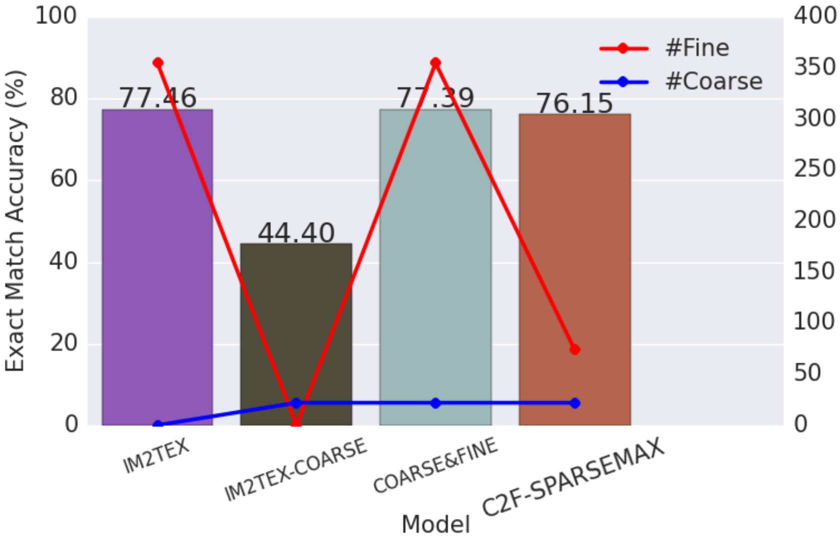
harvardnlp

Main Results

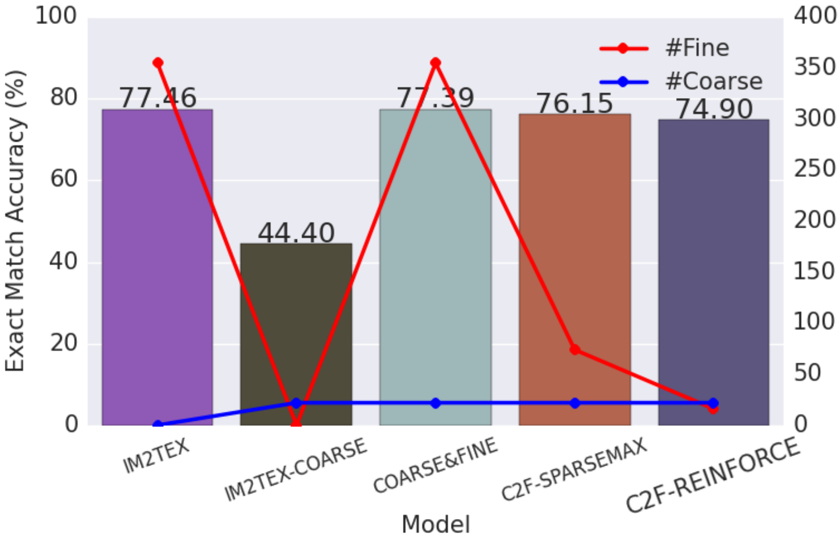


harvardnlp

Main Results



Main Results



harvardnlp

$$Z = \sum_{spins} \prod_{cubes} W(a|e,f,g|b,c,d|h),$$

$$\{\Psi \circ \mu, f\} = \left(\overline{X}_i f\right)\left(Y^i \Psi\right) \circ \mu \, ,$$

$$U_n(\theta,\phi) = \begin{pmatrix} \cos(\theta/2) & -e^{-in\phi}\sin(\theta/2) \\ \sin(\theta/2)e^{in\phi} & \cos(\theta/2) \end{pmatrix}$$

$$\sin\frac{\pi\alpha' s}{2} + \sin\frac{\pi\alpha' t}{2} + \sin\frac{\pi\alpha' u}{2} = -\frac{\pi^3}{16}\alpha'^3 stu + o(\alpha'^5),$$

$$Y(T,U) = \int_{\mathcal{F}} \frac{d^2\tau}{\Im\tau}\Gamma_{2,2}(T,U)\left(-6\left[\overline{\Omega}_2 - \frac{1}{8\pi\Im\tau}\right]\frac{\overline{\Omega}}{\overline{\eta}^{24}} - \frac{\overline{j}}{8} + 126\right) \, ,$$

harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

$$A_0^3(\alpha' \rightarrow 0) = 2g_d\, \varepsilon_\lambda^{(1)}\varepsilon_\mu^{(2)}\varepsilon_\nu^{(3)} \left\{\eta^{\lambda\mu}\left(p_1^\nu - p_2^\nu\right) + \eta^{\lambda\nu}\left(p_3^\mu - p_1^\mu\right) + \eta^{\mu\nu}\left(p_2^\lambda - p_3^\lambda\right)\right\}.$$

```
(A_{0}^{3}(\alpha^{\prime }\rightarrow 0)=2g_{d}\,\,\varepsilon^{(1)}_{\lambda}\varepsilon^{(2)}
_{\mu }\varepsilon^{(3)}_{\nu }\left\{ \eta ^{\lambda \mu}\left( p_{1}^{\nu }-p_{2}^{\nu }\right) +
\eta ^{\lambda \nu }\left(p_{3}^{\mu }-p_{1}^{\mu }\right)+\eta ^{\mu \nu }\left( p_{2}^{\lambda}
-p_{3}^{\lambda }\right) \right\} . \label{17}
```

$$\left\{\begin{array}{rcl}\delta_\epsilon B & \sim & \epsilon F\,, \\ \delta_\epsilon F & \sim & \partial\epsilon + \epsilon B\,, \end{array}\right.$$

```
\left\{\begin{array}{rcl}\delta_{\epsilon} B & \sim & \epsilon F
\, , \\\delta_{\epsilon} F & \sim & \partial\epsilon + \epsilon B
\, , \\\end{array}\right.
```

$$\int\limits_{\mathcal{L}_{d-1}^{d}} f(H)d\nu_{d-1}(H) = c_3 \int\limits_{\mathcal{L}_2^A}\int\limits_{\mathcal{L}_{d-1}^L} f(H)[H,A]^2 d\nu_{d-1}^L(H)d\nu_2^A(L).$$

```
\int \limits_{{\cal L}^{d}_{d-1}}f(H)d\nu_{d-1}(H) =  c_{3} \int \limits_{{\cal L}^{A}_{2}}
\int \limits_{{\cal L}^{L}_{d-1}}f(H)[H,A]^{2}d\nu_{d-1}^{L}(H)d\nu_{2}^{A}(L).
```

$$J = \left(\begin{array}{cc}\alpha^t & \tilde{f}_2 \\ f_1 & \tilde{A}\end{array}\right)\left(\begin{array}{ll}0 & 0 \\ 0 & L\end{array}\right)\left(\begin{array}{cc}\alpha & \tilde{f}_1 \\ f_2 & A\end{array}\right) = \left(\begin{array}{ll}\tilde{f}_2 L f_2 & \tilde{f}_2 L A \\ \tilde{A} L f_2 & \tilde{A} L A\end{array}\right)$$

```
J=\left( \begin{array}{cc}\alpha ^{t} & \tilde{f}_{2} \\ f_{1} & \tilde{A}
\end{array}\right) \left( \begin{array}{ll}0 & 0 \\ 0 & L\end{array}\right)
\left( \begin{array}{cc}\alpha & \tilde{f}_{1} \\ f_{2} & A\end{array}\right) =
\left( \begin{array}{ll}\tilde{f}_{2}Lf_{2} & \tilde{f}_{2}LA \\ \tilde{A}Lf_{2} &
\tilde{A}LA\end{array}\right)
```

$$\lambda_{n,1}^{(2)} = \frac{\partial \overline{H}_0}{\partial q_{n,0}}\ ,lambda_{n,j_n}^{(2)} = \frac{\partial \overline{H}_0}{\partial q_{n,j_n-1}} - \mu_{n,j_n-1}\ ,\quad j_n = 2,3,\cdots,m_n-1\ .$$

```
\lambda_{n,1}^{(2)}=\frac{\partial\overline{H}_0}{\partial q_{n,0}}\ ,\ \\lambda_{n,j_n}^{(2)}=\frac{
\partial\overline{H}_0}{\partial q_{n,j_n-1}}-\mu_{n,j_n-1}\ ,\      \ j_n=2,3,\cdots,m_n-1\ .
```

$$(P_{ll'} - K_{ll'})\phi'(z_q)|\chi >= 0$$

```
(P_{ll'} - K_{ll'}) \phi '(z_{q})|\chi > = 0
```

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

$$A_0^3(\alpha' \rightarrow 0) = 2g_d\, \varepsilon_\lambda^{(1)} \varepsilon_\mu^{(2)} \varepsilon_\nu^{(3)} \left\{ \eta^{\lambda\mu}(p_1^\nu - p_2^\nu) + \eta^{\lambda\nu}(p_3^\mu - p_1^\mu) + \eta^{\mu\nu}(p_2^\lambda - p_3^\lambda) \right\}$$

`(A_{0}^{3}(\alpha^{\prime }\rightarrow 0)=2g_{d}\,\,\varepsilon^{(1)}_{\lambda}\varepsilon^{(2)}_{\mu }\varepsilon^{(3)}_{\nu }\left\{ \eta ^{\lambda \mu}\left(p_{1}^{\nu }-p_{2}^{\nu }\right) + \eta ^{\lambda \nu }\left(p_{3}^{\mu }-p_{1}^{\mu }\right)+\eta ^{\mu \nu }\left(p_{2}^{\lambda} -p_{3}^{\lambda }\right) \right\} . \label{17}`

$$\left\{\begin{array}{rcl}\delta_{\epsilon} B & \sim & \epsilon F, \\ \delta_{\epsilon} F & \sim & \partial\epsilon + \epsilon B,\end{array}\right.$$

`\left\{\begin{array}{rcl}\delta_{\epsilon} B & \sim & \epsilon F \, , \\delta_{\epsilon} F & \sim & \partial\epsilon + \epsilon B \, , \\end{array}\right.`

$$\int\limits_{\mathcal{L}^d_{d-1}} f(H)d\nu_{d-1}(H) = c_3 \int\limits_{\mathcal{L}^A_2} \int\limits_{\mathcal{L}^L_{d-1}} f(H)[H,A]^2 d\nu^L_{d-1}(H)d\nu_2^A(L)$$

`\int \limits_{{\cal L}^{d}_{d-1}}f(H)d\nu_{d-1}(H) = c_{3} \int \limits_{{\cal L}^{A}_{2}} \int \limits_{{\cal L}^{L}_{d-1}}f(H)[H,A]^{2}d\nu_{d-1}^{L}(H)d\nu_{2}^{A}(L).`

$$J = \left(\begin{array}{cc}\alpha^t & \tilde{f}_2 \\ f_1 & \tilde{A}\end{array}\right)\left(\begin{array}{ll}0 & 0 \\ 0 & L\end{array}\right)\left(\begin{array}{cc}\alpha & \tilde{f}_1 \\ f_2 & A\end{array}\right) = \left(\begin{array}{ll}\tilde{f}_2 L f_2 & \tilde{f}_2 L A \\ \tilde{A} L f_2 & \tilde{A} L A\end{array}\right)$$

`J=\left(\begin{array}{cc}\alpha ^{t} & \tilde{f}_{2} \\ f_{1} & \tilde{A} \end{array}\right) \left(\begin{array}{ll}0 & 0 \\ 0 & L\end{array}\right) \left(\begin{array}{cc}\alpha & \tilde{f}_{1} \\ f_{2} & A\end{array}\right) = \left(\begin{array}{ll}\tilde{f}_{2}Lf_{2} & \tilde{f}_{2}LA \\ \tilde{A}Lf_{2} & \tilde{A}LA\end{array}\right)`

$$\lambda^{(2)}_{n,1} = \frac{\partial \overline{H}_0}{\partial q_{n,0}}, \ lambda^{(2)}_{n,j_n} = \frac{\partial \overline{H}_0}{\partial q_{n,j_n-1}} - \mu_{n,j_n-1}, \ j_n = 2,3, \cdots, m_n - 1.$$

`\lambda_{n,1}^{(2)}=\frac{\partial\overline{H}_0}{\partial q_{n,0}}\ ,\ \\lambda_{n,j_n}^{(2)}=\frac{ \partial\overline{H}_0}{\partial q_{n,j_n-1}}-\mu_{n,j_n-1}\ ,\ \ j_n=2,3,\cdots,m_n-1\ .`

$$(P_{ll'} - K_{ll'})\phi'(z_q)|\chi> = 0$$

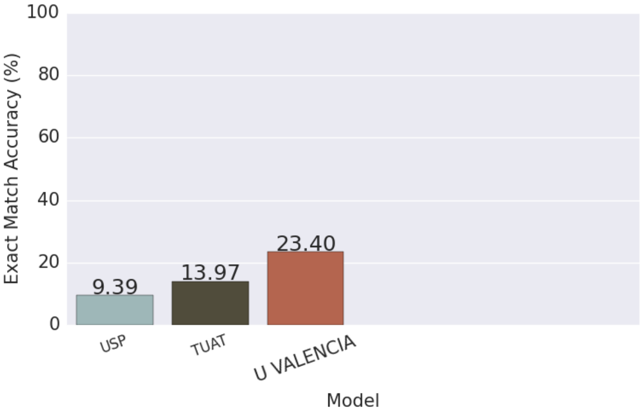
`(P_{ll'} - K_{ll'}) \phi '(z_{q})|\chi > = 0`

harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

CROHME 13

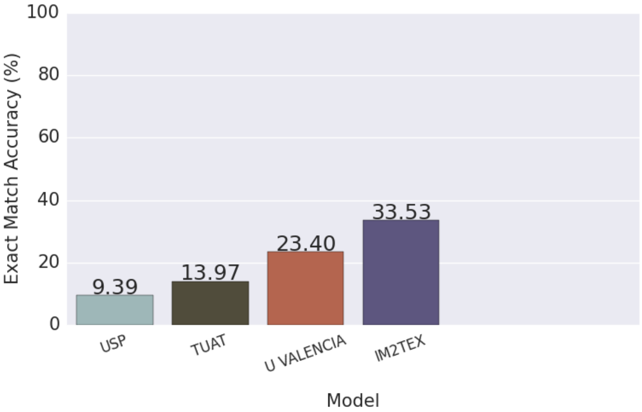


harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

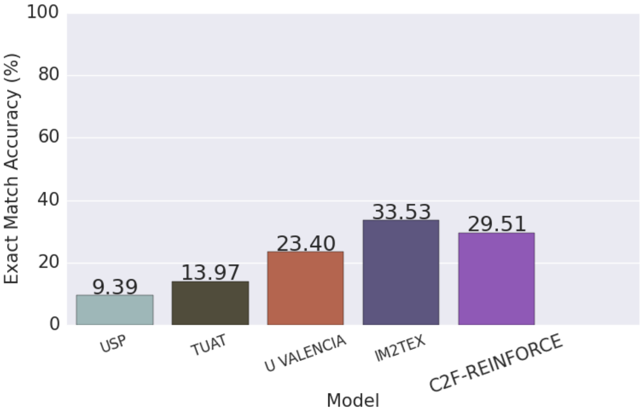
CROHME 13



Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

CROHME 13

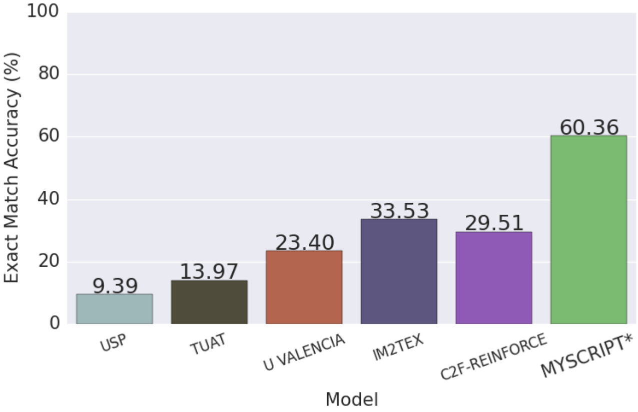


harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

CROHME 13 (*uses private in-domain handwritten training data)

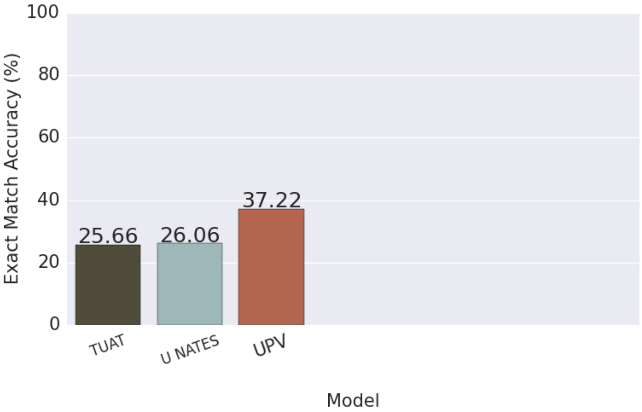


harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

CROHME 14

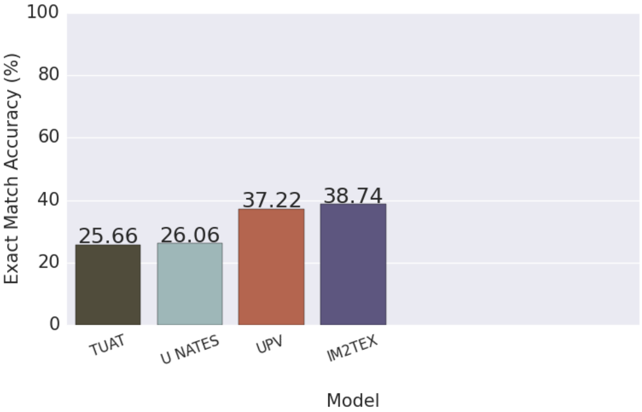


harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

CROHME 14

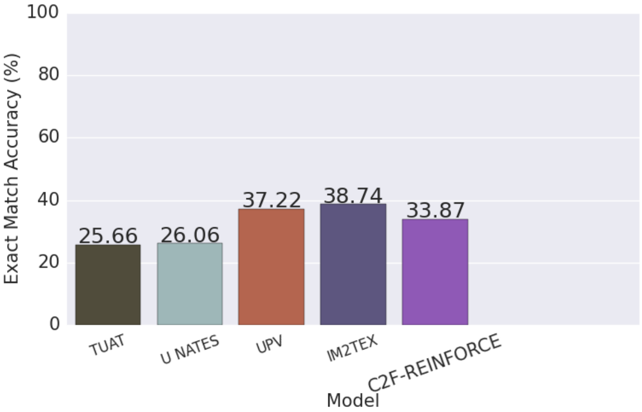


harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

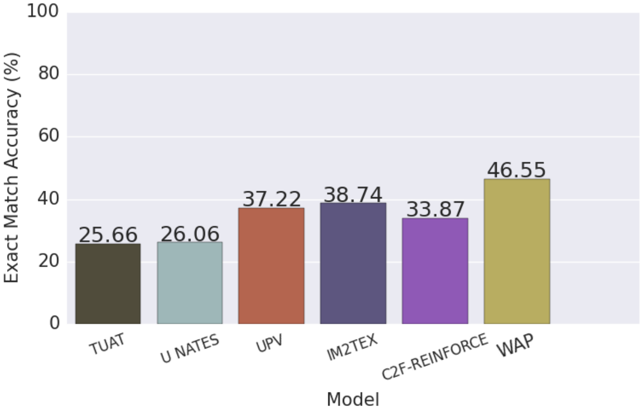
CROHME 14



Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

CROHME 14 (WAP: Zhang et al. [2017])

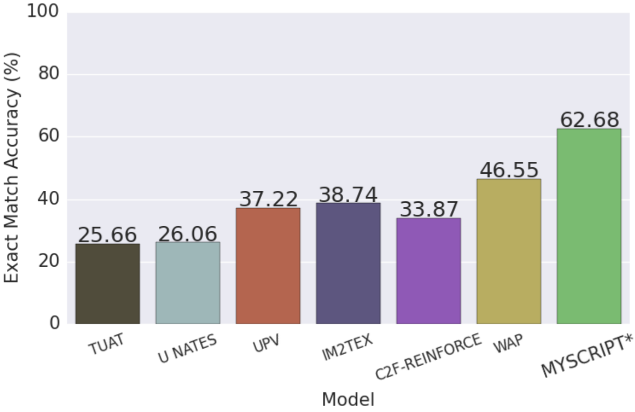


harvardnlp

Handwritten Formulas

- Synthetic handwritten formulas by using handwritten characters [Kirsch, 2010] as font, used for pretraining
- Finetune and evaluate on CROHME 13 and 14 (8K training set)

CROHME 14 (*uses private in-domain handwritten training data)



harvardnlp

Conclusions & Future Work

- The constructed dataset $\mathrm{IM2LATEX\text{-}100K}$ is rich structured and challenging
- A case study of multi-modal document recognition/generation
- Coarse-to-fine attention can be applied to other tasks

References

R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, number EPFL-CONF-192376, 2011.

R. Dale, D. Scott, and B. Di Eugenio. Introduction to the special issue on natural language generation. *Computational Linguistics*, 24(3):346–353, 1998.

J. Gehrke, P. Ginsparg, and J. Kleinberg. Overview of the 2003 kdd cup. *ACM SIGKDD Explorations Newsletter*, 5(2):149–151, 2003.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376. ACM, 2006.

A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

D. Kirsch. *Detexify: Erkennung handgemalter LaTeX-symbole*. PhD thesis, Diploma thesis, Westfälische Wilhelms-Universität Münster, 10 2010.[Online]. Available: http://danielkirs. ch/thesis. pdf, 2010.

G. Klein, Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprint arXiv:1701.02810*, 2017.

C.-Y. Lee and S. Osindero. Recursive recurrent nets with attention modeling for ocr in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2231–2239, 2016.

A. Martins and R. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623, 2016.

A. Mishra, K. Alahari, and C. Jawahar. Scene text recognition using higher order language priors. In *BMVC 2012-23rd British Machine Vision Conference*. BMVA, 2012.

B. Shi, X. Bai, and C. Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2016.

M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, and T. Kanahori. Infty: an integrated ocr system for mathematical documents. In *Proceedings of the 2003 ACM symposium on Document engineering*, pages 95–104. ACM, 2003.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

T. Wang, D. J. Wu, A. Coates, and A. Y. Ng. End-to-end text recognition with convolutional neural networks. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3304–3308. IEEE, 2012.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *Pattern Recognition*, 2017.

Q & A

- More visualizations:
  `http://lstm.seas.harvard.edu/latex/`
- Source code (part of OpenNMT):
  `http://opennmt.net/OpenNMT/applications/`

harvardnlp